One-Shot Learning with Bayesian Networks

Andrew L. Maas (amaas@andrew.cmu.edu)
Computer Science Department & Center for the Neural Basis of Cognition
Carnegie Mellon University

Charles Kemp (ckemp@cmu.edu)
Department of Psychology & Center for the Neural Basis of Cognition
Carnegie Mellon University

Abstract

Humans often make accurate inferences given a single exposure to a novel situation. Some of these inferences can be achieved by discovering and using near-deterministic relationships between attributes. Approaches based on Bayesian networks are good at discovering and using soft probabilistic relationships between attributes, but typically fail to identify and exploit near-deterministic relationships. Here we develop a Bayesian network approach that overcomes this limitation by learning a hyperparameter for each distribution in the network that specifies whether it is non-deterministic or near-deterministic. We apply our approach to one-shot learning problems based on a real-world database of immigration records, and show that it outperforms a more standard Bayesian network approach.

Keywords: Machine Learning; One-Shot Learning; Concepts and Categories; Bayesian Modeling

Introduction

Humans are able to discover and exploit relationships between attributes (e.g. nationality and language) and between attribute values (e.g. Brazilian and Portuguese) (Davies & Russell, 1987). Some relationships are near-deterministic, including the relationship between birth country and native language. We know, for example, that two individuals born in the same country are very likely to have the same mother tongue, and we know in particular that individuals born in Brazil are very likely to speak Portuguese. Other relationships are probabilistic, including the relationship between hair color and eye color. We know that these attributes tend to be related, and we know about specific relationships between values of these attributes (blondes often have blue eyes).

Suppose, for example, that after meeting several people from various countries, you meet a single person from Randeria, a country that is completely new to you. You observe that the person has blonde hair and speaks Randerian. Based on this single example, you may be very confident that the next Randerian you meet will speak the same language, but less confident that this second Randerian will also have blonde hair. Figure 1(a) shows a schematic representation of the observed data, and Figure 1(b) shows conditional distributions that capture our expectations about the language and hair color of the second Randerian. The Randeria problem just introduced is a special case of the more general problem of *one-shot learning* (Fei-Fei, Fergus, & Perona, 2003). Here we describe and evaluate a probabilistic model that can handle one-shot learning problems similar to the Randeria problem.

(a)

Nationality	Language	Hair
British	English	Red
British	English	Brown
Brazilian	Portuguese	Blonde
Brazilian	Portuguese	Brown
Spanish	Spanish	Brown
Spanish	Basque	Black
Randerian	Randerian	Blonde
Randerian	?	?

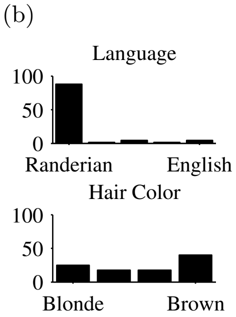

Figure 1: Randeria one-shot learning problem. (a) After meeting people from several different countries, you might discover that people from the same country tend to speak the same language. (b) Discovering the pattern in (a) supports one-shot learning about people from a new country. After observing a single Randerian, you might have strong expectations about the language spoken by a subsequent Randerian, but weak expectations about her hair color.

One-shot learning has been previously considered in the psychological literature. One prominent line of work has focused on “fast mapping” in word learning (Carey & Bartlett, 1978; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Empirical studies of word learning have documented that children are able to learn the meaning of some new words given a single training example and researchers have developed formal models (Colunga & Smith, 2005; Kemp, Perfors, & Tenenbaum, 2007) that help to explain this ability. Our approach grows out of this literature, and the work we describe builds on the hierarchical Bayesian model presented by Kemp et al. (2007). Hierarchical Bayesian models (Gelman, Carlin, Stern, & Rubin, 2003) can include representations at multiple levels of abstraction, and help to explain how humans acquire abstract knowledge that supports rapid or one-shot learning given exposure to a novel situation.

