Learning Additive and Substitutive Features

Ting Qian (ting_qian@brown.edu)
Joseph Austerweil (joseph_austerweil@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences, 190 Thayer Street
Providence, RI 02912 USA

Abstract

To adapt in an ever-changing world, people infer what basic units should be used to form concepts and guide generalizations. While recent computational models of human representation learning have successfully predicted how people discover features from high-dimensional input in a number of domains (Austerweil & Griffiths, 2013), the learned features are assumed to be additive. However, this assumption is not always true in the real world. Sometimes a basic unit is substitutive (Garner, 1978), which means it can only be one value out of a set of discrete values. For example, a cat is either furry or hairless, but not both. In this paper, we explore how people form representations for substitutive features, and what computational principles guide such behavior. In a behavioral experiment, we show that not only are people capable of forming substitutive feature representations, but they also infer whether a feature should be additive or substitutive depending on the observed input. This learning behavior is predicted by our novel extension to the Austerweil and Griffiths (2011, 2013)'s feature construction framework, but not their original model. Our work contributes to the continuing effort to understand how people form representations of the world.

Keywords: learning; additive features; substitutive features; Bayesian nonparametric modeling; feature learning

Introduction

People have the remarkable capability of forming concepts that enable them to generalize beyond what they have encountered so as to guide their behavior. To form these concepts, one must deal with uncertainty, not only of what objects are present, but also of what the basic units of objects are – or "features" – that represent the objects. Most theoretical frameworks of concept learning treat these basic units as immediately available to learners. However, there are an infinite array of properties that could be used as features to encode objects (Goodman, 1972), raising the question of whether people infer these basic units from their observations of the world, and if so, how. Recently, Austerweil and Griffiths (2011, 2013) presented a computational framework for explaining how people construct feature representations from raw sensory input. Their framework captures several theoretical aspects of human feature construction (e.g., arbitrary number of features and context sensitivity), as well as empirical studies of human feature construction. Their findings synthesize and complement previous research in the literature that shows people are able to infer features of objects from their environment (Schyns, Goldstone, & Thibaut, 1998).

Computational models of feature learning typically assume that features are independent and additive (Austerweil & Griffiths, 2013; Goldstone, Greganov, Landy, & Roberts, 2008). That is, given a set of features inferred from objects of the same concept, a novel object exhibiting a combination of those features should also be an instance of that concept as well. For instance, if we are given a group of cats and infer "having whiskers", "making meow sounds", "furry", and "hairless" as the features for the concept "cat", then a novel animal which both meows and looks furry is most likely a cat. However, this additive assumption can be problematic when features are substitutive (Garner, 1978). For example, a cat is either furry or hairless, but it cannot be both furry and hairless – they are two values of a "hair" feature. When learning features from raw sensory input, people are not told whether a feature is additive or substitutive, but must infer this while constructing features. How do people infer whether a newly constructed feature is additive or substitutive?

Previous work has identified psychological consequences of features being additive or substitutive (Garner, 1978; Gati & Tversky, 1982; Kemp, 2012). For example, Kemp (2012) found that some categories are easier to learn when they are defined as substitutive rather than additive features. In these studies, participants knew whether a feature was additive or substitutive based on prior knowledge. However, how do people learn whether a newly constructed feature is additive or substitutive in the first place (in which case, it might become prior knowledge in the future)? Building on the work of Austerweil and Griffiths (2011, 2013), we present a novel computational model for capturing how people construct features from raw sensory input, while learning whether those features *should* be additive or substitutive. The new model predicts a bias towards learning substitutive features when parts of objects are negatively correlated in the input, and we find support for this tendency in a behavioral experiment.

The outline of the paper is as follows. First, we review Austerweil and Griffiths (2013)'s feature construction framework. Next, we develop a novel Bayesian nonparametric model that constructs features while learning whether each of those features should be substitutive or additive. Then, we present a behavioral experiment testing a prediction of the proposed model and demonstrate that it explains human behavior better than the original model from Austerweil and Griffiths (2011). Finally, we discuss the implications of our results and some directions for future research.

