Inferring priors in compositional cognitive models

Eric J. Bigelow
Dept. of Brain & Cognitive Sciences,
University of Rochester
Rochester, NY 14627
ebigelow@u.rochester.edu

Steven T. Piantadosi
Dept. of Brain & Cognitive Sciences,
University of Rochester
Rochester, NY 14627
spiantadosi@bcs.rochester.edu

Abstract

We apply Bayesian data analysis to a structured cognitive model in order to determine the priors that support human generalizations in a simple concept learning task. We modeled 250,000 ratings in a "number game" experiment where subjects took examples of a numbers produced by a program (e.g. 4, 16, 32) and rated how likely other numbers (e.g. 8 vs. 9) would be to be generated. This paper develops a data analysis technique for a family of compositional "Language of Thought" (LOT) models which permits discovery of subjects' prior probability of mental operations (e.g. addition, multiplication, etc.) in this domain. Our results reveal high correlations between model mean predictions and subject generalizations, but with some qualitative mismatch for a strongly compositional prior.

Keywords: Concepts and categories; learning; Bayesian modeling; machine learning

Introduction

Structured "Language of Thought" (LOT) models have recently become popular cognitive theories across a wide variety of domains (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Katz, Goodman, Kersting, Kemp, & Tenenbaum, 2008; Kemp, Goodman, & Tenenbaum, 2008; Kemp, 2012; Ullman, Goodman, & Tenenbaum, 2012; S. Piantadosi, Tenenbaum, & Goodman, 2012; S. Piantadosi, Goodman, & Tenenbaum, under revision). In most of these accounts, learners are assumed to generate compositionally structured hypotheses in order to explain observed data. For instance, in Goodman et al. (2008), learners infer compositions of boolean operations and featural primitives (e.g. $RED \vee (DOTTED \wedge SMALL)$) to explain observed data, a model that explains several key phenomena in rule learning as the consequence of Bayesian rule induction. Other work models number word learning as the discovery of a counting algorithm, capturing children's developmental progression as a consequence of inferring the correct composition of operations to perform on sets (S. Piantadosi et al., 2012).

These types of LOT models typically assume fixed priors on hypotheses, which in turn provide an inductive bias for learners to prefer "simpler" compositions of primitives, consistent with behavioral tendencies (Feldman, 2000, 2003).

Of course, these priors amount to substantial assumptions about people's expectations at the start of rule learning. To test these assumptions, different particular LOTs have been tested to compare, for instance, LOT theories with distinct types of quantification or varying sets of boolean operations (Kemp, 2009, 2012; S. Piantadosi, 2011; S. T. Piantadosi, Tenenbaum, & Goodman, under review). Here, we develop a method for directly inferring the parameters of an LOT prior from behavioral data, much in the spirit of work recovering priors from behavioral data in psychophysics (Stocker & Simoncelli, 2006; Paninski, 2005). We provide a freely modifiable implementation in Python (S. T. Piantadosi, 2014) for further use and extension.

We assume that the prior parameters specify a generative model, namely a Probabilistic Context Free Grammar (PCFG). For instance, in the context of logic, we might have separate PCFG parameters corresponding to the production of a rule with disjunction ($\vee$) versus conjunction ($\wedge$). These parameters determine the relative likelihood of each operation; by inferring their values, we are able to determine how strongly subjects believe that each will be used in a novel, unobserved concept.

Our analysis technique relies on Bayesian tools, allowing us to infer both the likely parameters and the likely ranges of parameters from subjects' data. This allows us to determine exactly how much behavioral data tells us about the prior; we might discover that the behavioral data is not informative about subjects' priors, resulting in high variance in the posterior on PCFG parameters. Alternatively, we might discover that the prior probability of some but not all operations can be recovered from the data. This type of statistical inference permits inferences that are "just right" from subjects' data, indicating what a scientist should believe about otherwise unobservable cognitive operations.

The structure of this paper is as follows. First, we introduce the **Number Game**, an induction task providing a simple domain of concept-learning. Then, we discuss the structure and expressive potential of probabilistic context-free grammars as a representation for concept hypotheses. After this we present our method for Bayesian data analysis of grammar parameters and apply it to three complementary LOT formalizations.