Our hierarchical Bayesian approach is built on top of a standard method for learning Bayesian networks, also

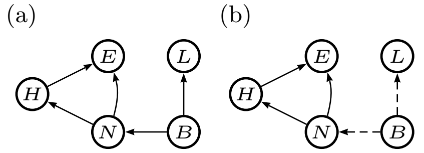

Figure 2: Models that capture relations among five attributes: birth country (B), language (L), nationality (N), eye color (E) and hair color (H). (a) A standard Bayes net can capture probabilistic relationships between attributes, shown here as solid arrows. (b) Our model learns Bayes nets that capture two kinds of relationships: near-deterministic relationships (dashed arrows) and probabilistic relationships (solid arrows).

known as Bayes nets. A Bayes net captures relationships between attributes using probability distributions that specify how the value of a given attribute is generated given the values of its parents. Our approach allows for two kinds of relationships: relationships where an attribute value is a soft probabilistic function of the values of its parent attributes, and relationships where an attribute value is generated in a near-deterministic way given the values of its parents (Figure 2b). By learning which relationships are probabilistic and which are near-deterministic, a Bayes net approach can account for one-shot learning while preserving the ability to handle probabilistic relationships.

After reviewing related work and introducing our approach, we apply it to an everyday problem that requires one-shot inferences—learning about people and their characteristics. Using demographic data for immigrants who arrived at Ellis island in the early twentieth century, we introduce two one-shot learning scenarios which correspond to real-world versions of the Randeria problem. We show that our model makes more intuitive inferences and predicts unobserved data better than a standard Bayesian network approach.

Logical Approaches To One-Shot Learning

One-shot learning has been previously considered by AI researchers, and the Randeria example introduced above is directly inspired by the work of Davies and Russell (1987). These researchers explore the role of determinations, or abstract logical statements that identify patterns of dependency between attributes. For example, the statement that “people of the same nationality speak the same language” is a determination that supports the conclusion that all citizens of Randeria are likely to speak the same language. Because this rule is defined over attributes, it is independent of any particular country and can be used to perform one-shot learning when exposed to a person from a new country. Russell (1989) discusses how determinations can be learned given a database such as the schematic example in Figure 1(a). The basic approach is to search through a hypothesis space of possible determinations and identify hypotheses that are consistent with the entries in the database.

A probabilistic approach to learning determinations can improve on existing work in several respects. First, a probabilistic approach can handle near-deterministic relations that are subject to noise and exceptions. Some citizens of Randeria may be English speakers who were born in the USA, and some countries (e.g. Spain) include different linguistic communities (e.g. Spanish speakers and Basque speakers). Second, a probabilistic approach can incorporate soft probabilistic relations, including the relationship between blonde hair and blue eyes. Russell (1989) allows for weighted determinations which can help to deal with uncertainty, but a probabilistic approach provides a principled treatment of reasoning under uncertainty. Finally, a probabilistic approach can provide a unified account of learning and using determinations. Logical approaches can rely on logical inference to explain how determinations are used, but must typically invoke some other principle to explain how these determinations are acquired.

There has traditionally been some tension between logical and probabilistic approaches to artificial intelligence, but several researchers have recently developed general-purpose frameworks that combine logic and probability (Milch et al., 2005; Richardson & Domingos, 2006). Some of these frameworks may be able to address the one-shot learning problems described earlier, but here we take a different approach. General-purpose frameworks are impressive in their scope, but the flexibility of these approaches often leads to very difficult learning problems. Here we describe a relatively simple probabilistic approach that relies on one of the best known formalisms for capturing relationships between attributes—Bayesian networks.

Learning Bayesian networks

A Bayesian network includes a graph and a set of distributions that specify probabilistic relationships between attributes. This section introduces a standard approach to learning and using these networks (Heckerman, Geiger, & Chickering, 1995).

A Bayes net can be represented as a pair (G, θ), where G is a directed acyclic graph over the attributes of interest and θ_i specifies the conditional probability distribution for attribute i, or the distribution over values of this attribute given the values of its parent attributes in graph G (Figure 3). Figure 2a shows a Bayes net graph structure over some of the attributes in the Randeria problem.

