Psychological Review
2013, Vol. 120, No. 4, 817–851

0033-295X/13/$12.00 DOI: 10.1037/a0034194

A Nonparametric Bayesian Framework for Constructing Flexible Feature Representations

Joseph L. Austerweil
Brown University

Thomas L. Griffiths
University of California, Berkeley

Representations are a key explanatory device used by cognitive psychologists to account for human behavior. Understanding the effects of context and experience on the representations people use is essential, because if two people encode the same stimulus using different representations, their response to that stimulus may be different. We present a computational framework that can be used to define models that flexibly construct feature representations (where by a feature we mean a part of the image of an object) for a set of observed objects, based on nonparametric Bayesian statistics. Austerweil and Griffiths (2011) presented an initial model constructed in this framework that captures how the distribution of parts affects the features people use to represent a set of objects. We build on this work in three ways. First, although people use features that can be transformed on each observation (e.g., translate on the retinal image), many existing feature learning models can only recognize features that are not transformed (occur identically each time). Consequently, we extend the initial model to infer features that are invariant over a set of transformations, and learn different structures of dependence between feature transformations. Second, we compare two possible methods for capturing the manner that categorization affects feature representations. Finally, we present a model that learns features incrementally, capturing an effect of the order of object presentation on the features people learn. We conclude by considering the implications and limitations of our empirical and theoretical results.

Keywords: representations, feature learning, Bayesian modeling, computational modeling, computational constructivism

Supplemental materials: http://dx.doi.org/10.1037/a0034194.supp

A fundamental tenet of cognitive psychology is that a person's reaction to a stimulus is determined by his or her representation of the stimulus and not by the stimulus itself. This explains why two people looking at the same piece of art (e.g., a Jackson Pollock painting) can react very differently (Chomsky, 1959; Neisser, 1967). Representations are the central device for explaining the differing reactions to the same stimulus. An art historian may enjoy a Jackson Pollock painting due to her representation of it as a rejection of painting with a brush and exclaim, "That is beautiful!" A layperson viewing the same painting might dislike it because he represents it as a cluttered mess of discordant colors and exclaim, "That is ugly." Although in this case the difference in reaction of the two people is due to differences in knowledge of art, arriving at different reactions to the same physical stimulus depending on its context is commonplace in human behavior. Thus, understanding how the mind forms representations for stimuli is a fundamental problem for cognitive psychology.

Before delving into how representations are computed from raw sensory inputs and how that influences behavior, it is important to be clear about what we mean by a representation. Unfortunately, representation is a notoriously difficult concept to define (Cummins, 1989; Markman, 1998; Palmer, 1978). Perhaps a good starting definition of a representation, which can be traced back at least to Aristotle (Pitt, 2008), "is something that stands in place for

Joseph L. Austerweil, Department of Cognitive, Linguistic, and Psychological Sciences, Brown University; Thomas L. Griffiths, Department of Psychology, University of California, Berkeley.

This work was supported by Air Force Office of Scientific Research Grant FA9550-07-1-0351 and National Science Foundation Grant IIS-084541. The first method for including categorization was first presented at the 21st Neural Information Processing Society conference (Austerweil & Griffiths, 2009). The initial simulations and experiments in Appendix B (Transformed Indian Buffet Process) and the Learning Spatially Invariant Features, One or Two Features? and Learning Which Transformations Apply sections were first presented at the 23rd Neural Information Processing Society conference (Austerweil & Griffiths, 2010). Much of the work discussed in the article was completed while the first author was a graduate student at University of California, Berkeley, and served as part of his doctoral dissertation. We thank Rob Goldstone, Michael Lee, and John Anderson for their comments on a previous version of the article. We thank József Fiser for providing the raw images from several experiments. We thank Karen Schloss, Stephen Palmer, Todd Gureckis, Tania Lombrozo, Adam Sanborn, Michael Jordan, Charles Kemp, Amy Perfors, Noah Goodman, Eleanor Rosch, and the Cognition Coalition at the University of California, Berkeley, for insightful discussions; Frank Wood for providing MATLAB code of the noisy-OR Indian buffet process model; and Brian Tang, David Belford, Shubin Liu, and Julia Ying for help with experiment construction, running participants, and data analysis.

Correspondence concerning this article should be addressed to Joseph L. Austerweil, Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Box 1821, 190 Thayer Street, Providence, RI 02912. E-mail: Joseph_Austerweil@brown.edu

something else" (Palmer, 1978, p. 262). Although this is vague, it gives the gist of what a representation is: something (e.g., a symbol or the activation of artificial neurons in layer of a neural network) that stands for something else (e.g., an object in the environment or symbol in a different cognitive process). The internal representation being active indicates the presence of what it represents (whether or not what it represents is present).

Features[1] are a form of internal representation that is widely used in psychological theories (Markman, 1998; Palmer, 1999; Tversky, 1977). They are elementary units that can be simple, such as the presence of a horizontal line at a particular location, or more complex, such as connectedness. They can be discrete, whether binary (present or absent) or one of a countable set of values (e.g., the style of a line might be dashed, dotted, or solid), or they can be continuous (e.g., the length of a line).[2] The response to a given input stimulus is generated by first encoding the values of the features for the input and then making a decision on the basis of those feature values. If two people represent the same stimulus using different features, then their response to that stimulus may be different.[3] How people determine the appropriate features for a stimulus and how those features are inferred from the sensory data are typically left as an open problem: Models of human cognition typically assume that people adopt a particular feature representation for a stimulus. However, a full explanation of human behavior demands an explanation of how different representations are determined for the same stimulus (Edelman, 1999; Garner, 1974; Goldmeier, 1936/1972; Goldstone, 2003; Goodman, 1972, Chapter 9; Kanizsa, 1979; Murphy & Medin, 1985; Navarro & Perfors, 2010; Schyns, Goldstone, & Thibaut, 1998).

In this article, we focus on one particular aspect of how representations are determined for an input stimulus: when the representing thing is a feature representation and the represented thing is a set of objects from the environment presented in images.[4] To tackle this problem, we investigate the principles that the mind uses to form representations of the continuous flow of images from its environment. Focusing on how people infer feature representations from sets of images allows us to explore the effects of context because the same image in two sets can be given different representations. Selfridge (1955) provided a classic example of this effect. When an image (halfway between "A" and "H") is presented with "C" and "T," it is represented as "A" (to form "CAT"), but when it is presented with "T" and "E," it is represented with "H" (to spell "THE"). We compare how people and models defined within our computational framework infer features to represent an object in different contexts, when the context is the set of other objects presented with the original object.

The outline of the article is as follows. First, we define a set of criteria for any model of human feature representation inference and provide evidence for each criterion based on previous theoretical, empirical, and computational work. Next, we review previous computational work on human feature learning. Then, we present a computational framework using ideas from nonparametric Bayesian statistics to make it possible to allow objects to possess infinitely many features, but only a finite subset of them are relevant in any given context. We then present a model in this framework by Austerweil and Griffiths (2011) that satisfies some of the criteria. However, this model is unable to learn transformation-invariant features, and is indifferent to categorization information and the object presentation order. We then extend the framework to define a model that infers features invariant over a set of transformations. The model predicts novel contextual effects, which we test in a behavioral experiment. It also provides a way to understand how people might learn the dependencies between the transformations of features in a set of objects. Afterward, we present two possible ways to extend models in the framework to include the effects of categorization information on feature learning. Finally, we formulate an incremental learning form of the model, which, like people, is sensitive to the order in which objects are presented. We conclude with a discussion of the implications and limitations of our empirical and theoretical results.

Criteria for a Computational Framework of Human Feature Representation Inference

In this section, we suggest criteria that any computational framework for understanding how people solve the problem of forming context-sensitive feature representations should be able to satisfy. After proposing each criterion, we present theoretical and empirical support for it from the human feature learning literature.

Criterion 1: Sensory Primitives

Even though it may be possible to define a set of features that can represent all the objects that a person could plausibly encounter in his or her lifetime,[5] researchers have argued that such an approach does not adequately capture the representational power and flexibility of human cognition. Instead, they argue for *flexible* feature theories (Goldstone, 2003; Goldstone, Gerganov, Landy, & Roberts, 2008; Hoffman & Richards, 1984; Schyns et al., 1998), where the set of possible features can adapt with experience. For example, Hoffman and Richards (1984) argued that people infer features of objects at changes in concavity, and Braunstein, Hoffman, and Saidpour (1989) provided empirical support for people segmenting (some) novel objects into features at these points. Flexible feature theorists argue that it is implausible for our perceptual systems to be hardwired in such a way that it would be

[1] There are two main ways to describe an object: as an observable stimulus (e.g., its image on the retina) and as an internal representation. Following Austerweil and Griffiths (2011), we use *properties* or *parts* to describe the observable stimulus and only use *features* when describing the internal components of the object representation.

[2] Researchers have distinguished between discrete and continuous valued properties, where continuous properties are called dimensions (e.g., Krumhansl, 1978). In this article, we consider both discrete and continuous properties to be features.

[3] Alternatively, people can produce different behaviors for the same stimulus if they use different procedures based on the same feature representation. As it is difficult to distinguish between these two possibilities (Anderson, 1978), we focus on cases where the same procedure is applied to different feature representations.

[4] Thus, we are not learning abstract representations, such as "rejection of painting with a brush," but rather more concrete representations, such as portions of an object's image. Though we focus on images in this article, in principle our approach can be used to find feature representations of sensory data of other types (e.g., Yildirim & Jacobs, 2012).

[5] Although Biederman (1987) argued for a limit to the number of objects that a person encounters in his or her lifetime, this is peripheral to whether or not a fixed feature set can represent all of the objects a person could potentially observe during his or her lifetime.

possible to recognize quickly objects that are now important to people (such as cell phones or cars) that are very different from the objects that were present during the evolution of our species (Goldstone, 2003). Thus, as it is difficult to define a set of primitives that is capable of representing all ways an object can be represented in all possible contexts, the framework should instead construct features from raw sensory data.

Criterion 2: Unlimited Features

Although similarity is used to motivate feature representations (as features provide a basis for computing similarity), it is also a source of criticism. One of the most serious criticisms of feature-based similarity theories is that infinitely many features can be used to represent any particular object (Goldmeier, 1936/1972; Goodman, 1972, Chapter 9; Medin, Goldstone, & Gentner, 1993; Murphy & Medin, 1985). For example, a book is a collection of articles, but it is also smaller than an airplane, larger than a mosquito, lighter than an elephant, and so forth. Any of these features could be relevant (e.g., being larger than a mosquito could be appropriate if your goal is to stop getting bitten by a mosquito; see Barsalou, 1985). This casts doubt on the explanatory power of feature-based similarity theories because determining the appropriate features (which in turn define similarity) is a nontrivial problem (Murphy & Medin, 1985). Many cognitive theories depend on similarity for later cognitive processes (e.g., generalization), and thus, this is a major issue for most explanations using feature representations. However, if there was a way to have an unlimited repository of features and select which features should be used in a manner that is independent of similarity, this would not be an issue.

Thus, consistent with the human capability of being able to identify a potentially infinite number of features to represent any object (Goldmeier, 1936/1972; Goodman, 1972, Chapter 9; Medin et al., 1993; Murphy & Medin, 1985), a computational framework should not place an arbitrary limit on the number of features that can be used. The number of features should be inferred based on the observed set of objects.