The Number Game

We consider concept learning in the Number Game (Tenenbaum, 1999, 2000), a simple domain of cognitively-interesting induction. In the number game, concepts correspond to subsets of integers from the domain $\{1, ..., 100\}$. Subjects observe some numbers D (data) in an unobserved concept C and are asked which other numbers are in C. For instance, given observed data $D = \{16, 2, 64, 8\}$, subjects might induce that the concept used to generate these was "powers of two". This an interesting domain because this problem is under-determined, meaning that there are many solutions ("all numbers" and "even numbers" are both consistent with the data, for instance). Despite this, subjects often have strong

intuitions that some concepts are more likely.

Tenenbaum (Tenenbaum, 1999) captures these intuitions using a Bayesian model which computes $P(C \mid D)$ according to Bayes rule, or $P(C \mid D) \propto P(C)P(D \mid C)$. Here, $P(C)$ is a prior on concepts representing subjects' beliefs about likely concepts before D is observed. $P(D \mid C)$ is a likelihood model of how likely D would be if C were the true concept. This work makes a *strong sampling assumption* that the elements of D are chosen by sampling uniformly from the set specified by C. Thus

$$P(D \mid C) = \left(\frac{1}{|C|}\right)^{|D|} \quad (1)$$

This explains why, for instance, $D = \{16, 2, 64, 8\}$ suggests "powers of two" rather than "even numbers". If C is "powers of two", this D would be chosen out of the set $C = \{2, 4, 8, 16, 32, 64\}$, giving a likelihood of $(1/6)^4$. If C were $\{2, 4, 6, \ldots, 100\}$, then the likelihood would be much less, $(1/50)^4$. This critical assumption that the likelihood of data depends on the cardinality of C is known as the **size principle** and is a natural consequence of the strong sampling assumption (Tenenbaum, 1999).

Given D, inferences about whether a new number x is in the concept are made by integrating across all possible concepts:

$$\sum_C P(x \mid C)P(C \mid D) = \sum_C P(x \mid C)\frac{P(D \mid C)P(C)}{P(D)} \quad (2)$$

Here, we estimate $P(D)$ by summing over a large number of concepts, representing the vast majority of posterior probability mass (see below). In our implementation, both $P(x \mid C)$ and $P(D \mid C)$ also include a noise parameter $\alpha = 0.90$ that generates elements of C 90% of the time, and elements from $\{1, \ldots, 100\}$ uniformly 10% of the time.

The Language of Thought in the Number Game

In a formalization of learning as inductive inference over a LOT representation language, a probabilistic context-free grammar (PCFG) can be used to model concepts as compositions of simple primitives. For our purposes, each tree generated by a LOT PCFG represents a concept hypothesis C. Generation probability for a tree gives the concept's prior probability $p(C)$, calculated as $p(C) = \prod_k \lambda_k$, where λ_k is the probability of the k'th rule used to generate the tree. As in all PCFGs, this probability is conditioned on the parent of the generated node. This prior allows us to implicitly specify an infinite concept space and assign higher probability to more concise LOT expressions.

The assumed PCFG represents a hypothesis about what mental representations might be like, as well as what types of concepts people intuitively find probable. By inferring the $\{\lambda_k\}$ from data, we are assuming part of the representation (the structure of the PCFG) and discovering part (the specific probabilities) from human data. Note that we may find that some λ_k are close to zero, meaning that we have assumed rules which are not psychologically justifiable. Additionally, we may write down different PCFGs and compare their performance in explaining human behavior. This allows us to deduce some further lessons about the structure of the PCFG over and above fitting probabilities. In particular, we consider three grammars capable of modeling number game concepts (see Table 1). Next, we describe these models in more detail. We will consider as an example the prior associated with the "odd numbers" concept, for each of the three grammars.