We assume here that all attributes are categorical, and represent θ_i as a conditional probability table (CPT) with one row for each setting of the parent attributes. Each row in θ_i specifies a multinomial distribution over

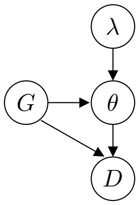

Figure 3: Graphical model for Bayes net structure learning. (G, θ) is a Bayes net, where G is a directed acyclic graph, and θ_i is a table that specifies the conditional probability distributions for node i in the graph. Each row in θ_i is drawn from a symmetric Dirichlet distribution with parameter λ_i.

values of attribute i, and we assume that these rows are independently drawn from a symmetric Dirichlet distribution with concentration parameter λ_i. A standard approach to structure learning sets $\lambda_i = 1$ for all attributes in the graph, which corresponds to a uniform prior over possible multinomial distributions for the rows in each CPT.

Suppose that we observe a data matrix D, where the rows in D represent independent samples from a Bayes net (G, θ). The posterior distribution over the components of the Bayes net is

$$p(G, \theta|D, \lambda) \propto p(D|G, \theta)p(\theta|G, \lambda)p(G) \qquad (1)$$

and we assume a uniform prior $p(G)$ over graph structures G. Since we use conjugate Dirichlet priors on the rows in each CPT, we can integrate out the parameters θ and work with the posterior distribution $p(G|D, \lambda)$ over graphs (Heckerman et al., 1995). We can sample from this distribution using standard MCMC techniques for structure learning (Giudici & Castelo, 2003). If we assume that any missing entries in D are missing at random, a bag of samples from $P(G|D)$ can be used to make predictions about these missing entries.

Bayesian networks have been widely used in the psychological literature to develop formal models of learning and reasoning (Glymour, 2001; Gopnik et al., 2004) The standard approach to learning these networks, however, cannot address one-shot learning problems like the Randeria problem. This limitation depends critically on the difference between attributes (e.g. nationality) and attribute values (e.g. Brazilian). Given enough data, the standard approach will be sensitive to near-deterministic relationships between attribute *values*. After observing many Brazilian individuals, for example, the standard approach will learn parameters for the network in Figure 2a that specify a near-deterministic relationship between being Brazilian and speaking Portuguese. No amount of experience, however, will allow the standard approach to exploit near-deterministic relationships between *attributes*. The standard approach can learn that Brazilians tend to speak Portuguese, and that Americans tend to speak English, and so on, but cannot arrive at the generalization that individuals from a given country tend to speak the same language. The next section introduces a Bayesian network approach that overcomes this limitation.

The Type-Learning Model

Our approach relies on the same basic machinery as the standard approach, except that we no longer assume λ is fixed to a single, known value for all attributes in the graph. Instead we assume that attributes come in one of two types: *non-deterministic* attributes are generated in a soft probabilistic way by their parents in the graph, but *near-deterministic* attributes are generated according to a near-deterministic function of their parent attributes. To capture the difference between these types of attributes, we assume λ_i will be smaller for near-deterministic attributes than for non-deterministic attributes. A small value of λ_i means that each row in CPT θ_i is expected have most of its probability mass concentrated on a single value of attribute i. Setting $\lambda_i = 1$, which is a standard practice when learning Bayes nets, means that each row of θ_i is drawn from a uniform prior over multinomial distributions.

A type-based approach could be implemented by assuming that each λ_i is drawn from one of two distributions: a distribution with a small mean for the near-deterministic attributes, and a distribution with mean 1 for the non-deterministic attributes. Here we take a simpler approach, and assume that $\lambda_i = 1$ for non-deterministic attributes but that $\lambda_i = 0.01$ for near-deterministic attributes. Note, however, that the type assignment for each attribute is not known in advance and must be learned.

A type-based approach can be contrasted with a type-free approach that assumes that the λ_i are independently generated from a continuous prior distribution such as an exponential distribution. These two approaches incorporate different inductive biases and should lead to slightly different predictions—for example, the type-based approach might be quicker to decide whether a given attribute is near-deterministic (low λ_i) or non-deterministic (high λ_i). Future work can consider whether a type-based or a type-free approach accounts better for human inferences. Note, however, that both approaches are consistent with our core proposal, which is that learning different values of λ_i for different attributes can allow a Bayes net approach to handle one-shot learning problems like the Randeria problem.