Criterion 3: Context Sensitivity

The features people use to represent stimuli change or are reweighted according to context and experience (Garner, 1974; Gibson, 1969; Goldmeier, 1936/1972; Goldstone, 2003; Kanizsa, 1979; Palmer, 1999; Schyns et al., 1998; Tversky, 1977). We already discussed one example of this effect in perception from Selfridge (1955). Tversky (1977) presented an example in the conceptual domain. He found that people partitioned the same nations, Austria, Sweden, and Hungary, into two groups differently when Norway was included rather then Poland. When Norway was included, the four nations split nicely into neutral (Austria and Sweden) and nonneutral members (Norway and Hungary) based on cold war political alliances, but when Poland was included, the four nations are segregated based on geographic proximity (Sweden and Norway vs. Austria and Poland). This difference was also seen in people's similarity judgments, and suggests that they represented the nations using different features depending on the nations in the set.

One reason that feature representations should be context sensitive stems from their computational role: Features denote the commonalities and differences between the objects in a set. So, feature representations are inferred with respect to a set of objects (Garner, 1974; Selfridge, 1955; Tversky, 1977). Note that this criterion predicts contextual effects: When an object is presented with two sets of objects, it may be represented with different feature representations. Thus, the features inferred by a computational framework for a stimulus should be similarly influenced by its context.

Criterion 4: Prior Expectations

Previous work has already identified a number of perceptual and conceptual expectations that people use to infer features that represent objects: People infer features that are "simpler" (Austerweil & Griffiths, 2011; Chater & Vitányi, 2003; Hochberg & McAlister, 1953), consistent with background knowledge of the function of objects (Lin & Murphy, 1997), consistent with learned categories (Pevtzow & Goldstone, 1994; Schyns & Murphy, 1994), between transitions in concavity of the object's contour (Braunstein et al., 1989; Hoffman & Richards, 1984), or contiguous (Goldstone, 2000) and based on Gestalt Prägnanz (perceptual "goodness"; Palmer, 1977). Thus, the computational framework should easily include perceptual and conceptual constraints, if such constraints are relevant. However, it should still be able to infer features when any of the above types of information are absent.

Criterion 5: Transformational Invariance

People are able to recognize two images as having the same feature, even when that feature occurs differently in the two images (e.g., due to differences in viewpoint). In other words, the framework should be able to learn features that are invariant over a set of transformations (Palmer, 1983; Rock, 1973; Rust & Stocker, 2010).

Criterion 6: Category Diagnosticity

Previous work has demonstrated that categorization training is a catalyst for inferring new feature representations (Goldstone, 2000; Goldstone & Steyvers, 2001; Lin & Murphy, 1997; Pevtzow & Goldstone, 1994; Schyns & Murphy, 1994; Schyns & Rodet, 1997). In these categorization training studies, a set of parts is diagnostic of a category the experimenter has in mind, but initially they are not inferred by participants as features. After repeated feedback, participants discover that the part is diagnostic, and it is inferred as a feature. For example, Pevtzow and Goldstone (1994) created a stimulus set where each object shares one part with two other objects. Participants learned categories where each object shared one diagnostic part with the other object in its category. After categorization training, they inferred the part diagnostic for categorization as a feature, but not the part nondiagnostic for their categorization training. Thus, the framework should be able to infer features diagnostic for categorization.

Criterion 7: Incremental Learning

People learn features incrementally as they are needed to represent novel images or explain category information (Schyns & Rodet, 1997). For example, in Experiment 2 of Schyns and Rodet (1997), participants learned to categorize three types of "Martian

cells" (circles with differently shaped blobs inside). Two of the blobs, *a* and *b*, were diagnostic for determining category membership. They created three categories using these two parts: *A* (the cell contained the blob *a*), *B* (the cell contained the blob *b*), and *AB* (the cell contained the blobs *a* and *b* connected in a particular spatial arrangement). They trained two groups of participants to learn the three categories in one of two orders: $A \rightarrow B \rightarrow AB$ or $AB \rightarrow A \rightarrow B$. Their hypothesis was that those participants who learned to categorize *AB* first would learn a single feature that was the conjunction of *a* and *b* in a particular spatial arrangement, and so they would not extend category membership to a new Martian cell containing *a* and *b* in a new spatial arrangement. On the other hand, those that learned *AB* last would already have learned the features *a* and *b* and thus would think of the category *AB* as the conjunction of two preexisting features *a* and *b* that were learned to perform the two previous categorizations. If it is represented simply as the conjunction of these two features, the two features should be allowed to be in any spatial arrangement, and so, in this case, the participants should extend membership of *AB* to a new Martian cell containing *a* and *b* in a new spatial arrangement. The results of the experiment supported their hypothesis: Participants who learned *AB* last extended category membership for *AB* to a novel Martian cell with *a* and *b* in new spatial arrangements more often than the participants who learned *AB* first. Thus, the order that data were presented to participants affected the features they inferred, because those who learned *AB* first inferred the features *a*, *b*, and *ab*, and those who learned *AB* last inferred the features *a* and *b*. Thus, there should be an explanation for how the order in which objects are presented can affect the features inferred by the model.

Prospectus

Although many previous models satisfy some of the criteria outlined above, no existing computational framework can address all of the criteria. In the next section, we discuss many of the previously proposed models for explaining human feature learning. Afterward, we present our approach to satisfy all of these criteria by formalizing the mathematical problem of learning feature representations. Our formalization of this problem follows Austerweil and Griffiths (2011), who proposed a model that satisfied a subset of these criteria. We demonstrate that it satisfies the first four criteria we identified, and then go beyond this previous work by defining a new set of models within this framework that demonstrate that our framework can address the remaining criteria.

Computational Approaches to Inferring Representations

In this section we review and critique computational proposals for inferring feature representations based on the criteria discussed in the previous section. One way of classifying computational models for inferring feature representations is to split them into two major types: *weight change* approaches and *structure change* approaches. First, we describe neural network and Bayesian weight change approaches, where the feature set is fixed but the importance of each feature is inferred. Afterward, we assess flexible neural network approaches that alter their own structure while learning to predict a set of observations. Finally, before moving on to our framework, a Bayesian structure change approach, we consider the psychological validity of using classic dimensionality reduction techniques from machine learning as a computational explanation of human feature learning.

Weight Change Approaches

One potential solution for capturing the effects of context on features is to assume a large set of features is known a priori (e.g., raw sensory data) and provide a systematic method for determining when and how the weights of these features change. This approach has been explored with both connectionist (Goldstone, 2003; Grossberg, 1987; Rumelhart & Zipser, 1985) and Bayesian (Zeigenfuse & Lee, 2010) models. One early proposal by Nosofsky (1984, 1986) was that feature weights are chosen to optimize performance for classifying objects into their categories. Indeed, this proposal has been mostly supported (Nosofsky, 1984, 1986, 1991), although it may not be the case for some category structures (Vanpaemel & Lee, 2012). One of the most successful models for changing the weights of features is ALCOVE (attention learning covering map; Kruschke, 1992; Lee & Navarro, 2002). According to the ALCOVE model, the similarity defined by generalized context model (GCM) feeds into a neural network, which categorizes an object based on its similarity to stored exemplars of each category. The neural network learns feature weights according to error-driven learning. Though this is very successful for the small feature sets typically used in categorization studies (e.g., Gluck & Bower, 1988), the number of potential features people may use is potentially infinite or, at the very least, exponential in the dimensionality of the raw sensory input (e.g., every combination of pixels in an image is potentially a feature).[6] Due to the bias–variance trade-off (Geman, Bienenstock, & Doursat, 1992), very strong constraints on the types of features (i.e., strong inductive biases, Griffiths, 2010) will be needed to learn features effectively from somewhat realistic sensory data, as the space of possible features is astronomically large if not infinite (Schyns et al., 1998; Schyns & Rodet, 1997).

Furthermore, although categorization is an important catalyst for feature representation change, different features can be inferred to represent the same object in different contexts without any category information (Austerweil & Griffiths, 2011). In other words, categorization is sufficient, but not necessary for changing the features used to represent an object. Another sufficient condition for feature representation change is the distribution of parts over the context of an object. Thus, methods based on ALCOVE are insufficient, as they do not infer feature weight values in the absence of category information (but see Pothos & Bailey, 2009, for extensions of the GCM that learn attention weights in the absence of category information).

Using competitive learning rules, artificial neural network methods have been developed to infer feature representations from the distribution of parts over a set of objects without any category information (Grossberg, 1987; Rumelhart & Zipser, 1985). The networks start with a fixed number of features whose weights are

[6] Nosofsky (2011) does not advocate using raw sensory data as features in GCM, but rather first learning features from similarity data. However, this leaves open the question of interest in this article: How do people learn features?

initialized to random values, which means that they are not "specialized" to detect any particular visual properties. When an object is presented to the network, the network attempts to recreate it using its features. According to competitive learning rules, the feature that best recreates the observed object is considered the "winner," and its weights are updated with error-driven learning based on the difference between the observed and recreated objects (at a faster rate than the weights of "losing" features). As this procedure is repeated over an object set, the features differentiate and become specialized to detect different visual properties. Once the procedure converges, each object is represented with some number of the network's vocabulary of inferred features.

Though early competitive learning approaches were successful at inferring features without category information, two problems with these approaches is that they do not take into account domain-specific biases (such as expecting that visual features are contiguous; Goldstone, 2000) and they are insufficient to explain the effects of categorization in feature representation change. Goldstone and colleagues (Goldstone, 2003; Goldstone et al., 2008) developed the CPLUS (conceptual and perceptual learning by unitization and segmentation) neural network model to tackle both of these issues. This model makes two modifications to traditional competitive learning: It incorporates a perceptual bias toward learning approximately continuous features by favoring neighboring pixels to share the same value and a conceptual bias toward learning features that respect categorization information (features that activate for objects in one category but not the other). This is an interesting approach that uses both distributional and category information to infer features.

Zeigenfuse and Lee (2010) described a Bayesian feature weight change approach, where they assume that human similarity judgments are determined in a similar manner as the GCM, but Bayesian inference is used to determine the weight of each feature's importance in human similarity judgments. When their method is given a set of objects in a particular domain (e.g., farm animals) represented by a given large set of features and human judgments as to the similarity of various pairs of objects in the domain (e.g., horse and cow), it infers a smaller number of features with non-negligible weights that adequately reproduce human similarity judgments. Although it is a useful tool for reducing the number of features needed to adequately capture human similarity judgments and for investigating which statistical properties yield large feature weights, it does not alter its feature representation depending on context or experience. Thus, it is not appropriate for our purpose of understanding how people infer feature representations.

Importantly, all of these approaches assume the structure of the model is known ahead of time (e.g., the number of features), which is an unrealistic assumption, as the number of features is not provided in our sensory input. Since there are typically many representations of different sizes consistent with sensory inputs, determining the number of features is an important aspect of forming a representation. As people infer features without this additional information, it is inappropriate to include as input to a model of human feature representation inference.