Independent Model Grammar	**Compositional Model Grammar**
$Start \xrightarrow{\lambda_0} Math$	$Start \longrightarrow Set$
$Start \xrightarrow{1-\lambda_0} Interval$	$Set \xrightarrow{\lambda_0} Set \cup Set$
	$Set \xrightarrow{\lambda_1} Set \cap Set$
	$Set \xrightarrow{\lambda_2} Set \setminus Set$
	$Set \xrightarrow{\lambda_3} Math$
	$Set \xrightarrow{\lambda_4} Interval$
$Math \xrightarrow{\lambda_{1-9}}$ Powers of $n \quad 2 \le n \le 10$	$Math \longrightarrow Map(\lambda x.Expr, Interval)$ [1]
$Math \xrightarrow{\lambda_{10-19}}$ Multiples of $n \quad 3 \le n \le 12$	$Expr \xrightarrow{\lambda_5} Expr \cdot Expr$
$Math \xrightarrow{\lambda_{20-29}}$ Ends with $n \quad 0 \le n \le 9$	$Expr \xrightarrow{\lambda_6}$ Ends with$(Expr, Expr)$
$Math \xrightarrow{\lambda_{30-39}}$ Contains Digit $n \quad 0 \le n \le 9$	$Expr \xrightarrow{\lambda_7}$ Contains Digit $(Expr, Expr)$
$Math \xrightarrow{\lambda_{40}}$ Prime numbers	$Expr \xrightarrow{\lambda_8}$ Prime$(Expr)$
$Math \xrightarrow{\lambda_{41}}$ Even Numbers	$Expr \xrightarrow{\lambda_9} Expr^{Expr}$
$Math \xrightarrow{\lambda_{42}}$ Odd Numbers	$Expr \xrightarrow{\lambda_{10}} Expr + Expr$
$Math \xrightarrow{\lambda_{43}}$ Squares	$Expr \xrightarrow{\lambda_{11}} x$
$Math \xrightarrow{\lambda_{44}}$ Cubes	$Expr \xrightarrow{\lambda_{12}} Const_E$
	$Const_E \xrightarrow{\lambda_{13-27}} n \quad 1 \le n \le 15$
$Interval \xrightarrow{1.0}$ Range$[m, n]: \ 1 \le m \le n \le 100$	$Interval \longrightarrow Range[Const_I, Const_I]$
	$Const_I \xrightarrow{1.0} n \quad 1 \le n \le 100$ [2]

Table 1: Three PCFG constructions: a Mixture Model (not shown), a simple model with independent probabilities for each rule ("Independent Model"), and a recursive compositional rule model ("Compositional Model"). The number above each arrow gives the *unnormalized* probability of each rule expansion. The Mixture Model Grammar is identical to the Independent with all parameters λ_k are fixed to 1.0 except λ_0.

Mixture Model

The **Mixture Model** implements a very simple PCFG which combines two types of grammatical productions: those generating **mathematical rules** and those generating ranges of numbers, called **interval-based** concepts. This setup follows Tenenbaum (Tenenbaum, 1999, 2000), who constructed a concept hypothesis space for the range of numbers 1 through 100 according to representative concepts generated by additive clustering within the domain of numbers 1 through 10 (Tenenbaum, 1996). The only distinctions between this concept hypothesis space and that of our model, are that our space includes two additional types of rule-based concepts: "Ends in n" and "Contains Digit n", and that where Tenenbaum assigns an Erlang prior across interval concepts, we assign a uniform prior for simplicity.

The Mixture Model grammar has a single parameter λ_0 which determines the relative probability of using a rule-based or a interval-based grammatical production. Once the type of the production is chosen, the specific production is

[1] While other λ here refer to parameters, this refers to the abstraction operator λ.

[2] In order to narrow the extremely broad hypothesis space of the compositional model, priors of 5.0 are assigned to constants $1, 10, 20, 30, \ldots, 90, 100$.

chosen uniformly as in (Tenenbaum, 1999, 2000). For example, to generate an "odd numbers" concept, we begin at *Start* and traverse to *Math* with probability λ_0. Following the allowed rules in Table 1, we expand *Math* to "Odd Numbers" with probability $1/44$, yielding an overall prior of $p(C) = \frac{\lambda_0}{44}$.

Independent Probabilities Model

The **Independent Probabilities grammar** generates the same set of concepts as the Mixture Model grammar, but with the key difference that each rule-based concept is produced by a rule with an independent probability parameter λ_k. This grammar also has a mixture parameter λ_0 that biases the grammar towards interval- or rule-based concepts. As in the Mixture Model, interval concepts are assigned a uniform prior. The parameter space for the Independent Model allows a more intricate representation of individuals' priors, where relative bias associated a priori with particular concepts can be inferred. For example, in the Independent Model we might infer that an individual has a stronger bias associated with the concept "multiples of 4" over the concept "multiples of 7".