Modeling Feature Learning

Inferring latent features in binary images

Viewing feature learning as Bayesian nonparametric inference is one proposed explanation of how people discover the features of objects (Austerweil & Griffiths, 2013). For the particular problem of feature learning with binary images, the Indian Buffet Process (IBP; Griffiths & Ghahramani, 2011) with a noisy-or likelihood function (the IBP

noisy-or model; Wood, Griffiths, & Ghahramani, 2006) is typically used (Austerweil & Griffiths, 2013). According to this model, the learning problem is formalized as finding the most probable assignment of features to objects $\mathbf{Z}$ and feature images $\mathbf{Y}$ given the raw sensory input of a set of objects $\mathbf{X}$. $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are all defined to be binary matrices (see Figure 1). $\mathbf{X}$ is the data matrix, where each row corresponds to the image of an object, and each column contains the pixel values at that location for all objects. So, $X_{nd} = 1$, indicates that the dth pixel of the nth object is "turned on" (i.e., it is non-blank). $\mathbf{Y}$ is the feature image matrix where each row is the image of its corresponding feature, and each column indexes the pixel locations. So, $Y_{kd} = 1$, indicates that the dth pixel of an object should be "on" if that object has feature k (subject to the noise introduced by the model; more details below). Finally, $\mathbf{Z}$ is the feature ownership matrix, where each row corresponds to an object, and each column corresponds to a feature. So, $Z_{nk} = 1$, indicates that object n has feature k, which in turn suggests that the pixels that feature k can turn on are likely to be present in object n.

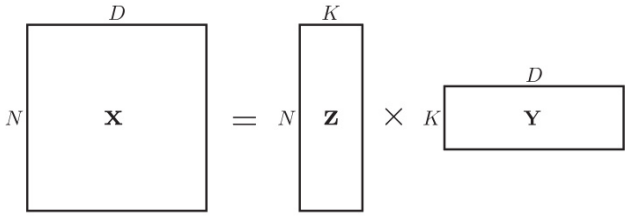


Figure 1: A schematic illustration of the $\mathbf{Z}, \mathbf{Y}, \mathbf{X}$ matrices in the IBP noisy-or model. Figure reprinted from Austerweil and Griffiths (2011) with permission.

One challenge in learning a latent feature representation under this model is that one does not know *a priori* the number of latent features that best accounts for the collection of objects $\mathbf{X}$. Instead, that number needs to be inferred from the data as well. The inference problem is also highly underconstrained because only $\mathbf{X}$ is observed, while both $\mathbf{Y}$ and $\mathbf{Z}$ need to be inferred from $\mathbf{X}$. In terms of Bayesian inference, this means that the joint posterior distribution of $\mathbf{Y}$ and $\mathbf{Z}$ needs to be estimated solely based on the observed data $\mathbf{X}$:

$$p(\mathbf{Y}, \mathbf{Z}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{Z}, \mathbf{Y})p(\mathbf{Y})p(\mathbf{Z}) \tag{1}$$

Using Bayes' rule, Equation (1) shows how the inference of $p(\mathbf{Y}, \mathbf{Z}|\mathbf{X})$ can be decomposed into two subtasks: to find the most likely $\mathbf{Y}$ and $\mathbf{Z}$ matrices, one should maximize the likelihood $p(\mathbf{X}|\mathbf{Z}, \mathbf{Y})$, corresponding to how well our feature representation captures the observations, and the prior probabilities p($\mathbf{Y}$) and p($\mathbf{Z}$). In the IBP noisy-or model, the prior distribution on $\mathbf{Z}$ is the IBP, which allows an infinite number of features to be inferred, but includes a penalty for overly complex representations. Details on how the IBP prior achieves this goal, including the culinary metaphor that provides the intuition of IBP can be found in Austerweil, Gershman, Tenenbaum, and Griffiths (2015). Here we describe the generative process definition. First, the probability of the nth object having a pre-existing feature k is proportional to number of objects that already have feature k (i.e., m_k in Equation 2), divided by the number of objects observed so far (i.e., n):

$$p(Z_{nk} = 1|\mathbf{Z}_{-nk}) \propto \frac{m_k}{n} \tag{2}$$

where $\mathbf{Z}_{-nk}$ is the feature assignments without Z_{nk}.

The nth object also can take on novel features as well. The probability that the nth has f novel features, which have not been observed in the first $n-1$ objects, is $p_{\text{poisson}}(f; \alpha/n)$. That is, this probability is evaluated as the chance of obtaining the sample f from a Poisson distribution with a mean of α/n. Here, α is a free parameter of the model, which we set to 1 for models throughout this paper. The IBP noisy-or model defines the prior distribution on $\mathbf{Y}$ by treating elements in $\mathbf{Y}$ as independent and identically distributed, each with a prior probability of θ to take the value 1 ($Y_{kd} \sim \texttt{Bernoulli}(\theta)$). The value of θ is set to 0.02 here, constraining the model to prefer $\mathbf{Y}$ with empty feature images (i.e., values set to 0) unless the feature images describe $\mathbf{X}$ well.