Structure Change Approaches

Though the model's structure of current feature learning approaches is chosen a priori by the researcher (e.g., assuming a certain number of features), there is no reason to believe that it is impossible to develop a model that can alter its own structure (e.g., create new features as necessary). In fact, one method that would most likely be successful would be to specify a large number of potential features with a penalty for nonzero weights, which would bias them toward leaving a large number of the features unused. In fact, this is mathematically similar to the framework we introduce in the next section.

A similar approach exists in the neural network literature, which uses a mechanism for altering the architecture of the neural network, called cascade-correlation (Fahlman & Lebiere, 1990). Cascade-correlation recruits new nodes in the neural network when the current architecture does not adequately capture the pattern of input and outputs. This has been used successfully to capture developmental stages (Shultz & Sirois, 2008) and intermediate representations used for categorization (Love, Medin, & Gureckis, 2004). A flexible connectionist model for learning features using either mechanism would be very interesting and most likely could capture many of the feature learning results in this article; however, it would still have difficulty learning features whose images are transformed when instantiated in objects.

One model that can be interpreted as learning a set of features without presupposing a certain number of them a priori is the Bayesian chunk learner (Orbán, Fiser, Aslin, & Lengyel, 2008). It learns to group objects that appear to have been generated by a common cause. This Bayesian model infers the number of causes that "chunk" objects together, which is, in some sense, equivalent to learning features. However, it is given preprocessed visual information (a binary indicator for each object, which represents it occurring in the given scene). Thus, it does not answer questions that concern the visual input (e.g., it cannot explain how features with transforming images could be learned), and so it cannot explain how people learn primitives. This will prevent it from being able to explain how people infer features using both spatial and distributional information in an experiment that we report later in this article.

A different approach would be to apply dimensionality reduction techniques from machine learning, which find a smaller set of dimensions to represent objects (Abdi, Valentin, & Edelman, 1998; Edelman & Intrator, 1997; Hansen, Ahrendt, & Larsen, 2005). Unfortunately, Austerweil and Griffiths (2011) demonstrated that the results of two previously advocated techniques, principal component analysis (PCA)[7] and independent component analysis (ICA), are inconsistent with how people form low-dimensional representations based on their generalization patterns in several behavioral experiments. In these experiments, participants observed objects created from a set of parts, where the parts were either strongly or weakly correlated (but never perfectly correlated or independent). Then, participants judged how likely novel objects that were created from the set of parts were to be

[7] Another popular dimensionality reduction technique is projection pursuit (Friedman & Tukey, 1974), where an object set is reduced to a smaller set of dimensions that optimize some measure of the "interestingness" of each dimension. There may be some definition of interestingness of a dimension that may be consistent with the results of Austerweil and Griffiths (2011), but when the variance of stimuli on the dimension is the definition of the interestingness of a dimension, projection pursuit is equivalent to PCA (Carreira-Perpiñán, 1997).

members of the original set. Participants generalized to the novel objects when they observed objects whose parts were weakly correlated (reflecting that they inferred the parts as features), but did not generalize to the novel objects after observing the objects whose parts were strongly correlated (reflecting that they inferred the objects as features). Unlike people, the features inferred by PCA and ICA reflect the weak correlations between parts of the observed images, and thus, the features inferred by PCA and ICA may not be behaviorally relevant.

To explain the observed patterns of generalization behavior in their experiments, Austerweil and Griffiths (2011) proposed an ideal observer model that learns a feature representation from the raw sensory data of a set of objects without specifying the number of features ahead of time. This model is a nonparametric Bayesian model (see Gershman & Blei, 2012, for a review of nonparametric Bayesian models), which is able to achieve this goal by allowing a potentially infinite number of features to represent the set of objects but penalizing feature representations that use more features. Thus, the model infers a feature representation with as few features as possible that is still able to recreate the observed set of objects. As Austerweil and Griffiths demonstrated the ability of this model to explain their and previous experimental results, this model serves the foundation for the model we present in the next section. However, unlike people, this model fails Criteria 5–7: It is unable to learn transformation-invariant features (e.g., it learns a different feature each time a part occurs in a different part of the visual image), can only use categorization information in an ad hoc manner, and is insensitive to the order that objects are presented. In the rest of the article, we formulate a computational framework that includes this model and allows us to develop novel models that overcome these challenges.

Summary

Due to the importance of features in psychological theories, there have been a number of computational proposals for how people infer feature representations. These proposals vary in a number of ways, from fixed to flexible feature representations, connectionist to Bayesian models, and having origins in psychology or machine learning. Although many of the models capture interesting psychological phenomena, none of the previous models capture all of the previously discussed criteria. In the remainder of the article, we develop a computational framework capable of capturing these criteria.

A Nonparametric Bayesian Framework for Inferring Flexible Feature Representations

In this section, we describe a rational analysis of how features should be inferred to represent a set of objects (Anderson, 1990; Oaksford & Chater, 1998). A rational analysis compares the behavior produced by a cognitive process to the ideal solution of the underlying computational problem. It is at the abstract *computational level* (Marr, 1982), which focuses on the optimal solution to the mathematical problem of inferring context-sensitive feature representations. This constrains the *algorithmic level* or the cognitive processes people might be using to infer feature representations, which we explore in more depth later in the article.

To perform the rational analysis, we present a computational framework for solving the problem of context-sensitive inference of feature representations using nonparametric Bayesian models. We first review previous models defined within this framework by Austerweil and Griffiths (2011) that satisfied some of the criteria, and then go beyond this previous work by defining a set of new models that can address the remaining criteria.

Problem Formalization

Our computational problem is as follows: find a feature representation for a given set of objects satisfying the criteria outlined above. The set of objects are grouped into a matrix $\mathbf{X}$, whose N rows contain each object and D columns contain their observable properties. Although our framework is agnostic to the specific modality (it has been applied to auditory inputs and extended to multisensory data; Yildirim & Jacobs, 2012), for the purpose of this article, the observable properties of each object $\mathbf{x}$ (a row of the matrix) are an array of pixels representing the intensity of light reflected onto the retina. The two-dimensional array is converted into a one-dimensional vector of size D.[8] In this article, we further simplify the input to assume that each pixel is binary, though it is simple to use other types of observable properties (e.g., grayscale images; Austerweil & Griffiths, 2011).

One way to view this problem mathematically is as a problem of matrix factorization (Austerweil & Griffiths, 2011). Tentatively, let us assume that we know the number of features is K (we will relax this assumption later). As illustrated in Figure 1, the goal of a feature learning model is to reconstruct $\mathbf{X}$ as the product of two matrices: (a) a $N \times K$ binary feature ownership matrix $\mathbf{Z}$, where $z_{nk} = 1$ if object n has feature k, and (b) $K \times D$ feature "image" matrix $\mathbf{Y}$ that encodes how each feature gets instantiated in an object (e.g., if a feature of a mug was its handle, the corresponding $\mathbf{y}$ would be the handle's image.). Informally, this amounts to recreating each observed object $\mathbf{x}$ by superimposing the images $\mathbf{Y}$ of the features it has (given by its feature ownership vector $\mathbf{z}$). In this view, the model is solving the matrix equation $\mathbf{X} = \mathbf{ZY}$ for two unknown variables, the matrices $\mathbf{Z}$ and $\mathbf{Y}$ given only one variable, the matrix $\mathbf{X}$. Thus, the solution is underconstrained (it is analogous to solving a linear regression equation for the predictors and their weights simultaneously), meaning that additional information is necessary to solve for $\mathbf{Z}$ and $\mathbf{Y}$.

The solution to this problem is given by Bayesian inference (Geisler, 2003; Griffiths, Kemp, & Tenenbaum, 2008). Here the observed set of objects is our data, and the matrices $\mathbf{Z}$ and $\mathbf{Y}$ form our hypothesis. So, applying Bayes' rule gives us the following solution:

$$P(\mathbf{Z}, \mathbf{Y} \mid \mathbf{X}) \propto P(\mathbf{X} \mid \mathbf{Z}, \mathbf{Y}) P(\mathbf{Z}) P(\mathbf{Y}). \quad (1)$$

Thus, we have decomposed our original problem into three subproblems: find feature ownership and image matrices $\mathbf{Z}$ and $\mathbf{Y}$ such that they reconstruct the observed set of objects $\mathbf{X}$ well (designated by the "likelihood" $P(\mathbf{X}|\mathbf{Z},\mathbf{Y})$), and capture our prior expectations as to what makes a good feature ownership matrix ($P(\mathbf{Z})$) and a good feature image matrix ($P(\mathbf{Y})$). This is our proposed computational framework. A model in the computational framework is specified by how these three components are defined (as well as the method of inference).

[8] This is just for mathematical convenience. No information is lost, and it can be converted back to a two-dimensional array if necessary.

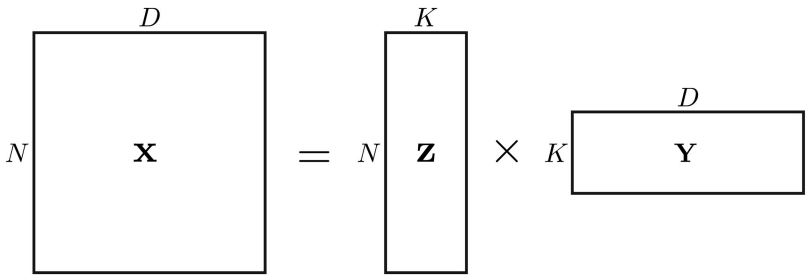

Figure 1. Formally, the problem of feature learning reduces to a matrix factorization problem. Represent a matrix **X**, whose rows are the objects **x**, using the product of a binary matrix **Z** (the feature ownership matrix), whose rows correspond to the features each object has, and a matrix **Y** that encodes the observable consequences of having each feature (the feature image matrix). In particular, each element of $(\mathbf{ZY})_{nd}$ encodes the number of object n's features whose image has pixel d on. This is not quite correct, as **X** is a binary matrix. More precisely, as $(\mathbf{ZY})_{nd}$ increases, so should the probability that $x_{nd} = 1$. In this article, this is done by assuming that x_{nd} is noisy-OR distributed (Pearl, 1988) with parameter given by $(\mathbf{ZY})_{nd}$.

To relax the assumption that we know the number of features a priori, we use a nonparametric Bayesian model as our prior probability distribution on **Z**. Nonparametric Bayesian models infer probability distributions that are potentially infinitely complex (in that the number of parameters required by a nonparametric Bayesian model to infer a distribution is infinite) but prefer simpler probability distributions (Jordan, 2010). Intuitively, the result is that the model infers the minimal amount of structure necessary to represent the observed data because the model trades off encoding the observed data with a bias toward simplicity.

In our particular case, the most common nonparametric Bayesian model for **Z**, a matrix with a finite number of rows (the objects) and an infinite number of columns (the K features), is the Indian buffet process (IBP; Griffiths & Ghahramani, 2005, 2011). It generates matrices with an infinite number of columns, but only a finite, random, number of them have at least one nonzero element.[9] As features are only part of a feature representation for a set of objects when they are actually used by at least one object, the infinite columns containing all zeros have no consequence. Thus, using this process, we can relax the assumption of knowing the number of features a priori and infer the number of features needed to represent the observed set of objects.