To generate the "odd numbers" concept, the same rules are used to generate the concept as in the Mixture Model but with different associated priors: $p(C) = \frac{\lambda_{42}}{\sum_{k=1}^{44} \lambda_k}$.

Compositional Model

The **Compositional grammar** is a recursive LOT model that expresses concepts by freely composing primitives, based on the general approach of (Goodman et al., 2008). The Compositional Model has a significantly wider concept hypothesis space, and its parameter space reflects the probability of using each primitive operation. These primitives include values and operations that can compose to create a range of numerical concepts, including all those of the other two models, as well as set operations between mathematical expressions and intervals. For example, one might infer that humans have a particularly high bias associated with the "times" operator, and a very low prior with the "powers of n" operator. Note that by the assumed statistical structure of a PCFG, every place that "times" is used, it will have the same probability. Thus, despite having a richer concept space, the Compositional grammar cannot model some of the concept-specific priors of the Independent Model — e.g. the context free assumptions of the PCFG mean that 2 in $(2 \cdot x)$ and 2 in $(2 + x)$ both are equally likely (i.e. irrespective of their use in the context of addition or multiplication).

This grammar requires a considerably deeper tree to represent the "odd numbers" concept. From *Start* we go to *Set*, then *Math* with a probability λ_3, which maps an *Expr* across the range of integers 1 to 100. This *Expr* goes to $Expr + Expr$ with probability λ_{10}; one of the two *Expr* terminates at 1 with probability λ_{13}, and the other goes to $Expr \cdot Expr$. This final expression terminates with x and 2, with respective probabilities λ_{11} & λ_{14}. In plain English, "multiples of 2, plus 1, for integers one to one-hundred." While this may seem complicated, an advantage of this model is its capacity to capture a wide range of complexities with fewer free parameters. For example, the following three concepts are ordered by increasing complexity when generated by the Compositional grammar: "powers of 2", "powers of 2 plus 1", "only primes from the set of powers of 2 plus 1"

Bayesian Data Analysis

The primary contribution of this paper is to apply data analysis techniques to an interesting domain of inductive concept learning in cognitive science, assessing parameters of the prior, $\{\lambda_k\}$, given subjects' behavioral data. The prior is the most psychologically laden aspect of LOT models because it specifies people's assumptions without any data, as well as—in compositional models—their subjective expectations about how likely each operation is to be used. Similar data analysis has been used previously to infer the parameters of the prior's PCFG (S. Piantadosi, 2011; S. T. Piantadosi et al., under review), although largely with the goal of comparing different languages. Here, we focus on the complementary question of inferring in detail the relative expectations among different primitives, or different classes of primitives.

To do this, we present a method of Bayesian data analysis (Kruschke, 2010) to infer a posterior distribution on each rule probability, given the human responses. We'll use y_i to represent the number of "yes" counts for whether a query x_i (e.g. $x_i = 16$) is in the concept, given data D_i (e.g. $D_i = \{2,4,8\}$). We'll use N_i to denote the total number of human responses for query x_i, fixed by experimental design. Let $G = \{\lambda_k\}$ be the set of rule production probabilities. For a given G, we implicitly define a space of concept hypotheses that may be generated by this grammar. We wish to find

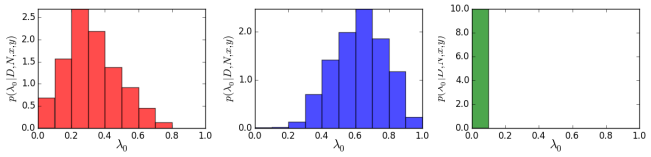
$$\begin{aligned} P(G \mid y, D, N, x) &\propto P(G) P(y, \mid G, D, N, x) \\ &= P(G) \prod_i P(y_i, \mid G, D_i, N_i, x_i). \end{aligned} \tag{3}$$

Where the product comes from assumed independence of responses, conditioned on D_i and x_i. In this equation, $P(G)$ is a prior on parameters. We assign a $Gamma(2,1)$ prior to each λ_k, representing its *unnormalized* probability. Human responses are then assumed to come from a binomial likelihood $P(y_i, \mid G, D_i, N_i, x_i)$,