The likelihood function $p(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$ is defined with respect to the chance of generating the observed data $\mathbf{X}$ given a specific configuration of $\mathbf{Y}$ and $\mathbf{Z}$. In the IBP noisy-or model, the matrix product of $\mathbf{Z}$ and $\mathbf{Y}$ is first computed. This product is a matrix of the same dimensions as $\mathbf{X}$, whose elements can be interpreted as "weights" that indicate the total strength possessed by the current latent feature representation (i.e., $\mathbf{Z}$ and $\mathbf{Y}$) to turn on various pixels of observed objects. Intuitively, if the weights are large where elements in $\mathbf{X}$ are in fact 1, and the weights are small where elements in $\mathbf{X}$ are 0, then the corresponding $\mathbf{Z}$ and $\mathbf{Y}$ may be close to the optimal feature representation. Formally, the likelihood function is

$$p(\mathbf{X}|\mathbf{Y}, \mathbf{Z}) = \prod_{n,d} |x_{nd} - (1-\varepsilon)(1-\lambda)^{\mathbf{z_n y_d}}|, \tag{3}$$

where λ (set to 0.95) is the probability that a feature whose image has the current pixel on and is being used to represent the current object whose image succeeds to turn that pixel on in the current object, ε is the probability that a pixel in an image is on by chance (set to 0.05). As a result of jointly maximizing the likelihood function and the prior probabilities of $\mathbf{Y}$ and $\mathbf{Z}$, the IBP noisy-or model trades off keeping the feature representation as simple as possible with the model's ability to explain the data $\mathbf{X}$.

The additive nature of the IBP noisy-or model

Although the IBP noisy-or model has been shown to capture how the distribution of parts affects the feature representations people form (Austerweil & Griffiths, 2011), the features found by the model are always additive. For example, consider a scenario where half of the objects $\mathbf{X}$ have part A but not part B, the other half have part B but not part A, and no object has both part A and part B. The IBP noisy-or model

will correctly discover these features and, through the learning of the feature image matrix $\mathbf{Y}$, encode the information that both feature A and feature B are valid features for this group of objects. With this representation, the model will not only assign a high probability to any new object with either feature A or feature B, it will also regard a novel test object with both feature A and feature B as highly probable. The model considers feature A and feature B to be additive, because features are assumed to be independent of each other. The negative correlation between feature A and feature B is ignored by the model.[1]

A substitutive variation of the IBP noisy-or model

To add the capability of inferring substitutive features to the IBP noisy-or model, we propose a simple extension to the original model. Instead of there being only one feature image matrix $\mathbf{Y}$, the new model has two $\mathbf{Y}$ matrices: $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Therefore, a single feature has two alternative feature images (i.e., the two values the feature can take), each represented by the corresponding row vector in $\mathbf{Y}_1$ and that in $\mathbf{Y}_2$. Additionally, we used a new indexing matrix $\mathbf{F}$ that is the same size as $\mathbf{Z}$ and whose elements encode which image matrix should be used if an object takes a feature. The elements of $\mathbf{F}$ take on the value of 0 when the corresponding value in $\mathbf{Z}$ is also 0, and the value of 1 or 2 when the corresponding value in $\mathbf{Z}$ is 1. That is, for features that are present in an object, as indicated by $\mathbf{Z}$, the values in $\mathbf{F}$ indicate which of the two $\mathbf{Y}$ matrices (1 or 2) is its *value*.

The feature learning problem is then to infer $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{F}$ from the observed $\mathbf{X}$. Similar to Equation (1), we use Bayes' rule to decompose the posterior into simpler terms:

$$p(\mathbf{Y},\mathbf{Z},\mathbf{F}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{Y},\mathbf{Z},\mathbf{F})p(\mathbf{F}|\mathbf{Z})p(\mathbf{Y})p(\mathbf{Z}) \qquad (4)$$

The prior distribution on $\mathbf{Z}$ is the same IBP prior as in the original IBP noisy-or model. Conditioned on object n taking feature k ($Z_{nk} = 1$), F_{nk} is equally likely to be 1 or 2. If object n does not take feature k ($Z_{nk} = 0$), then $F_{nk} = 0$ with probability 1. So, its value for the infinite number of features that are not assigned to any object is 0. The prior on each $\mathbf{Y}$ is the same as in the original model. The calculation of the likelihood $p(\mathbf{X}|\mathbf{Y},\mathbf{Z},\mathbf{F})$ is also similar to the case of the original model, except that feature images are retrieved conditioned on the values of $\mathbf{F}$ and $\mathbf{Z}$ rather than on $\mathbf{Z}$ alone.