An Initial Model

On the basis of our analysis above, we define a model in the computational framework by specifying three probability distributions: the likelihood $P(\mathbf{X}|\mathbf{Z},\mathbf{Y})$, the prior on feature ownership matrices $P(\mathbf{Z})$, and the prior on feature images $P(\mathbf{Y})$. Once these three probability distributions are defined, we can adapt approximation techniques from machine learning to infer feature representations for a given set of objects (which we discuss further in Appendices A and F, and the Capturing Incremental Feature Learning section). As we discuss later in this article, how these three components are defined, what sorts of other data are incorporated with them (e.g., categories), and the approximation method used to infer the feature representations determine the behavior of the feature learning model. We begin by first explaining in depth the prior on feature ownership matrices and then provide an example of the simplest possible model in this framework, the initial model introduced by Austerweil and Griffiths (2011), and evaluate it against our criteria.

Feature ownership matrices are binary matrices, and so (supposing again that the number of features is known a priori) one simple method for defining a probability distribution on them is to flip a coin for each entry, putting a 1 in that entry for "heads" and 0 for "tails." One issue with this simple model is that all features have the same probability of being in an object. This is easily remedied by using a different "coin" for each column k, where the probability of heads for the coin used in column k is a parameter π_k. With this probability model for binary matrices, we can relax the assumption of knowing the number of features a priori. If we simultaneously increase the number of columns K to infinity and decrease the probability of a feature being in an object π_k to 0 at corresponding rates, the result is a probability model that only generates matrices with a finite number of columns having at least one nonzero element and an infinite number of columns containing all zeros. This is because for there to be "nice" limiting behavior as K increases, the probability of heads has to approach so close to 0 that the probability of there being a 1 in all future columns approaches 0 (i.e., at some point, the infinite "remaining features" will not be in any object).

The probability model resulting from this limiting construction is equivalent to an implicit probability distribution on binary matrices given by a sequential culinary metaphor, where objects (or rows) are "customers" and features (or columns) are "dishes" in an Indian buffet. Imagine an Indian buffet where the dishes are arranged sequentially (in order of the first customer that took a dish). Customers enter the Indian buffet one at a time, and we record in a matrix the dishes that they take according to the following set of rules. The first customer samples a number of dishes that is drawn from a Poisson distribution with parameter α. Each successive customer i takes a dish with probability $\frac{m_k}{i}$, where m_k is the number of people who previously sampled dish k, and

[9] More precisely, a matrix generated by the IBP has a finite number of columns with at least one nonzero element with probability 1. However, this is a mathematical technicality with no consequence for our purposes.

then sample Poisson new dishes with parameter $\frac{\alpha}{i}$ (where dividing by i captures the intuition that as more customers enter the restaurant, it should be less likely that a new dish is sampled). The number of previous customers who have sampled the dish, m_k, is normalized by the "arbitrary" customer index i because $i - 1$ is the number of previous customers who could have taken the dish (and the denominator is one larger due to our uniform prior beliefs over the probability that a customer takes the feature). This encodes the intuition that features that have been sampled by a large number of the previous customers are more likely to be sampled by new customers. The recorded matrix has the same probability under the culinary metaphor and the original limiting process.

Figure 2 illustrates how the culinary metaphor works for an example with three customers. The circles in Figure 2A denote dishes, the numbers around each dish denote the customers taking that dish, and the image above each dish is the observable consequence associated with taking that dish (we will explain how these are generated shortly). The culinary metaphor generates a corresponding feature ownership matrix $\mathbf{Z}$, which is shown in Figure 2B. The observable consequences of each feature $\mathbf{Y}$ are displayed in Figure 2C. Finally, Figure 2D demonstrates how each object is created by superimposing the observable consequences of the features it has.

The explicit form for the probability of the feature ownership matrix[10] after N customers who have taken K_+ dishes at least once is

$$\frac{\exp\{-\alpha H_N\}}{\prod_{h=1}^{2^N-1} K_h!}\alpha^{K_+}\prod_{k=1}^{K_+}\frac{(N-m_k)!\,(m_k-1)!}{N!}, \tag{2}$$

where K_h is the number of features with "history" h (the history is the natural number equivalent to the column of the feature interpreted as a binary number), K_+ is the number of columns with at least one nonzero entry (number of "used" features), and H_N is the Nth harmonic number ($H_i = \sum_{j=1}^{i} j^{-1}$). For example, the history for the second feature in Figure 2B is 3 ($(1, 1, 0) = 2^0 + 2^1 = 3$). Note $K_h!$ is 1 unless more than one feature has history h. We refer the reader to Griffiths and Ghahramani (2011) for the full derivation details and instead provide intuitions for the relationship between the culinary metaphor and Equation 2. The term $\exp\{-\alpha H_N\}(\alpha)^{K_+}$ is due to the sampling of new dishes (compare it to the form of a sum of Poisson distributed random variables), the term $\prod_{k=1}^{K_+}\frac{(N-m_k)!(m_k-1)!}{N!}$ encodes the probability customers chose previously sampled dishes, and $\frac{1}{\prod_{h=1}^{2^N-1} K_h!}$ is a normalization constant that accounts for dishes that are "equivalent" to the IBP (i.e., customers who made the same decisions for two or more dishes).

To define the likelihood, $P(\mathbf{X}|\mathbf{Z},\mathbf{Y})$, we form our observation matrix $\mathbf{X}$ that consists of the values of D observable binary properties (e.g., pixels in an image) for N objects ($\mathbf{X} = [\mathbf{x}^T_1, \ldots, \mathbf{x}^T_N]$, where $\mathbf{x}^T_i \in \{0, 1\}^D$). The likelihood compares the reconstructed object using the inferred feature representation ($\mathbf{ZY}$) to the observed set of objects. For simplicity, the observable properties in this article will always be binary pixels of an image (though "concept primitives" and grayscale pixels were used in Austerweil & Griffiths, 2011). Each element of $(\mathbf{ZY})_{nd}$ encodes the number of object n's features whose image has pixel d on (has the value 1). As $(\mathbf{ZY})_{nd}$ increases, so should the probability that $x_{nd} = 1$. Thus, the appropriate likelihood for our purposes is the noisy-OR distribution (Pearl, 1988; Wood, Griffiths, & Ghahramani, 2006), which seems to capture the assumptions people have for how an observed effect is produced by multiple hidden causes (Cheng, 1997; Griffiths & Tenenbaum, 2005). The total number of features that can turn on pixel d in object n is given by the inner product of the feature assignment vector for the object $\mathbf{z}_n$ with the column vector $\mathbf{y}_d$, which indicates whether or not pixel d is in the different feature images. Assuming each pixel is generated independently, the probability that a pixel is on is

$$P(x_{nd} = 1 \,|\, \mathbf{Z}, \mathbf{Y}, \lambda, \varepsilon) = 1 - (1 - \lambda)^{\mathbf{z}_n \mathbf{y}_d}(1 - \varepsilon), \tag{3}$$

where ε is the probability a pixel is turned on by chance and λ is the efficacy of a feature (the probability a feature with the pixel on in its image turns on the pixel in its object). One interpretation of Equation 3 is that it assumes each pixel is off a priori, but each feature that object n has with pixel d on turns the pixel on with probability λ and it is on by chance with probability ε.

The last portion of the model to specify is the prior on feature images $\mathbf{Y}$. This is where domain-specific expectations (e.g., perceptual biases such as good continuation) can be included into our framework. However, for most of our simulations, a simple "knowledge-less" prior on feature images suffices, where each pixel is on with probability ϕ independent of the other pixels in the image. Formally, this is a Bernoulli prior on each pixel with parameter ϕ specifying the probability the pixel is on or $P(\mathbf{Y}) = \prod_{k,d} \phi^{y_{kd}}(1 - \phi)^{1-y_{kd}}$. This prior distribution on $\mathbf{Y}$ in the model ignores spatial factors. However, Austerweil and Griffiths (2011) showed that it is a useful starting assumption, and that a proximity bias can be used to improve the psychological plausibility of the inferred feature images (although the learned feature ownership matrices were unchanged).

Evaluating Nonparametric Bayesian Models Against the Criteria

We will now argue that the first two criteria are satisfied by the framework itself, as a result of the problem formulation and solving it using methods from nonparametric Bayesian statistics. Austerweil and Griffiths (2011) considered models defined by the IBP as its feature ownership prior and either the knowledge-less or proximity bias feature image prior. We claim that these models satisfy Criteria 1–4, but not the other three criteria. This motivates us to develop models that satisfy the challenges posed by the remaining criteria.

We now evaluate how well the proposed computational framework satisfy the criteria. Criterion 1 (sensory primitives) is satisfied by using the IBP as our prior on the feature ownership matrix. This is because it identifies a feature image with each dish, which is inferred from the observable properties of the corresponding objects of the customers that took the dish. Through the limiting construction of the IBP, Criterion 2 (unlimited features) is satisfied

[10] Technically, this is the probability of all matrices equivalent to $\mathbf{Z}$ in the sense that they contain the same set of column vectors (i.e., ignoring any differences due to the order of the columns).

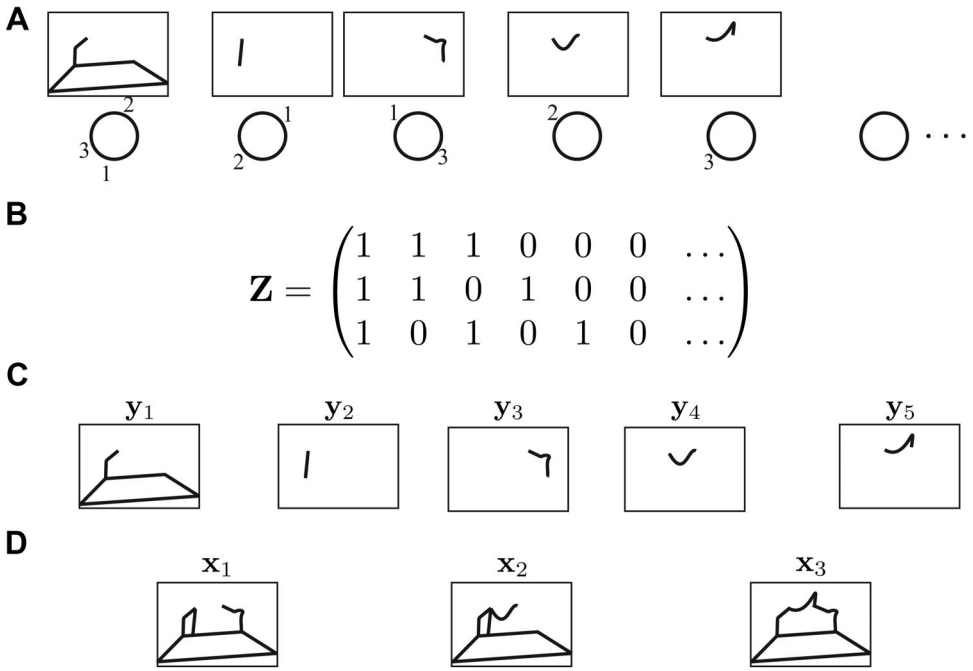

Figure 2. An illustration of the relation between the culinary metaphor, the Indian buffet process (IBP), and the model. (A) The culinary metaphor for the IBP. The numbers are customers (objects), and the circles are dishes (features). A number adjacent to a table represents its corresponding object taking that feature. A feature image is generated for each dish, which appears above the circle for each dish. (B) The equivalent feature ownership matrix represented by the culinary metaphor above. (C) The feature images generated from the feature image prior for each feature. (D) The reconstructed objects using **Z** and **Y** defined in Figures 2B and 2C, respectively.

as the number of features approaches infinity.[11] Although an infinite number of possible features might seem at first alarming and mystifying, for any finite set of objects, only feature representations with a finite number of features have nonnegligible probability. The IBP penalizes each feature that is actually used (it is assigned to at least one object). There are two ways that additional features are penalized by the IBP. The first way is from the third term in Equation 2, as it is the product of probabilities (for each feature, there is an additional $\frac{(N - m_k)!(m_k - 1)!}{N!}$ term, which is less than 1), and so each additional feature decreases the probability of the feature ownership matrix. Second, if α is set such that $\alpha/N < 1$, then the IBP implements a bias toward using fewer features, as there is a $(\alpha/N)^{K_+}$ term, which decreases as the number of features increases. Thus, it encodes a simplicity bias, capturing an important aspect of human perception and cognition (Chater & Vitányi, 2003; Hochberg & McAlister, 1953; Lombrozo, 2007), and hence it also satisfies an aspect of Criterion 4 (prior expectations).