Since the type assignments that determine λ are not known in advance, we work with a posterior distribution created by summing over all possible values of λ:

$$\begin{aligned} p(G, \theta|D) &\propto p(D|G, \theta)p(\theta|G)p(G) && (2)\\ &= \sum_{\lambda} p(D|G, \theta)p(\theta|G, \lambda)p(G)p(\lambda) && (3) \end{aligned}$$

We use a uniform prior over type assignments, which amounts to a uniform prior over the two possible values of λ_i for any attribute i. Standard MCMC techniques for structure learning can be extended to sample from $P(G, \lambda|D)$, but for the small data sets considered here we compute Equation 3 by enumerating all possible values of λ. As for the standard approach in Equation 1, the parameters θ can be integrated out for any given value of λ, and we make inferences about missing values in D using a bag of samples from the learned distribution $P(G, \lambda|D)$.

Related Work

A special case of our general approach has previously been discussed in the psychological literature. Kemp et al. (2007) describe a Bayesian model that can discover, for example, that objects in the same category tend to have the same same shape—in other words, that the relationship between category label and shape is near-deterministic. Their model, however, works with a restricted class of Bayes nets where there is an arrow from the category label attribute to each other attribute, and where no other edges are allowed. The model developed here can handle Bayes nets with arbitrary structure, including networks that specify relationships between attributes (e.g. hair color and eye color) that do not correspond to category labels.

Our emphasis on near-deterministic relationships is consistent with previous suggestions that humans assume by default that causal relationships will be deterministic (Schulz & Sommerville, 2006). Previous researchers have developed probabilistic approaches that can exploit deterministic relationships when they are present. Closest to our own approach is the work of Lucas and Griffiths (2007), who describe a hierarchical Bayesian model that can learn whether causal observations are better explained by a deterministic relationship or a noisy-OR relationship between variables. Note, however, that this model does not handle settings where a single network includes both near-deterministic and non-deterministic relationships, and cannot address one-shot learning problems like the Randeria problem considered here.

Our approach to one-shot learning relies critically on the concentration parameters λ_i used to define the Dirichlet priors on the Bayes net parameters θ. We know of no previous work that explores one-shot learning with Bayesian networks, but several previous researchers have emphasized the role of the Dirichlet priors. One line of work explores structure learning in the standard setting where there is a single value of λ for all nodes in the network, and has demonstrated that the value of this parameter plays an important role in determining the graph structure G that maximizes $P(G|D)$ (Steck, 2008; Silander, Kontkanen, & Myllymäki, 2007). When λ is very small, the best graph structure will often have very few edges, and as λ increases the number of edges in the inferred graph will also tend to increase. This result suggests that the value of λ matters, and supports the idea that predictive accuracy may be improved by choosing different λ_i values for near-deterministic and non-deterministic nodes.

Table 1: Passenger Data Attributes

Attribute	Example	# Values
Nationality	Spain	24
Race	Spanish	16
Language	Spanish	12
Birth Country	Spain	24
Complexion	Dark	2
Hair	Black	4
Eyes	Brown	7

Previous authors have explored the possibility of learning a single λ parameter for the entire network (Giudici & Green, 1999), but there are few attempts to learn different values of λ_i for different attributes. One possible reason is that this approach is inconsistent with the assumption of likelihood equivalence, or the assumption that networks in the same Markov equivalence class should receive the same prior probability (Heckerman et al., 1995). Although likelihood equivalence is often appealing, it will not always apply in settings where prior knowledge is available about network parameters. Our setting is one example, and the knowledge in this case specifies that some relationships are near-deterministic but that others are probabilistic.

Experiments

We evaluate our approach in two ways using a real-world data set. First, we directly model the Randeria problem to show the practical consequences of modeling near-deterministic relationships. Second, we use a larger test set to demonstrate the quantitative differences between inferences made by our model and a standard Bayes net approach.

Passenger Data

Our experiments used a real-world version of the data set shown schematically in Figure 1(a). The data specify physical and cultural properties of immigrants who arrived at Ellis Island during the 1920s and 1930s, and were extracted from passenger manifests available at `ellisisland.org`. We took manifests for 4 ships and created a data set with 85 people and 7 categorical attributes[1]. Table 1 shows each attribute, its number of possible values, and example values for one person. The relationships between the attributes include both near-deterministic relationships (country determines language) and soft probabilistic relationships (hair color predicts eye color). Note, however, that the near-deterministic relationships are not perfectly clean (e.g. not everyone from Spain speaks Spanish).