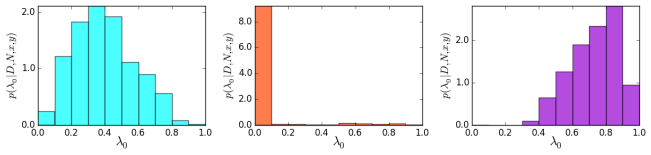
$$\binom{N_i}{y_i} [P(y_i, \mid G, D_i, N_i, x_i)]^{y_i} [1 - P(y_i, \mid G, D_i, N_i, x_i)]^{N_i - y_i} \tag{4}$$

Together, these equations specify a data analysis model over grammar probabilities that can take human inferences (e.g. 9 out of 10 people believe 16 is in the concept $\{4,8\}$ came from), and work backwards to discover the relative probability of compositional PCFG components behind these.

For the Compositional Model, we use a Markov Chain Monte Carlo sampling technique, the Metropolis-Hastings (MH) algorithm, to generate a representative subset of the infinite concept hypothesis space when computing the marginal (integral) (2). In computing the marginal, we ran 255 independent MH chains at 300,000 steps each, with each chain conditioned according to a unique concept example D from our data. All chains were initialized with the same uniform



(a) 15 in concept? $\langle 3\ yes, 3\ no \rangle$ (b) 15 in concept? $\langle 3\ yes, 30\ no \rangle$ (c) 15 in concept? $\langle 30\ yes, 3\ no \rangle$



(d) 64 in concept? $\langle 3\ yes, 3\ no \rangle$ (e) 64 in concept? $\langle 3\ yes, 30\ no \rangle$ (f) 64 in concept? $\langle 30\ yes, 3\ no \rangle$

Figure 1: Posterior distribution of parameter space for the Mixture Model grammar $p(\lambda_0 \mid D, x, y, N)$. $D = \{16\}$ for each example here, and x, y, N are each single-item sets, for either $x = \{15\}$ (top row) or $x = \{64\}$ (bottom row).

prior G_0 (approximately uniform, with higher probability assigned to terminal rules). The 1000 hypotheses with highest posterior scores for each chain (thus containing nearly all posterior probability mass) were all joined by union, yielding a total set size of 9,946. This formed a finite hypothesis space that well-approximated each posterior ($P(C \mid D)$).

MH was also used to infer the grammar parameters G in 3 for all models. Ten chains each were run for the Compositional and Independent Models, each for 50,000 steps, with 50,000 steps of burn-in; convergence was usually observed for these after less than 20,000 steps. Convergence in the Mixture Model (which has only one parameter) was usually reached after less than 500 samples were drawn, so four chains for this model were each run for 10,000 steps.

An example statistical inference

To demonstrate our approach, we first apply this data analysis to a simple toy example. Imagine you see a single example from a concept: $\{16\}$ and are asked whether 15 or 64 are part of the same concept. 15 is close to 16 in magnitude, the metric of "similarity" considered by Tenenbaum. On the other hand, 64 can be grouped with 16 according to numerical rules—for example, "even numbers", "powers of two", "powers of 4", or "perfect squares" are all candidate concepts.

We ran this simulated data set on the Mixture Model, shown in Table 1. We then used our data analysis to infer this model's single λ_0 parameter from a few data sets with intuitive answers, serving as a proof of principle for our data analysis methods. For instance, if subjects assume 15 is in the concept (given only that they know 16 is), we should infer that λ_0 is low, meaning that similarity (distance) based concepts are the most psychologically salient ones. On the other hand, if our simulated data has that subjects believe 64 is likely instead, we should see a *high* λ_0, indicating that the psychologically highest prior concepts are rule-like.

Figure 1 shows the posterior distribution on the λ_0 for six artificial datasets. Figure 1a shows the distribution on λ_0 if we find 3 subjects saying "yes" and 3 saying "no" to 15, conditioned on $\{16\}$ being in the concept. In this case, λ_0 is biased towards 0, representing a preference for interval-based concepts. A bias emerges here because mathematical concepts have trouble explaining this data, so little is gained by assigning them a high prior. However, if most subjects say "no" to 15 (Figure 1b), the model correctly implies that λ_0 must be higher: rejecting 15 means that people likely down-weight the prior probability of ranges, many of which include 15.