The latent states (features) generated by the IBP are independent a priori (unlike the Chinese restaurant process [CRP]; Anderson, 1990; Griffiths, Sanborn, Canini, Navarro, & Tenenbaum, 2011; Sanborn, Griffiths, & Navarro, 2010). However, both the feature ownership matrix and feature images inferred by the model are dependent given the observable properties of a set of objects (through the likelihood function). Thus, the features are dependent under the IBP a posteriori (i.e., with the information provided by observing the objects). The current feature ownership matrix is used to infer feature images directly from the raw sensory data. Based on the other objects presented with an object, the feature images or feature ownership matrix could be different. For example, when the image in Figure 3A is presented in the context of Figure 3B, the IBP model represents it with a single feature shown in Figure 3D, which is the object itself. Conversely, when the image in Figure 3A is presented in the context of Figure 3C, the IBP model represents it with the three features shown in Figure 3E, which are three of the six parts used to generate the whole set of objects. As the model infers different features for an object appropriate for its context (the set of other presented objects), it satisfies Criterion 3 (context sensitivity). Furthermore, on the basis of the patterns of generalization behavior across several experiments, Austerweil and Griffiths (2011) suggested that this model infers context-sensitive representations in a similar manner to people.

Having shown that the first three criteria are satisfied by the model, we now consider how prior expectations are encoded in the model (Criterion 4: prior expectations). From our previous discussion, we have already established that one type of expectation (simplicity) is encoded by the Bayesian model. We now turn to

[11] This is true even though the number of unique feature images is finite because the IBP does not prevent two features from having the same image.

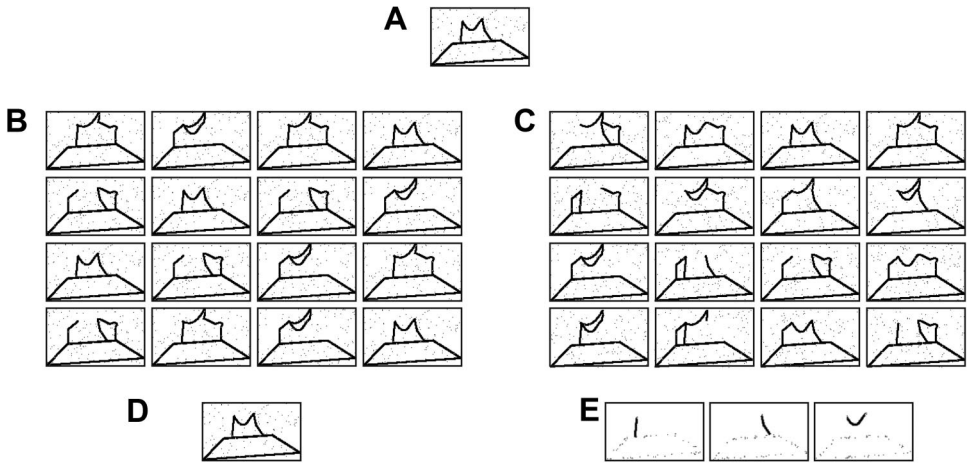

Figure 3. Context effects from Austerweil and Griffiths (2011). When the image (A) is presented with the images in context (B), participants and the Indian buffet process (IBP) represent the image using the object itself as a feature (D). Conversely, when the image (A) is presented with the images in context (C), participants and the IBP represent the parts as features (E).

how to include domain-specific knowledge and category information, satisfying the other portions of Criterion 4. One way to include domain-specific knowledge is to encode it in the feature image prior. For example, Austerweil and Griffiths (2011) showed that the initial model did not infer contiguous features (which was shown by Goldstone, 2000, to be a human perceptual bias in feature learning). When this model is given the images in Figure 4A (Set 1a from Experiment 1 of Shiffrin & Lightfoot, 1997), it infers features with "speckled holes" when the independent feature image prior is used (see Figure 4B). To address this issue, they defined a feature image prior $P(Y)$ to favor adjacent pixels to share the same value (called the Ising model; Geman & Geman, 1984). When the Ising model is used as the feature image prior, the resulting model infers more psychologically valid features that are contiguous (see Figure 4C). As both conceptual and perceptual biases can be incorporated in the model, it satisfies the Criterion 4.

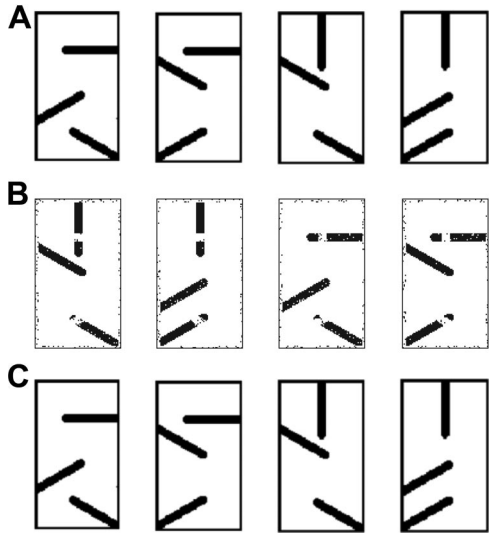

Figure 4. Incorporating perceptual biases into the model by varying the feature image prior from Austerweil and Griffiths (2011). (A) The four images given to the model, which were the Set 1a images in Experiment 1 of Shiffrin and Lightfoot (1997). (B) Features inferred by the model using the independent Bernoulli feature image prior. The features are not completely psychologically valid due to the "speckled holes." (C) Features inferred by the model using the proximity bias as a feature image prior. The features are more psychologically valid, as the speckled holes are filled in.

Although models defined by Austerweil and Griffiths (2011) satisfy the first four criteria, they do not satisfy the last three criteria. These are the simplest models that can be constructed from the proposed framework, suggesting that other models in the framework should satisfy these first four criteria. In the remainder of the article, we develop new models in the proposed framework that satisfy the other three criteria (transformation invariance, category sensitivity, and incremental learning). We first demonstrate how to learn transformation-invariant features and incorporate various expectations about how the transformations of different features are related (Criterion 5). Then we explore how to account for the effect of categorization on the learned features (Criterion 6). Afterward, we tackle one issue with models in the computational framework, which is that the probability distribution over feature representations is *order invariant*, meaning they infer the same features regardless of the order that objects are presented. This is not consistent with Criterion 7 (incremental learning), as there are reported effects of the order of object presentation on the features learned by people (Schyns & Rodet, 1997). We address this issue by formulating a rational process model for feature learning (in the spirit of Sanborn et al., 2010). This model is an incremental form of the transformation-invariant model and appropriately infers different features depending on the order that objects are presented.

Spicing It Up: Inferring Transformation-Invariant Features

We now address how to learn features that are invariant over a set of transformations (Criterion 5: transformation invariance). In

our environment, stimuli do not occur identically every time they appear in our raw sensory data. For example, imagine that on your desk, next to your computer, is your coffee mug. The position of the image of the coffee mug on your retina varies whenever your eye saccades. Or if you move your head toward the coffee mug while fixating on it, the coffee mug's image on your retina changes as well. Although your retinal image changes in each of these cases, objects and their properties in the environment do not change (they are invariant).

Fortunately the retinal image varies according to predictable rules or *transformations*. When your eye moves to the left, the retinal image before the saccade is the same as the retinal image after the saccade after each point is shifted to the right by some amount (see Figure 5 for a possible set of psychologically relevant transformations). Additionally, imagine that you move closer to the coffee mug. When you move closer, the mug's image scales to be larger, but otherwise it is the same. In both of these cases, given the original image and knowledge of the transformation, the resulting image is perfectly predictable. Thus, we can tell objects are equivalent when we rotate our heads even though the retinal image has been rotated (Rock, 1973), or when we navigate the world and the image of objects is translated and scaled in different ways (Palmer, 1983). In other words, people are able to use transformation-invariant features, and a computational model of human feature learning should be able to learn them too (Criterion 5).

Defining a Model That Includes Transformations

Although we have shown that the initial model is very successful at explaining human feature learning when those features occur identically on every presentation, features frequently occur differently in our sensory inputs across presentations. This presents a challenge for the usefulness of the initial models because they cannot learn features that occur differently across appearances. Currently, every transformed occurrence of a feature would be treated as a distinct and unique feature even though they are really the same feature after taking into account a simple transformation. Thus, the previously discussed models are not sufficient to explain human feature learning.

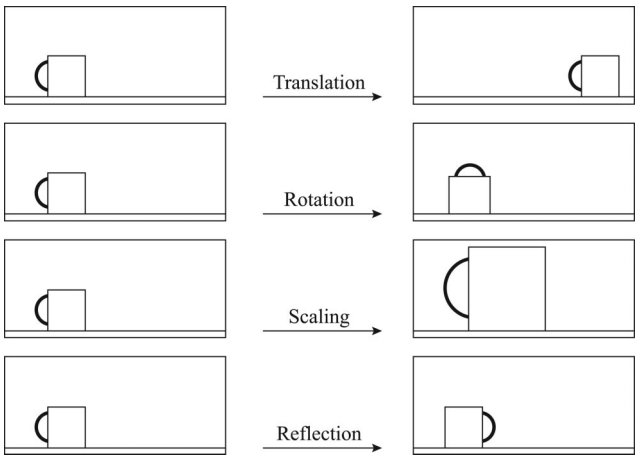

Figure 5. A possible set of transformations used by the human perceptual system. Adapted from "The Psychology of Perceptual Organization: A Transformational Approach," by S. E. Palmer, in *Human and Machine Vision*, p. 273, edited by J. Beck, B. Hope, and A. Rosenfeld, 1983, New York, NY: Academic Press. Copyright 1983 by Academic Press.

The images of objects and features change, but in predictable ways. So, if we include the possibility of features being transformed into the IBP model, we could learn features that are invariant to transformations in appearance. We will now demonstrate that this can be done by adding an extra step to the culinary metaphor for the IBP discussed above.