Crucially, this new model is capable of representing a substitutive feature, because the elements in the $\mathbf{F}$ matrix are either 1 or 2, but not both. For example, if a pair of parts, A and B, are negatively correlated across objects in the input (meaning that they almost never occur together), this new model will strongly favor a representation of a single feature with two alternative features images, one corresponding to Part A and the other corresponding to Part B. The model favors this representation over one with two additive features because the IBP prior favors feature representations with fewer features. Given this representation, test objects with either Part A or Part B will be regarded as highly probable, because the learned representation is exactly "one feature", in the form of A or B. Test objects with both parts will however be considered rather improbable, because the model lacks the necessary representation (which would be a two-feature representation) to account for those two parts simultaneously. Note that additive features can also be learned by the model when it is appropriate. This occurs when the pixels are all off for one of a feature's images. Thus, in some sense, it is a generalization of the original additive IBP noisy-or model.

The inherently additive IBP noisy-or model and our newly proposed model are two hypotheses for how people learn feature representations. The question of interest is, in what circumstances, if any, do people form substitutive representations of negatively correlated parts in objects? Our behavioral experiment aims to find an answer to this question.

Behavioral Experiment: Learning Additive or Substitutive Features in Vertical Bar Images

The goal of this experiment is to investigate whether people form additive or substitutive feature representations given 1) the co-occurrence patterns of parts within each image and 2) the way that parts are distributed across observed images. According to our model, the prediction is that people should prefer an additive representation for parts that occur independently. People are expected to prefer a substitutive feature representation for negatively correlated parts – that is, those that are never observed together in the same image, even when they have been observed separately in the set of all training images. Correspondingly, this experiment consists of two conditions - an additive condition and a substitutive condition - which test these two predictions respectively.

Methods

Participants Forty Amazon Mechanical Turk workers participated in this experiment (20 in each condition). Each was paid $0.20 for about 90 seconds of work.

Stimuli We designed artificial stimuli in the form of images containing vertical bars within a square box. Each box had 4 slots where vertical bars can appear. In both conditions, participants were exposed to a total of six stimuli (i.e., six square boxes with vertical bars in them). Four of the six stimuli were shared between the two conditions (see Figure 2). These four images are selected so that the locations of bars are counterbalanced and images with 1, 2, and 3 bars were all observed.

The crucial difference between the additive condition and the substitutive condition is the two additional images.[2] In the

[1] Existing connectionist feature learning methods (e.g., CPLUS; Goldstone et al., 2008) would also struggle learning substitutive features. One way to extend them to learn substitutive features would be to use a gating mechanism (Frank, Loughry, & O'Reilly, 2001).

[2] We also ran a baseline condition, which consisted of only the 4 training images shown in Figure 2. The results were identical to the substitutive condition reported here. This suggests people are biased towards the substitutive interpretation of features.

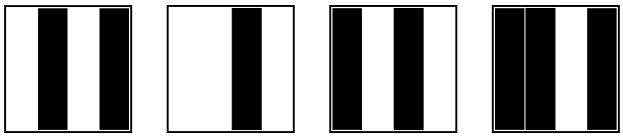
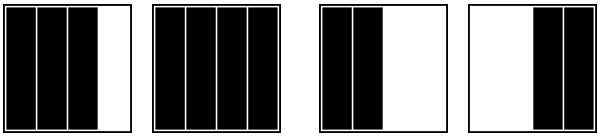


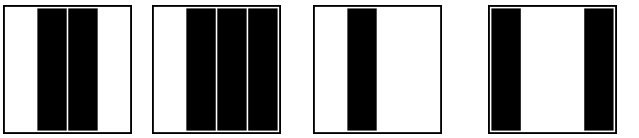
Figure 2: Four training stimuli were shared between the additive condition and the substitutive condition.