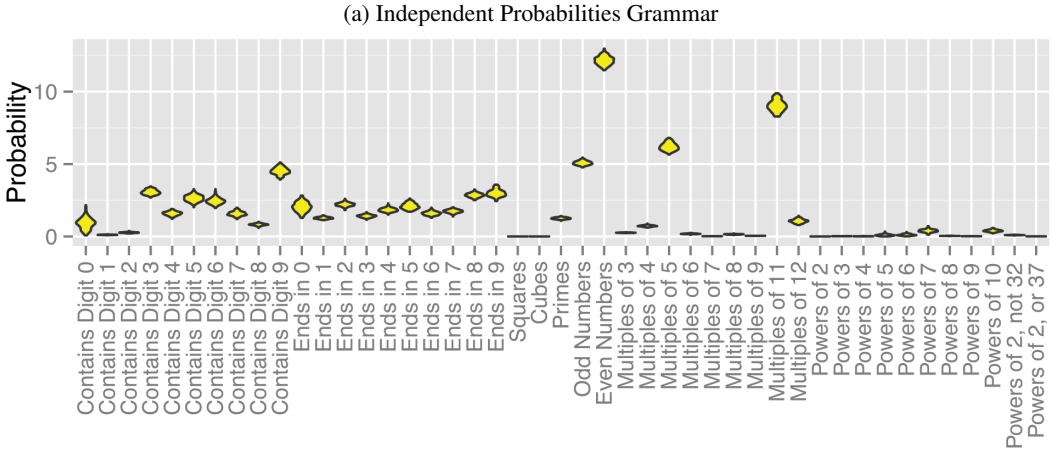
Therefore, more "yes"es than "no"s will indicate that interval-based concepts probably have a high prior bias. We see the posterior distribution in Figure 1c reflects this expectation, since the distribution over mixture ratios is greatly skewed towards interval-based concepts (low λ_0). We also see this skew towards interval-based concepts in Figure 1a, but with a more uniform distribution. With very few data points, it is unclear whether these responses are truly reflective of the prior, or simply result from noise. With a more proportional mixture ratio, we should see fewer "yes"es than "no"s due to an initial preference towards rule-based concepts, as reflected in the distribution of Figure 1b.

Imagine you are instead asked whether 64 is in the target concept given $D = \{16\}$. Maybe the pattern is "square numbers" or "powers of 2" - each of these has a small set size in the domain of 1 to 100, so it is likely that 16 would be output if one of these was the concept. Since "rule-based" concepts have such small set sizes, these are normally preferred over interval concepts given very little data (Tenenbaum, 1999). We therefore expect that with a proportional bias between interval and rule concepts, a majority of observers should assume 64 is in the target concept given $D = \{16\}$, as seen in Figure 1f. This will not be the case only when interval concepts are given a strong bias over rule concepts (Figure 1e).

Analysis of a large-scale Number Game experiment

We analyzed a large dataset of human responses for our experiment, comprised of 606 participants with demographics typical of Amazon Mechanical Turk workers (Bigelow & Piantadosi, 2016). Subjects were tested on input sets D that were generated by using a small set of "primordial" sets (all integers, evens, odds, squares, cubes, and primes) and then mapping a variety of functions (e.g. $f(n) = n+1$, $f(n) = n+2$, $f(n) = n-1$, $f(n) = 2 \cdot n$, etc.) across each. Data sets D were generated by randomly sampling sets of length 2, 3, and 4 from each resulting concept. On each trial, participants were told a set D was generated by a specific program, then indicated whether it was likely the program would generate each target from a random sequence of 30 numbers in the range $\{1, \ldots, 100\}$. No feedback was given as to the correctness of responses. Responses were trimmed for the 31 subjects with the lowest total typicality values - measured as $log(p)$ for "yes" responses and $log(1-p)$ for "no", where p is the fraction of positive responses - yielding a total 258,750 generalizations across 255 unique concepts.

Figure 2 shows an overall model summary of our results, giving the relationship between the model's predictive distri-



bution (x-axis) and humans' probability of responding "yes" (y-axis), to each of the generalizations in the data. The left column shows the initial grammar parameters — i.e., before $\{\lambda_k\}$ has been fit to human data; the right column shows the MAP grammars. This plot makes clear several aspects of our data. First, the correlations are high in the left column, meaning that the default model fits the human data relatively well. However, the model fit does improve in the right column, indicating that there is some variance to be gained by adjusting these parameters. Note that in all subplots the reported R^2 valued are on the box plot means, meaning that they do *not* represent the full variance in the data. The box plot also shows that there is considerable variance in responses for each bin, meaning that there is a substantial variability over and above the means that the model does not capture.