The additional step to the culinary metaphor is as follows: When a customer takes a dish (a feature is used by an object), she "spices" the dish (transforms the feature image) according to a set of spices available by the restaurant to every customer to put on every dish. The spice, or transformation, is not observed, but is drawn from a distribution over a predefined set (e.g., all possible translations, rotations, etc.), and so it must be inferred based on the data as well. It is a function whose input is a feature image and output is a new feature image. To calculate the likelihood, each feature image is transformed according to its sampled transformation. Then those transformed feature images are superimposed and used in the same way to reconstruct the observable properties of the objects via the inferred feature representation and to determine its likelihood. Otherwise, the model is essentially the same.

Figure 6 illustrates an example feature representation generated by the extended culinary metaphor, where we have defined the set of possible transformations to be right translations.[12] As before, Figures 6A–6C show how customers enter the restaurant drawing dishes, resulting in a corresponding feature ownership matrix, and associate feature images with each dish. Unlike the initial model, the extended model draws a transformation for every object and feature (even if the object does not use the feature) as shown in Figure 6D. This results in the reconstructed objects shown in Figure 6E.

This new process is called the transformed IBP (tIBP; see Appendix B for more technical details). Although the model knows what transformations are possible a priori, what the features are and how the features are transformed in each object are not known a priori. In other words, the model learns both a set of features and for each object how the learned features are transformed. By including the different transformations hypothesized by psychologists (like those in Figure 5), the tIBP can overcome the limitations of the IBP and, like people, learn features that are invariant over transformations typical of our environment.

By including some of the transformation types in Figure 5, the tIBP is able to learn transformation-invariant features, predict novel contextual effects, and explore assumptions that the perceptual system makes about the structure of transformations in a set of objects. First, we demonstrate the difference in the features learned by the IBP and the tIBP in a classic machine learning task: learning features from images containing horizontal and vertical bars in random locations (Ghahramani, 1995; Rumelhart & Zipser, 1985; Spratling, 2006). Next, we model results from Fiser and Aslin (2001, 2005), showing that people learn spatially invariant features (Fiser & Aslin, 2001) using the tIBP. Afterward, we test a novel

[12] Imagine the image to be a torus, so if an image is shifted past the size of a dimension, it wraps back to the beginning of the dimension.

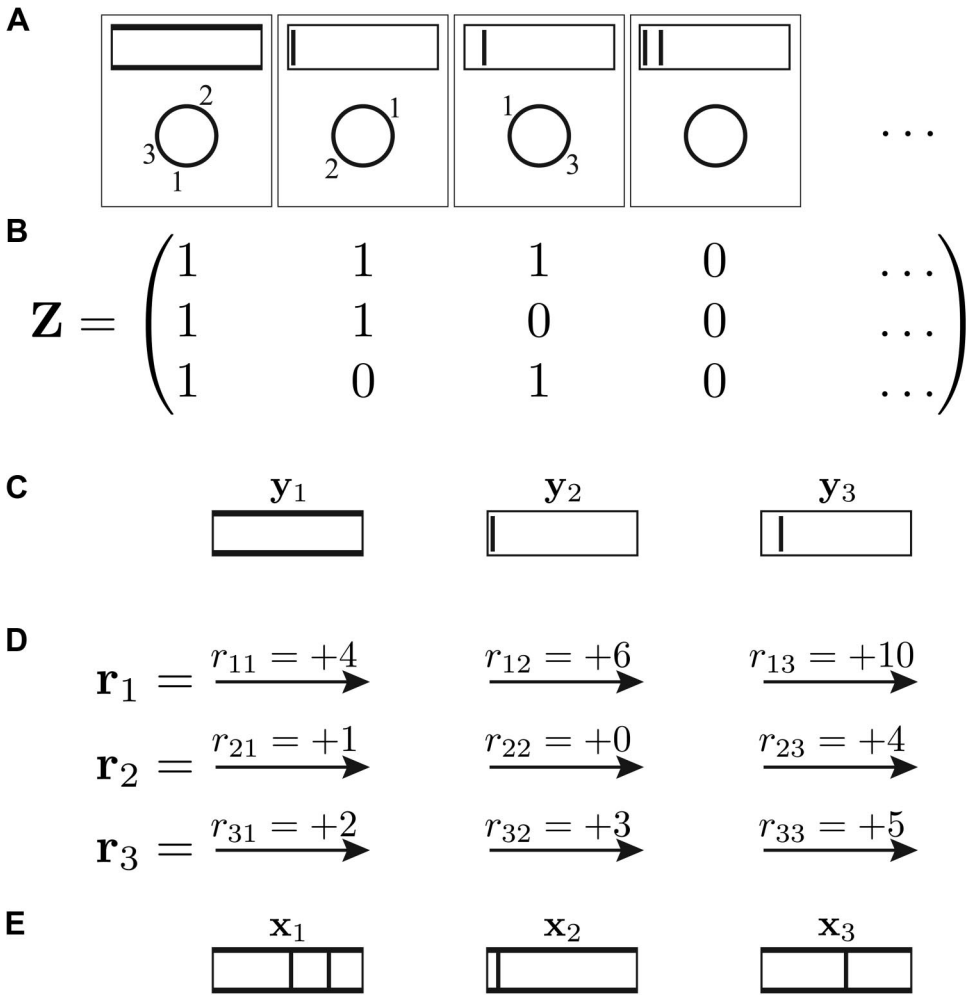

Figure 6. An illustration of the relation between the culinary metaphor, the transformed Indian buffet process (tIBP), and the model. Note that the only difference between this and the illustration of the Indian buffet process is the "spice" (transformation) that each customer (object) draws each time she takes a dish (feature). (A) The culinary metaphor for the tIBP process. The numbers are customers and the circles are dishes. A number adjacent to a table represents its corresponding object taking that feature. A feature image is generated for each dish, which appears above the circle for each dish. (B) The equivalent feature ownership matrix represented by the culinary metaphor above. (C) The feature images generated from the feature image prior for each feature. (D) The transformations drawn (uniformly over a set of translations to the right) for each customer for each dish. The number corresponds to how many pixels the feature image (above the transformation) is shifted to the right when it appears in a reconstructed object (its row number). Note that although every object–feature pair gets a transformation, the transformation only has a consequence when the object has that feature. (E) The reconstructed objects using $\mathbf{Z}$, $\mathbf{Y}$, and $\mathbf{R} = [\mathbf{r}^T_1, \mathbf{r}^T_2, \mathbf{r}^T_3]$ defined in Figures 6B and 6C, respectively.

context effect that is predicted by our model in a behavioral experiment.

We then explore how the perceptual system assumes that the transformations of features in the set of objects are dependent. Previous results by Smith (2005) suggest that people generalize the direction of one type of transformation (translation) to another type of transformation (scaling). Additionally, we demonstrate that people and the tIBP learn which types of transformations are allowed from the observed set of objects. In both cases, the model explains human behavior as a prior assumption about how feature transformations are dependent.

Models Learning Invariant Versus "Variant" Features

A classic problem in machine learning is learning features from images composed of horizontal and vertical lines in random locations (Ghahramani, 1995; Rumelhart & Zipser, 1985; Spratling, 2006). For example, Ghahramani (1995) applied his model, which reconstructs a separate feature for the horizontal and vertical lines in each position. Though this is a good solution, it is not ideal. Ideally, a horizontal line and a vertical line should be inferred as features. With these features, the model will be able to generalize

to images containing horizontal and vertical lines in novel locations. Figures 7A and 7B show the feature representations inferred by the tIBP and IBP, respectively, when given 100 images composed of a vertical bar and a horizontal bar each occurring in the image with probability .8 in a position drawn uniformly at random over the image. Not surprisingly (since the IBP assumes features occur identically in every image), the IBP model infers a feature for different positions of the lines (similar to the solution of Ghahramani, 1995) and some of the objects themselves. The tIBP model infers a horizontal bar and a vertical bar that occur in different positions—a better solution.

Learning Spatially Invariant Features

People form representations of objects, even when the objects' observable properties occur in different locations. To see whether statistical learning could be a potential mechanism for how people learn higher order visual representations, Fiser and Aslin (2001, 2005) investigated whether people learn different types of base units of scenes from merely observing them. In this section, we first describe their methodology, then go through a number of their studies and demonstrate how the tIBP similarly infers appropriately sized units from observing the pixel images of the scenes. Our evaluation of the tIBP in this section is qualitative, as our goal is to illustrate how the same framework constructs similar spatially invariant representations, and makes similar predictions as people from observations.

To test how people learn representational units from observing scenes, Fiser and Aslin (2001, 2005) defined a repository of 12 parts, similar to those shown in Figure 8A. First, they defined a set of "base units," which were subsets of parts from the repository that always co-occurred together in a particular spatial arrangement. Then they showed participants a set of scenes that were generated by including some number of the base units in random locations. For example, in the first two experiments of Fiser and Aslin (2001), they showed participants scenes that were composed of three base units out of a set of six possible base units. The base units were composed of two parts from the repository that always co-occurred together in a particular spatial arrangement. They called these "base pairs" which is a special type of base unit (see Figure 8B for an example set of six base pairs and Figure 8C for an example scene). After participants observed the set of scenes, they were asked to choose which of two scenes was more familiar to the previously observed scenes: a scene that contained a single base pair in a novel location (unobserved images) or a scene containing a set of parts that occurred together in the observed location (a part of an observed image) but was not a base pair. For the aforementioned example, after observing 144 randomly generated scenes, participants judged base pairs in novel locations (an unobserved image) as more familiar than a pair of parts that "accidently" occurred together (but was not a base pair) in a previously observed location. As the base pairs in novel locations had higher familiarity than parts of previously observed images, Fiser and Aslin suggested that participants learned the base pairs as translation-invariant units (or features) from passively viewing scenes in their experiments.

We now illustrate that a model using the tIBP infers the base pairs as translation-invariant features and makes similar familiarity judgments as participants did in Fiser and Aslin (2001). We did not use the exact images as Fiser and Aslin did because they were too large (1200 $\times$ 900 pixels), and simply reducing their resolution to a tractable size rendered the parts indistinguishable. Therefore, we preserved the same statistical structure by recoding each part to be 3 $\times$ 3 pixels, and then formed the same configurations that they used in their images. This resulted in scenes that were 15 $\times$ 15 pixels (we embedded the 3 $\times$ 3 grid in a 5 $\times$ 5 grid for consistency in modeling their later experiments). Figure 8B shows the parts grouped into their base pairs, and Figure 8C shows an example scene generated from these parts, one of the 144 scenes that the model observed, and the model infers the base pairs as features given the 144 scenes (see Appendix B for details). Not only does the model reconstruct the base pairs that were used to generate the images, but because it uses the tIBP, it also learns that the base pairs can occur in any location. We next compared the model's solution with people's familiarity judgments from Fiser and Aslin and found the same pattern. Given the feature representation inferred from observing the 144 scenes (the six base pairs), the probability of observing a base pair in a new location (a novel scene) was much larger than observing two parts that occurred

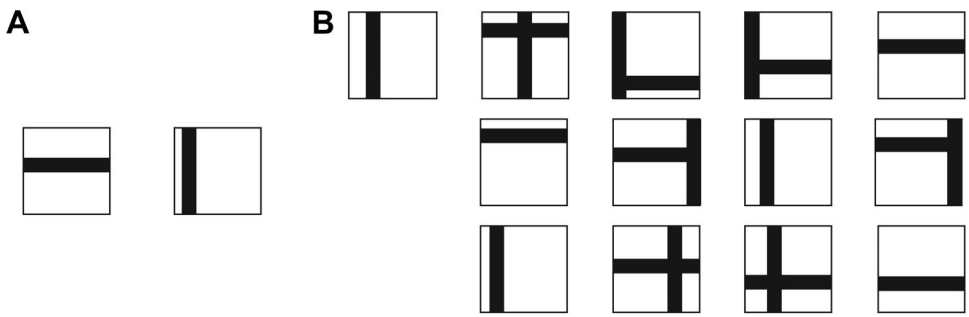

Figure 7. Comparing the features learned by two models defined within our computational framework. The features learned by the transformed Indian buffet process (tIBP; A) and Indian buffet process (IBP; B) given images containing a vertical and horizontal bar that randomly occur in random positions. Note that the tIBP is able to learn a vertical and horizontal bar due to the set of predefined transformations it is given a priori, but the IBP must create a different feature each time it observes a bar in a novel position.