(a) Additive condition (b) Substitutive condition

Figure 3: Two additional training stimuli were different between conditions. In the substitutive condition, the second and third vertical bars were perfectly negatively correlated.

additive condition, as shown in Figure 3a, these additional images demonstrate that all vertical bars can co-occur in a stimulus, most evidently shown by the image where all four bars appeared together. In the substitutive condition, as shown in Figure 3b, these additional images are consistent with a pattern that is already in the four shared images: the second and third vertical bars are never observed together.



(a) co-occurring (b) single (c) non-occurring

Figure 4: Test stimuli were grouped into three types depending on the configuration of the second and third bars.

In both conditions, participants rated the likelihood of observing the four novel test stimuli while the training stimuli remained visible (more details in the procedure section below). Because the crucial difference between the additive and substitutive conditions is whether the second and third bars co-occurred in the training set, these test stimuli were grouped into 3 different groups defined by the arrangement of those two bars for evaluating participant ratings: a "co-occurring" groups where both the second and third bars are in a test stimulus, which included the two stimuli in Figure 4a; a "single-occurring" group where either the second or third bar is in a test stimulus (see 4b), and a "non-occurring" group where neither of the two bars is in a test stimulus (see 4c).

Procedure The procedure is identical in both conditions (they only differed in which two extra images were given). At the beginning of an experiment, participants were presented with the 6 training images appropriate to their conditions along with the following cover story:

> Recently a group of archaeologists found a cave with a collection of different images on its walls. The archaeologists believe the images could have been left by a prehistoric civilization. The images are shown below. Please take a few moments to investigate the images. You'll be asked questions about these images later.

Participants had to spend at least 30 seconds studying the training images before they were able to continue (although they could study the images for as long as they wanted). Afterwards, they were given the following test instructions:

> It looks like there are many more images on the cave wall that the archaeologists have not yet had a chance to record. If the archaeologists explored the cave wall further, which images do you think they would be likely to see?
>
> You will be presented with a few images, and your task is to rate how likely you believe it is that each image will be discovered in that cave, based on the images that you just studied.

Each test trial presented one test stimulus, and to minimize memory effects, training images were also shown alongside the test stimulus. For each test stimulus, participants were asked "How likely do you believe this image will be discovered in the cave?" and instructed to rate the likelihood using a scale ranging from 1 to 5. They were clearly instructed that 1 meant least likely and 5 meant most likely. Once participants committed to a rating, the experiment was programmed in such a way that they could not go back to previous test trials to modify their ratings.

Results and Discussion

Figure 5 shows the average ratings of participants for each group of test stimuli in both conditions. In the additive condition (see Figure 5a), participants rated the test stimuli from all three groups equally large ($F(2,77) = 1.60, p > 0.2$), regardless of whether the second and third bars were in the same image (the co-occurring group), only one of the two bars was in an image (the single group), or neither of the two bars was in an image (the non-occurring group). Crucially, there was no difference in the ratings between test co-occuring and single test stimuli (mean difference = -0.075, post-hoc Tukey test $p > 0.5$), indicating that participants treated the second and third bars as additive features, allowing them to appear in the same image. Participants did not distinguish between single and non-occurring stimuli either (mean difference = -0.5, post-hoc Tukey test $p > 0.3$).

In contrast, participants in the substitutive condition gave significantly different ratings to the test stimuli of the three different groups ($F(2,77) = 3.29, p < 0.05$; see Figure 5b). In particular, stimuli of the co-occurring group present received a much lower rating than stimuli of the single group (mean difference = -0.88, post-hoc Tukey test $p < 0.05$). This difference in rating suggests that participants formed a substitutive feature representation for the second and third bars –

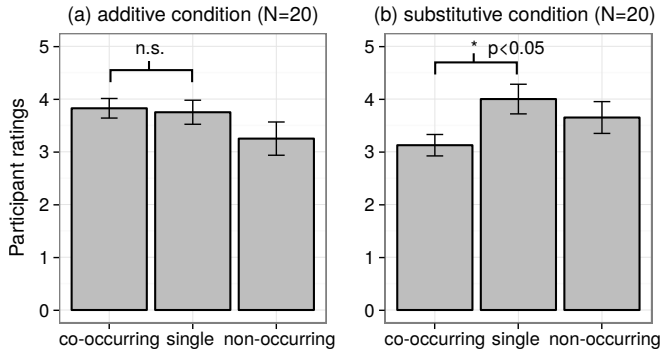


Figure 5: Subjects in the additive condition rated images with both the second and third bars as high as images without only one of those bars. Meanwhile, subjects in the substitutive condition gave images with both the second and third bars a significantly lower rating.

only one of them can be in an image at a time. No significant difference was found between the ratings of the co-occurring and non-occurring groups (mean difference = -0.53, post-hoc Tukey test $p > 0.3$), or between the single and non-occurring groups (mean difference = 0.35, $p > 0.5$).