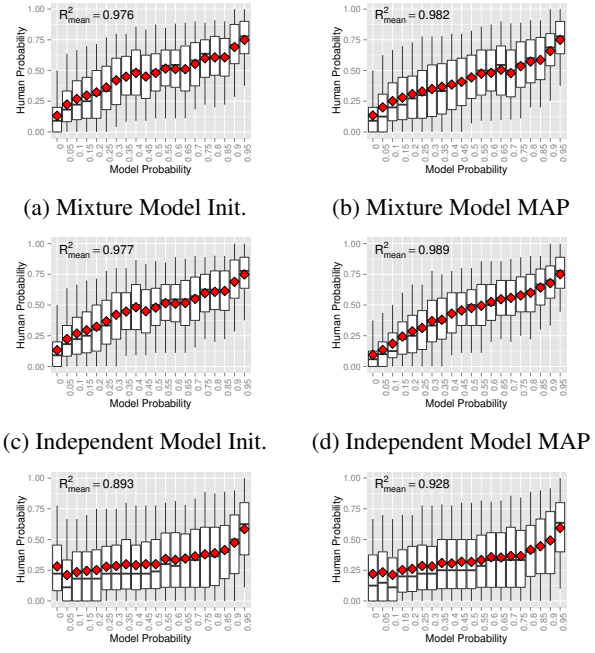


Figure 2: Box plots showing the distribution of human responses (y) binned by model predictions (x). Model predictions are shown for initial (a,c,e) and MAP (b,d,f) grammar priors. Red diamonds show mean values; box plots show 25^{th}, 50^{th}, & 75^{th} percentiles, and whiskers show 5^{th} and 95^{th}. Correlations were computed on mean values; box plots show high variance over and above these.

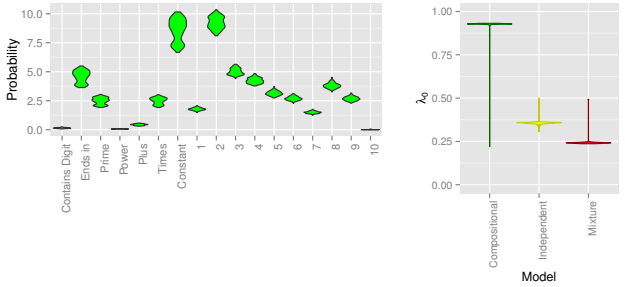


Figure 3: Posterior distribution of unnormalized grammar production probabilities λ_k. Bar width in this graph represents sample density relative to a specific parameter λ_k. Note that in the Independent Model, all rules shown are normalized relative to each other. In the Compositional Model, constant primitives (1-10) are weighted relative to other constants, and operators ("Ends in", "Times", "Constant") are weighted relative to other operators (see Table 1).

The overall shape of these plots tells us whether the model makes qualitatively correct predictions. The Mixture Model does well in this respect, as does the Independent Model, although the Independent Model must be interpreted with care since it freely fits the prior on a large number of concepts. The initially good fits of the Mixture and Independent Models indicate that an approximately uniform prior across these models' hypothesis spaces fits the data well; clearly, inferring MAP G will improve fit for some models more than others.

The Compositional Model seems to miss the qualitative match, which means falling far from the line $y = x$ even though the general increasing trend yields a respectable correlation. The qualitative mismatch is particularly salient for model predictions < 0.5, meaning the Compositional Model seems to systematically mischaracterize people's likely "no" responses. This may be because the form of the prior is highly constrained in the Compositional Model to follow the PCFG. For example, assigning a high prior to the concept "odd numbers" $(2n+1)$ requires high priors on its four components "times", "plus", "2", and "1". But high priors on these would also assign high prior to $1n+2$ ("numbers greater than two") due to the independence assumptions of a PCFG. It is possible that our Compositional Model performs poorly because we have included the wrong primitives or wrong structure in our particular PCFG. Fits such as these may in principle be used to quantitatively compare hypothesized LOTs.

These results highlight the impact of grammar design on determining the hypothesis space for models, and we expect this to be an area of future work. Another area of future work may be to explore the limitations of PCFGs. It may be that using a context-sensitive grammar will be necessary to fully capture human data, or models of another family entirely.