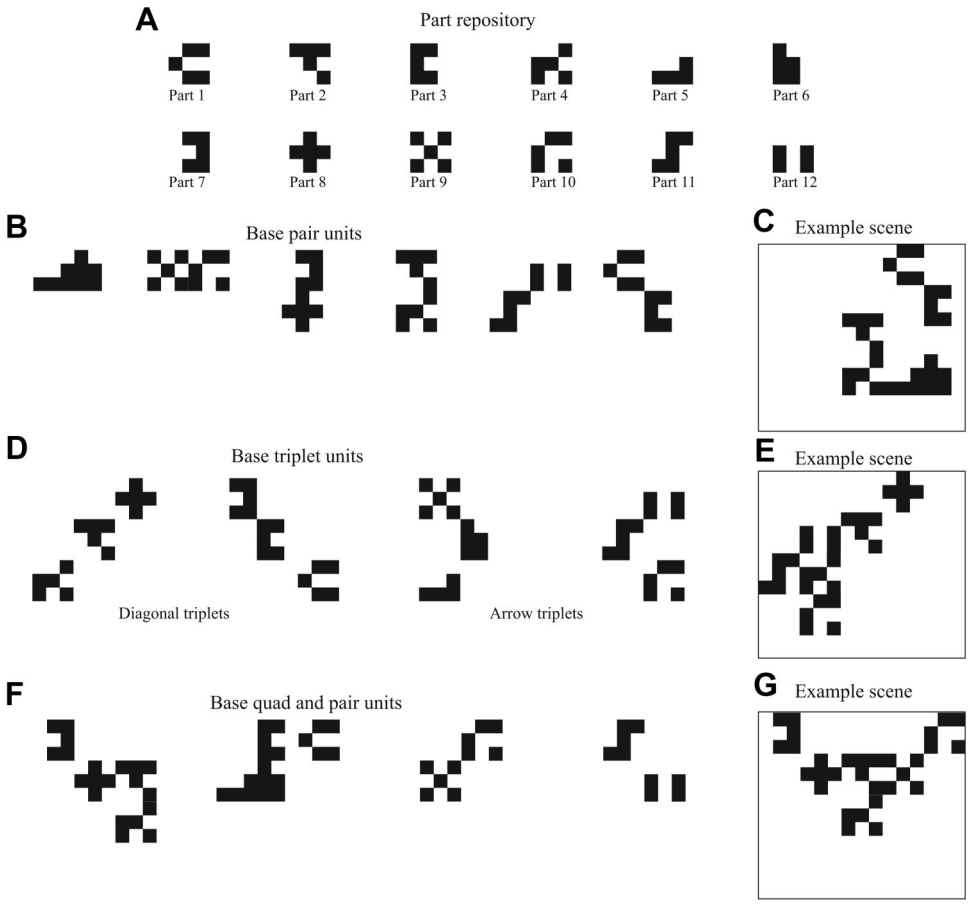

Figure 8. Learning spatially invariant units. (A) The repository of 12 parts that we use to explore the experiments of Fiser and Aslin (2001, 2005). (B) "Base pairs" as the base units. These are used in the baseline and frequency-balanced simulations (Fiser & Aslin, 2001, Experiments 1–3). (C) An example scene generated when the base pairs are the base units. (D) "Base triplets" as the base units. These are used in the triplet simulations (Fiser & Aslin, 2005, Experiment 1). (E) An example scene generated when the base triplets are the base units and each scene contains one diagonal base triplet and one arrow base triplet. (F) "Base quads and pairs" as the base units. These are used in the quadruple and pair simulations (Fiser & Aslin, 2005, Experiment 4). (G) An example scene generated when the base quadruples and pairs are the base units and each scene contains one base quad and one base pair.

in the same location together in a previously observed scene.[13] This was true for all six base pairs.

To investigate how robustly people can learn higher order visual representations using statistical learning, Fiser and Aslin (2001, 2005) used a similar experimental paradigm while (a) varying the type of statistics given to participants, (b) varying the number of parts in each base unit, and (c) using base units with different numbers of parts within the same experiment. By generating scenes that had both rare and more frequent base units, Fiser and Aslin (2001), in Experiment 3, equated the joint probability of the two parts in a rare base unit with a part in the rare base unit and a part from a frequent base unit. The base units were the same base pairs as used in Experiment 1 of Fiser and Aslin. Although the joint probabilities were equal, the parts of the rare base pair had a higher conditional probability (because parts within the same base pairs always co-occur) than a part of the frequent base pair with the part of the rare base pair. Participants observed over 200 scenes generated by combining three out of the six base pairs such that it equated the joint probability of some of the parts in different base pairs, by making some base pairs more probable than others. After observing these scenes, the participants judged the true base pairs to be more familiar than two parts from different base pairs, despite both options having the same joint probability. This suggests that the participants used conditional probabilities rather than joint probabilities to form units.

Does the tIBP model similarly form units using conditional probabilities rather than joint probabilities? To evaluate whether the tIBP model forms units based on conditional or joint probabilities, we gave the tIBP model the same scenes that participants

[13] The probability of new images given a feature representation inferred from a set of images is straightforward to calculate. See Appendices A and B for details.

observed in Experiment 3 of Fiser and Aslin (2001), downsized in the same manner as before, and found that it infers the six base pairs used to generate the scenes. Furthermore, conditioned on the features inferred by the tIBP model given the training scenes, the probability of the true base pairs was much larger than the nonbase pairs with equated joint probability. Thus, like people, the tIBP model forms units using conditional probabilities rather than joint probabilities.

To test whether people can learn triplet base units (base units composed of three parts), Fiser and Aslin (2005), in Experiment 1, had participants observe scenes containing "diagonal" and "arrow" triplet base units in random locations (as shown in Figure 8D), under the constraint that the base units were intertwined (as shown in Figure 8E). Participants preferred base triplets to a triplet of parts in a spatial configuration that never occurred during the training phase, but did not prefer a scene containing a pair embedded within a base triplet to a pair of parts that were never adjacent during the training phase. Thus, the participants represented the triplets using a single unit, but did not explicitly represent the pairs of parts embedded within the triplets.

Does the tIBP model infer base triplets from the same scenes participants observed in Experiment 1 of Fiser and Aslin (2005) and find base triplets more familiar than three parts in a novel configuration, but not find pairs embedded within a base triplet more familiar than a pair of parts that never were adjacent during training? To answer this question, we gave the tIBP model the same scenes that participants observed in Experiment 1 of Fiser and Aslin, downsized in the same manner as before, and found that it inferred the four triplet base units that were used to generate the scenes. So, conditioned on the features inferred by the tIBP model given the training scenes, the probability of true base triplets was larger than the fake base triplets (three parts in a configuration that never occurred during training). Next, we compared the conditional probability of embedded pairs to the probability of fake pairs (pairs never observed to be adjacent during training) and found that the ratio of the probability of an embedded pair to its fake pair comparison to be much lower than the ratio of the probability of a base triplet to the probability of a fake triplet. Thus, similar to the participants in Experiment 1 of Fiser and Aslin, the tIBP model infers triplets but does not infer pairs of parts embedded within the triplets as features.

Finally, Experiment 4 of Fiser and Aslin (2005) demonstrated that people could learn base units of different sizes simultaneously. In this experiment, participants observed scenes containing a base quadruple and base pair (base units and an example scene are shown in Figures 8F and 8G, respectively). After training, participants preferred the base quadruples to random arrangements of four parts, and they preferred the base pairs, but not embedded pairs (two adjacent parts within a quadruple), to two parts that were not observed to be adjacent to each other during training. Does the tIBP construct base quadruples and base pairs from the scenes that participants observed in Experiment 4 of Fiser and Aslin? To investigate this question, we gave the tIBP the same scenes that participants observed, downsized in the same manner as before, and found that it inferred the two base quadruples and two base pairs as features. Next, we found that it assigns a higher conditional probability to the base quadruples and pairs than to fake quadruples and pairs (four or two random parts in configurations that did not occur in the observed scenes), given the features inferred from the training scenes. Also, the ratio of the probability of an embedded pair to its fake pair comparison was found to be much lower than the ratio of the probability of a true base pair to the probability of a fake pair. Thus, the tIBP model infers spatially invariant features of different sizes (pairs and quadruples) and is sensitive in a similar manner as people. Furthermore, in both the simulations of Experiments 1 and 4 of Fiser and Aslin, the model acts in accordance with the embeddedness constraint, where subunits embedded in larger units are not explicitly represented. We refer the reader to Appendix B for a more detailed description of the simulations.

One or Two Features?

Although previous empirical work on visual statistical learning has investigated how people learn different types of visual units based on different types of covariance of parts across scenes, it has not considered the influence of coherent spatial transformation of units across scenes. This ignores the problem of recognizing spatially invariant units because it assumes this problem is already solved. Typically, spatially invariant units are given as inputs to whatever statistical learning technique is being used (e.g., Orbán et al., 2008). However, the mind does not have the luxury of nature providing it spatially invariant units in its input and, thus, should be able to form different units based on different patterns of coherent spatial transformations, even when the statistical information is the same. In this section, we explore a case where people and our model infer different representations based on coherent patterns of spatial transformations over the set of objects, even though the statistical information with respect to spatially invariant units is equivalent across conditions.

Having transformable features raises an interesting new problem: If components of an image can be mapped on to one another using one or more transformations, when should the components be perceived as different features, as opposed to different instantiations of the same feature? This problem is illustrated in Figure 9, where an object containing two vertical bars can be represented as (a) having a single feature containing two vertical bars (the *unitized* feature) or (b) having two features each containing a single vertical bar with its own translation (the *separate* features). An intuitive heuristic for solving this problem is to choose the feature representation with the smallest number of features that can adequately encode the observed objects. Figure 10 shows how this

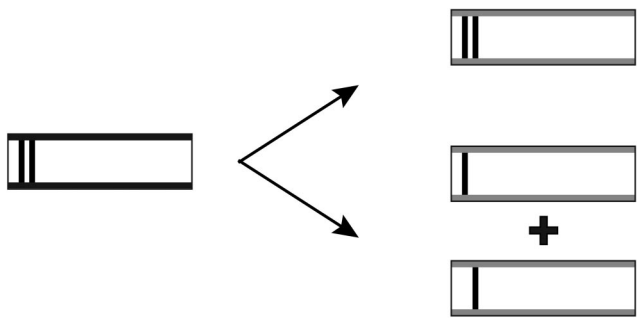

Figure 9. Does the image have one feature containing two vertical bars or two features that are both a vertical bar? Allowing transformations introduces new ambiguities for learning features.