Comparison to Model Predictions Overall, the results suggest that people can infer a substitutive feature representation for parts that are consistently negatively correlated in the input. They can also infer an additive feature representation when it is appropriate. Qualitatively, this contrasts with the prediction of the original IBP noisy-or model, which only forms additive representation of features regardless of the distribution of parts within and across objects. To examine the extent to which the original IBP noisy-or model and our proposed substitutive extension can describe human feature learning, we compared participant ratings to the predictions of these two models. To simplify the computational workload, the training images were downsampled to a resolution of 4 by 4 pixels without affecting the distributional information in these images. A Gibbs sampler was implemented for each model and run for 2000 iterations, where the 1001-2000th posterior samples were extracted as hypotheses of latent representations inferred by the models. We then searched among the 1000 hypotheses and extracted the one that maximized the generative probability of the test stimulus t:

$$p(t|\text{training stimuli}) = \max p(t|\text{inferred feature reps}) \quad (5)$$

where a "feature rep" is the inferred $\mathbf{Z}$ and $\mathbf{Y}$ matrices in the original IBP noisy-or model, and $\mathbf{Z}$, $\mathbf{F}$, and two $\mathbf{Y}$ matrices in our substitutive extension to the IBP noisy-or model. These probabilities were then normalized to a 1-5 scale so that the results are comparable to participant ratings.

Unsurprisingly, model predictions were more extreme and did not show as much variation as in participant ratings. However, we can still assess the qualitative similarity between the model ratings as shown in Figure 6 and participant ratings shown in Figure 5. As expected, the original IBP noisy-or model (Figure 6a) rated co-occurring, single, and non-occurring stimuli equally high regardless of whether the training condition was additive or substitutive. This is because the original IBP noisy-or model inferred four independent features, each of which represented a vertical bar at one of the four possible locations in a stimulus. Although such an additive representation correctly predicted the average rating behavior of participants in the additive condition ($R^2 = 0.99$), it failed to explain the lower ratings for the co-occurring stimuli in the substitutive condition ($R^2 = 0.44$).

The substitutive IBP noisy-or model, successfully predicted participant ratings in both conditions (Figure 6b). Given the substitutive images, it rated co-occurring stimuli much lower than the other two types of test stimuli ($R^2 = 0.83$), because it represented the negatively correlated parts (i.e., the second and third bars in the training images) as two alternative, but never co-occurring, values of a single feature. At the same time, it also predicted behavior relatively well in the additive condition ($R^2 = 0.86$). The R^2 value is slightly lower for the additive condition, which is most likely due to its built-in tendency towards substitutivity by having two $\mathbf{Y}$ matrices hardcoded in the model. However, it still qualitatively captured human performance in the additive condition. Thus, our newly proposed model described the overall pattern of human feature learning in our experiment better than the original IBP noisy-or model.

Conclusions and Future Directions

In this paper, we explored how people learn additive or substitutive features depending on the distribution of parts over objects using computational models and a behavioral experiment. In the experiment, one group of participants received images without negatively correlated parts (i.e., the additive condition), and the other group received images where two parts were never seen together (i.e., the substitutive condition). Our results demonstrated that the latter group readily treated the negative correlated parts as substitutive features: novel images with both those parts present were given a significantly lower rating than novel images with only one of those parts. However, what computational principles do people use to form substitutive feature representations? To answer this question, we proposed an extension to the original IBP noisy-or model (Austerweil & Griffiths, 2011), which allows a single feature to be realized in two alternative, but never co-occurring forms. The resulting substitutive IBP noisy-or model successfully captured the rating pattern of participants in the substitutive condition, improving upon the original IBP noisy-or model that failed to do so.