The most interesting aspect of our analysis is the posterior distribution on grammar productions for each model, giving what we should believe about people's relative probability of each kind of operation. Figure 3 shows these posterior distributions. The independent probabilities model shows that human data is best fit when the highest priors are associated with "Even Numbers", "Multiples of 11" (numbers with repeated digits), "Multiples of 5", "Odd numbers." Interestingly, nonzero weight is also assigned to "Primes", "Multiples of 12", and "Multiples of 4".

In the Compositional Model, human data is best fit by high priors associated with the "Ends In", "Prime", and "Multiples" ($a \cdot b$) operations. The highest prior among operations is associated with "Constant", which is to be expected as this allows for smaller hypothesis trees. Among the constants, the highest prior is associated with 2, followed by a general decreasing trend for large numbers that is reminiscent of the frequency distribution of numbers in language (Dehaene & Mehler, 1992; S. T. Piantadosi, in press).

Figure 3c shows that λ_0 is much higher for the Compositional Model than the others, as to be expected since many of its hypotheses are redundant with interval concepts (e.g. $n+2$ corresponds to "numbers greater than 2", equivalent to the interval $[3,100]$). The Independent Model's λ_0 is higher than the Mixture Model's, which is also expected as λ_k for each rule-based concept was conditioned on the same data.

Conclusion

Our work has shown how a structured inductive LOT model may be combined with a Bayesian data analysis to infer likely parameters of human subject's priors. Our analysis has revealed both an ability to freely infer priors on concepts (the Independent Model) as well as those that use the PCFG more productively, decomposing a concept's prior into a product of the priors of its parts (the Compositional Model). In these cases, Figure 3 represents our inference of these values from subject data. While the mean predictions of the Compositional Model are highly correlated with human means, the qualitative fit is worse. However, the method developed here of analyzing large-scale experiments with Bayesian data analysis provides a way forward for fitting, refining, and comparing LOT models of human cognition.

References

Bigelow, E., & Piantadosi, S. (2016). A large dataset of generalization patterns in the number game. *Journal of Open Psychology Data*, *4*(1).

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, *12*(6), 227.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science*, *32*(1), 108–154.

Katz, Y., Goodman, N., Kersting, K., Kemp, C., & Tenenbaum, J. (2008). Modeling semantic cognition as logical dimensionality reduction. In *Proceedings of Thirtieth Annual Meeting of the Cognitive Science Society.*

Kemp, C. (2009). Quantification and the language of thought. *Advances in neural information processing systems*, *22*.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, *119*, 685–722.

Kemp, C., Goodman, N., & Tenenbaum, J. (2008). Learning and using relational theories. *Advances in neural information processing systems*, *20*, 753–760.

Kruschke, J. (2010). *Doing bayesian data analysis: A tutorial introduction with r*. Academic Press.

Paninski, L. (2005). Nonparametric inference of prior probabilities from bayes-optimal behavior. In *Advances in neural information processing systems* (pp. 1067–1074).

Piantadosi, S. (2011). *Learning and the language of thought.* Unpublished doctoral dissertation, MIT.

Piantadosi, S., Goodman, N., & Tenenbaum, J. (under revision). Modeling the acquisition of quantifier semantics: a case study in function word learnability.

Piantadosi, S., Tenenbaum, J., & Goodman, N. (2012). Bootstrapping in a language of thought: a formal model of numerical concept learning. *Cognition*, *123*, 199–217.

Piantadosi, S. T. (2014). *LOTlib: Learning and inference in the language of thought.* available from https://github.com/piantado/LOTlib.

Piantadosi, S. T. (in press). A rational analysis of the approximate number system. *Psychonomic Bulletin and Review*.

Piantadosi, S. T., Tenenbaum, J., & Goodman, N. (under review). The logical primitives of thought: Empirical foundations for compositional cognitive models.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, *9*(4), 578–585.

Tenenbaum, J. B. (1996). Learning the structure of similarity. *Advances in neural information processing systems*, 3–9.

Tenenbaum, J. B. (1999). *A bayesian framework for concept learning*. Doctoral dissertation, Massachusetts Institute of Technology.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. *Advances in neural information processing systems*, *12*, 59–65.

Ullman, T., Goodman, N., & Tenenbaum, J. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*.