[1] The data set is available online at `www.andrew-maas.net`

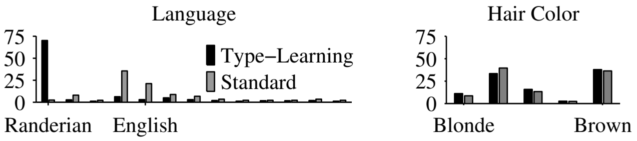

Figure 4: Conditional distributions on the language and hair color of a new person given only the information that she is Randerian. These marginals are analogous to those in Figure 1(b), but are computed by models trained on real-world passenger data.

Our first experiment addresses the Randeria problem schematically described in Figure 1. Our second experiment explores prediction of missing attributes when these hidden attributes were specifically chosen to create one-shot learning problems similar to the Randeria example. Both experiments rely on learning the structure of a Bayesian network, and we first present structure-learning results for the passenger data.

Learning Model Structure

Structure learning for the standard model can be achieved by drawing a MCMC sample from $P(G|D, \lambda)$, where each λ_i is set to 1. For the type-learning model we drew an MCMC sample from $P(G|D, \lambda)$ for each possible setting of λ. Given these samples, we constructed an approximate posterior $P(G, \lambda|D)$ by computing the relative posterior probabilities of each pair (G, λ) then normalizing.

Both models learned distributions on graph structures which capture some of the intuitive relationships between the seven attributes. For example, both models predict with high confidence that there is an edge between the birth country and nationality attributes. The structures assigned high probability by the type-learning model tend to have more edges than the structures preferred by the standard model. Adding more edges allows the model to explain certain attributes as near-deterministic functions of their parents.

For any training set D, we use the above training technique to obtain structure distributions $P(G|D, \lambda)$ for the standard model and $P(G, \lambda|D)$ for the type-learning model. These distributions serve as the basis for predictions about unobserved attributes.

Meeting a Randerian

Our first test directly corresponds to the Randeria problem mentioned in the introduction. We took the passenger data already described and added a record for a single Randerian—an individual with blonde hair, a fair complexion, and blue eyes, but a new nationality, race, language and birth country. Using the training technique described in the previous section, the models infer structure distributions and network parameters. Both models were then asked to predict the language and hair color of a second individual that was known to be Randerian, but had no other attributes observed. Figure 4 shows the marginal distributions over language and hair-color for both models.

Only the type-learning model was able to confidently predict that a second Randerian would also speak Randerian based on the single training instance provided. When predicting hair color, both models produce similar distributions over the possible values. Despite allowing for near-deterministic relationships, the type-learning model correctly realizes that hair color is not a near-deterministic function of nationality.

One-Shot Learning Tests

Figure 4 suggests that the type-learning model matches our intuitive notion about correct performance on the Randeria problem, and our next analysis explores a setting where model success can be assessed more objectively. We took the passenger data and created a series of one-shot learning problems for each attribute value. For example, we create a one-shot learning problem for the case where Language=French by removing all French-speaking passengers except one from the training set. The test set contains all of the French speakers that were removed, and the task is to predict the language of each individual given all of their other attributes. In other words, we explore whether the models can confidently identify French speakers after observing a single example of this category. We repeated this process for each value of each attribute in the passenger data.

To evaluate the models we measure both model accuracy and model confidence. We expect that near-deterministic relations will allow confident predictions based on a single training instance, and use Kullback-Leibler(KL) divergence as a metric of model confidence. We considered the models' inferred marginals as approximating distributions to the true marginal, KL(true||inferred). The true marginal is a point-mass distribution which assigns all of its probability to the correct attribute value. In this case, the KL-divergence simplifies to $-\log[p(v_t)]$ where $p(v_t)$ is the probability a model assigns to the true attribute value.