Building upon the behavioral paradigm and computational framework introduced in this paper, we plan to further investigate the issue of substitutive representation in human feature learning. First, we will test the generality of the results by using more natural perceptual and conceptual stimuli. This should also help address the potential confound of there be-

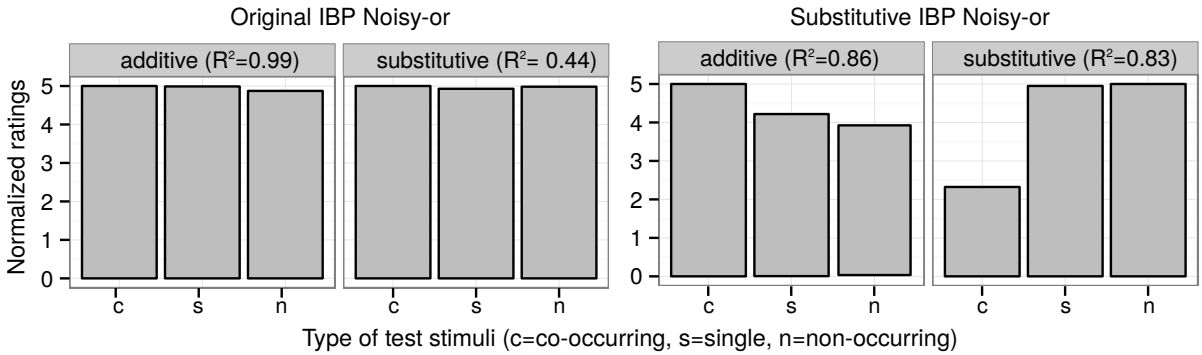


Figure 6: The both the original IBP noisy-or model and our model captured the additive condition reasonably well, but only our model predicted subjects' ratings in the substitutive condition.

ing more bars in the substitutive condition than additive condition (this can also be addressed by having six possible bars and equating the number of bars in each image across conditions). We will also pursue how well the substitutive model explains human performance while parametrically manipulating the negative correlation of the parts. In the substitutive condition of the current experiment, the parts of interest are perfectly negatively correlated. However, presumably, such a strong correlation is rare in the real world, and people should adjust their representation between extreme additivity and extreme substitutivity based on the observed correlation. Further, we will generalize the substitutive IBP noisy-or model to learn substitutive features with more than two values (e.g., line styles can be "solid", "dashed", or "dotted"), as previous work has demonstrated people can learn categories with three-valued features (Aitkin & Feldman, 2006). One potential way to address this limitation is to generate a set of feature image matrices from a Dirichlet Process (DP; Ferguson, 1973). This results in the number of feature image matrices also being inferred, which will enable the model to represent multi-valued substitutive features, thus extending the current model's assumption that substitutive features can only have two possible values. Together with the current findings, future work will further illuminate how people construct representations beyond the simple case of binary additive features.

References

Aitkin, C. D., & Feldman, J. (2006). Subjective complexity of categories defined over three-valued features. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (p. 961-966). New York: Psychology Press.

Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B., & Griffiths, T. L. (2015). Structure and flexibility in bayesian models of cognition. In *Oxford handbook of computational and mathematical psychology.* Oxford University Press.

Austerweil, J. L., & Griffiths, T. L. (2011). A rational model of the effects of distributional information on feature learning. *Cognitive Psychology*, *63*, 173–209.

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric bayesian framework for constructing flexible feature representations. *Psychological Review*, *120*(4), 817–851.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*, 209-230.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cognitive, Affective, & Behavioral Neuroscience*, *1*(2), 137-160.

Garner, W. R. (1978). Aspects of a stimulus. In *Cognition and categorization* (pp. 99–133). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gati, I., & Tversky, A. (1982). Representation of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(2), 325–340.

Goldstone, R. L., Greganov, A., Landy, D., & Roberts, M. E. (2008). Learning to see and conceive. In L. Tommasi, M. Peterson, & L. Nadel (Eds.), *The new cognitive sciences* (p. 163-188). Cambridge, MA: MIT Press.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects.* New York: The Bobbs-Merrill Co.

Griffiths, T. L., & Ghahramani, Z. (2011). The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, *12*, 1185–1224.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, *119*(4), 685–722.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. (1998). Development of features in object concepts. *Behavioral and Brain Sciences*, *21*, 1–17.

Wood, F., Griffiths, T. L., & Ghahramani, Z. (2006). A non-parametric bayesian method for inferring hidden causes. In R. Dechter & T. Richardson (Eds.), *Proceedings of the 22nd uai* (pp. 536–543). Arlington, VA: AUAI Press.