Table 2 shows the results of the one-shot learning tests for both models. As expected, the type-learning model

Table 2: One-shot learning tests. Each model was shown a single instance with a given attribute value (e.g. a single French-speaking passenger) and asked to make inferences about all other instances with this attribute value.

Missing	**KL Divergence**		**Accuracy (%)**	
Attribute	**TL**	**Standard**	**TL**	**Standard**
Nationality	1.46	2.72	73	58
Race	1.74	2.16	63	36
Language	1.38	2.16	60	60
Country	1.23	2.32	82	45
Complexion	1.99	1.96	13	18
Hair	3.22	3.28	0	0
Eyes	3.26	3.33	0	0

made more confident inferences for attributes with near-deterministic relations given only a single training example. Given a single instance of a passenger from a new country, for example, the model achieves high accuracy and confidence (as measured by a low KL divergence) when predicting the country attribute for subsequent passengers from that country. In contrast, the standard model was often unable to make confident one-shot inferences. Although this model made inferences from the single target instance at a rate better than chance, it had substantially lower confidence and accuracy for attributes with near-deterministic relations. Both models performed comparably for the three non-deterministic attributes. We do not expect one-shot learning to be possible for these attributes, and accuracy was low in all cases.

Conclusion

Humans often make accurate inferences given a single example of a novel situation, and we presented a model that attempts to match this ability. Our model uses a Bayes net to capture relationships between attributes, and learns which of these relationships are soft and probabilistic and which are near-deterministic. The ability to exploit near-deterministic relationships gives our approach a different inductive bias than a standard Bayes net approach, and we showed that this inductive bias supports one-shot learning about novel situations.

Here we focused on a specific one-shot learning problem—the Randeria problem—that is motivated by real-world inferences made by human learners. Future studies can design behavioral experiments to test our approach, and can explore, for example, how people make inferences about unobserved entries in the passenger data that we analyzed. Future experimental studies can also explore one-shot learning in other settings. Kemp et al. (2007) describe a special case of our approach that helps to explain word-learning data collected by Smith et al. (2002), and our current approach should account for all of the findings captured by this previous model. This previous model, however, can only learn Bayesian networks that belong to a very restricted class. Future studies of one-shot learning can test our prediction that people can learn and reason about a much broader class of relationships.

Acknowledgments We thank Daniel Navarro and two anonymous reviewers for comments on the manuscript. Andrew Maas was supported by an Interdisciplinary Training Grant in Computational Neuroscience (NIH/NIDA R09 DA023428).

References

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and reports on child language development*, *15*, 17–29.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychological Review*, *112*(2).

Davies, T. R., & Russell, S. J. (1987). A logical approach to reasoning by analogy. In *IJCAI 10* (pp. 264–270).

Fei-Fei, L., Fergus, R., & Perona, P. (2003). A Bayesian approach to unsupervised one-shot learning of object categories. In *ICCV 9.*

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall.

Giudici, P., & Castelo, R. (2003). Improving Markov Chain Monte Carlo model search for data mining. *Machine Learning*, *50*, 127–158.

Giudici, P., & Green, P. (1999). Decomposable graphical Gaussian model determination. *Biometrika*, *86*, 785-801.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology.* Cambridge, MA: MIT Press.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1-31.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, *20*(3), 197–243.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*, 307–321.

Lucas, C., & Griffiths, T. (2007). Learning the functional form of causal relationships. In *Proceedings of the 29th annual conference of the cognitive science society* (p. 1810). Austin, TX: Cognitive Science Society.

Milch, B., Marthi, B., Russell, S., Sontag, D., Ong, D. L., & Kolobov, A. (2005). BLOG: Probabilistic models with unknown objects. In *IJCAI 19* (pp. 1352–1359).

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine Learning*, *62*, 107–136.

Russell, S. J. (1989). *The use of knowledge in analogy and induction.* London: Pitman.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: causal determinism and children's inferences about unobserved causes. *Child Development*, *77*(2), 427–442.

Silander, T., Kontkanen, P., & Myllymäki, P. (2007). On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In *UAI 23.*

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Steck, H. (2008). Learning the Bayesian network structure: Dirichlet prior vs data. In *UAI 24* (p. 511-518).