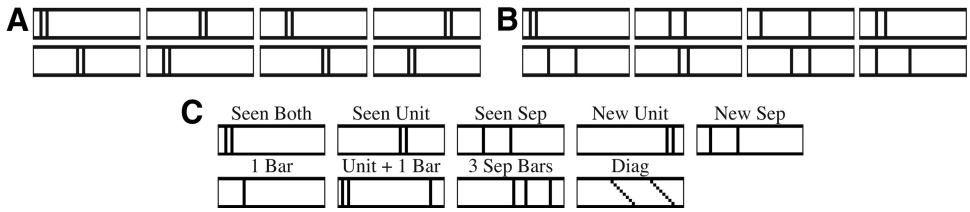

Figure 10. Resolving ambiguity when learning features invariant over translations by inferring the smallest feature representation that encodes the whole set of objects. (A) The unitized set. These objects were made by translating the unitized feature. (B) The separate set. These objects were made by independently translating the two features. The number of times each vertical bar occurs is equal over the unitized and separate object sets. (C) The test set, which participants generalized to after observing either the unitized or separate object set. Unit = unitized; Sep = separate; Diag = diagonal.

same object can be represented with different features, depending on the structure of the other objects in the set. In the unitized set (Figure 10A), all of the objects can be represented by the single two-bar feature, whereas the objects in the separate set (Figure 10B) require at least two single-bar features that translate independently. Although the objects in Figure 10A can be represented with two single-bar features (as in Figure 10B), that representation seems less good because it requires more features, and it would be a surprising coincidence that the two vertical bars are always the same distance if they varied independently.

The other objects expected to be in a set of objects depends on how those objects are represented. When the objects are represented by the unitized feature (a single feature consisting of two vertical bars a fixed distance apart), new objects with two vertical bars that same distance apart are expected. These objects are also consistent with separate feature representation (a feature for each vertical bar), and thus, we should expect participants in both conditions to generalize to new unitized objects. However, objects with two vertical bars at arbitrary distances apart (*New Sep*) can only be represented with the separate feature representation. Thus, if participants generalize object set membership to the *New Sep*, we can infer that they represent the objects using the separate feature representation (and if they do not, they are using the unitized feature representation).

We recruited 40 participants through Amazon Mechanical Turk (https://requester.mturk.com), which were randomly assigned to observe either the unitized or separate object set and paid $0.50 for their participation. Participants were told that they were given images that the Mars rover took while exploring a Martian cave (an adapted version of the cover story used in Experiments 1 and 2 of Austerweil & Griffiths, 2011). Unfortunately, three participants did not complete the task, and so there were 18 participants in the unitized condition and 19 participants in the separate condition. After observing the eight objects appropriate to their condition, they were asked to rate, from 0 to 6, how likely the Mars rover was to encounter nine other objects (presented in a random order) in a new portion of the Martian cave. The test images were organized into the following image types: *Seen Both* (an image in both the unitized and separate sets), *Seen Unit* (an image only in the unitized set), *Seen Sep* (an image only in the separate set), *New Unit* (an image that can be represented with either the unitized or separate feature set), and four control images, which are shown in Figure 10C.

The average participant ratings on the test images are shown in Figure 11A. The separate group generalized to the *Seen Sep*, $t(35) = 6.40$, $p < .001$, and *New Sep* images, $t(35) = 5.43$, $p < .001$, more than the unitized group, but otherwise the two groups generalized in the same manner. This supports our predictions that the separate group will use the separate feature representation, whereas the unitized group will use the unitized feature representation (even though the separate feature representation is also consistent with the objects they observed). As both groups generalized to the appropriate new test images, the participants inferred translation-invariant features.

The predictions made by the tIBP model given either the separate or unitized images are shown in Figure 11B (see Appendix B for simulation details).[14] The fit of the model is quite good, as the Spearman's rank order correlation between the human results and the model's predictions is .85 (using only one parameter[15] fit to human responses). One discrepancy between the model and human results is that the model predicts that participants in the unitized group should rate test objects that fit the unitized feature representation higher than participants in the separate group rate test objects that fit the separate feature representation. The model makes this prediction because there are more possible objects consistent with the separate feature representation than the unitized feature representation, and so each object is assigned lower probability by the model with the separate feature representation than the unitized feature representation. However, participants in both training conditions gave approximately equal rating to test objects, consistent with the conditions' corresponding feature representation, $t(35) = 0.89$, $p = .38$. The reason for this discrepancy is not clear. It is possible that people just rate all objects consistent with their current feature representation equally, which could be addressed in the framework by developing a more complex mapping from the raw model probabilities to ratings on a 0–6 scale. However, this conflicts with the results of Austerweil and Griffiths

[14] A form of the tIBP that learned relations about how features in the same image are transformed might produce similar results without forming a unitized feature given the unitized images. This is an intriguing idea, which may be difficult to distinguish from the formation of a unitized feature. We thank one of our anonymous reviewers for pointing this out.

[15] Although there are multiple parameters used by the model, only the value of one of them, ϕ, was chosen to minimize the discrepancy between the model and human responses.

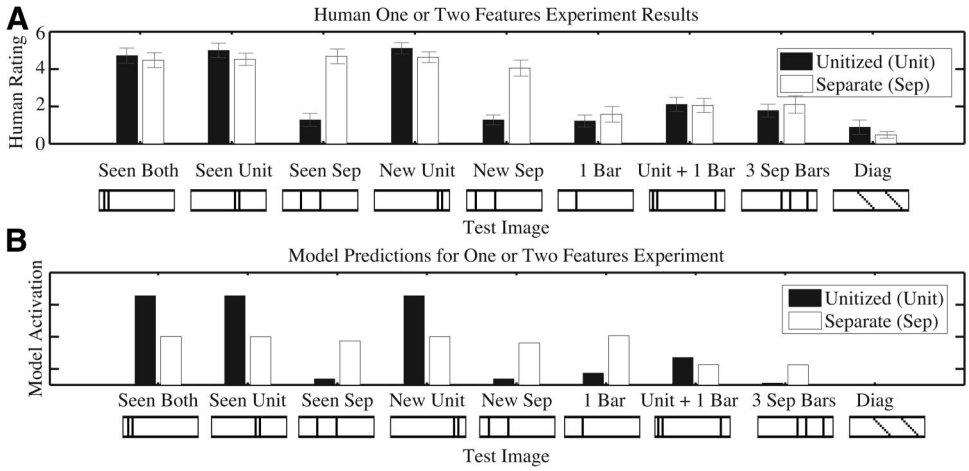

Figure 11. Human judgments and model results for learning features invariant over translations depending on the object set. (A) Human judgments. Participants who observed the unitized objects only generalized to the images that had the two vertical bars in the same distance apart. Participants who observed the separate objects generalized to any image with two vertical bars (regardless of the distance between the vertical bars). (B) The results of the transformed Indian buffet process model. Diag = diagonal.

(2011), who did find support for participants giving a larger rating to each test object when they were encoded with a feature representation with fewer features (participants who used the objects as features tended to rate the seen tests higher than those who used the parts as features). Understanding the relationship between the number of possible features (which in turn implies more possible objects) and the degree of generalization is an interesting question for future research.

One other discrepancy between the model and participants results is the *1 Bar* test image. The model incorrectly predicts that the separate group of participants should generalize the most to the *1 Bar* test image, as it only has one feature and the model prefers images with fewer features. These results indicate that while inferring the appropriate representation for a set of objects, people also infer other expectations about the set of objects (e.g., the number of features per object). This is an interesting phenomenon for future research, which we discuss further in the General Discussion. Note that previous statistical learning models do not predict any differences between conditions because once the parts are coded as invariant, the conditions are effectively equivalent. So they cannot explain these results. Thus, the degree of correspondence in how the tIBP and people similarly change their representations and behavior depending on the context is impressive.

Beyond Transformation Independence: Translations Affecting Scalings

In our environment, transformations do not always occur independently of each other. For example, imagine that while you look at the computer and coffee mug on your desk, you rotate your head 45°. The retinal images of the computer and coffee mug are not transformed independently; they are both transformed by a rotation of the same amount (roughly 45°). In other words, the computer and coffee mug share the same set of coordinate axes or *reference frame* (Marr & Nishihara, 1978; Palmer, 1975, 1989; Rock, 1973).

Another two types of transformations that are coupled are translations and scalings of a single moving object. Imagine fixating on an object while it moves toward you. As the object translates over time toward you, its image on your retina grows (the image scales equally in both dimensions). In its simplest form, the tIBP assumes that the transformations occur independently, and thus, it would not capture how people expect the translation and scaling of the image of an object to be coupled. However, it is easy to extend it to incorporate dependencies between transformations. If we assume that the translation and scale transformations are both normally distributed with their own means, but share the same covariance matrix, then we would expect the direction of variation in one transformation to be the similar to the direction of variation in the other transformation. Thus, if we observe an object translate horizontally, but not vertically, then the model with this added assumption will expect variation in horizontal, but not vertical, scalings (and vice versa if it observes vertical translations).

Do people generalize the direction that a feature is scaled based on the direction that it is translated? In Experiment 3 of Smith (2005), participants (children around 2.5 years old) learned that a novel object was a "zup" and watched an experimenter move the object in either a horizontal or vertical direction. Additionally, participants moved the object themselves in the same direction that they observed the experimenter to move it. Then they were asked whether or not they believed each of six test objects was also a zup. The test objects were identical to the given exemplar of a zup except that H1 was horizontally elongated, H2 was even more horizontally elongated, H− was shrunk horizontally, and V1, V2, and V− were similarly transformed but on the vertical dimension. As previously discussed, if translation and scaling transformations are coupled (e.g., because the object is translating fast and so a "blur" in the same direction occurs; McCarthy, Cordeiro, & Caplovitz, 2012), then participants should be more likely to believe

an object scaled in the same direction as the motion of a zup than in the opposite direction of the motion. Thus, participants who observed the given exemplar to move horizontally should generalize more to the horizontally elongated objects than the vertically elongated objects, and vice versa for those participants who observed the given exemplar move vertically (see Appendix C for details). This is essentially the result that Smith reported, as shown in Figure 12A.

To understand the dependency between translations and scalings within our framework, we added the extra assumption that was described above to the framework: The distribution generating translations and the distribution generating scalings share the same covariance matrix (see Appendix C for more details). As shown in Figure 12B, the model with this extra assumption is able to leverage the information learned about translations when observing scalings and generalize zup-hood in a similar manner as participants in Experiment 3 of Smith (2005). Indeed, the quantitative fit between the model predictions and human results is quite good, with a Spearman's rank order correlation of .94 and Pearson's product–moment correlation of .90.

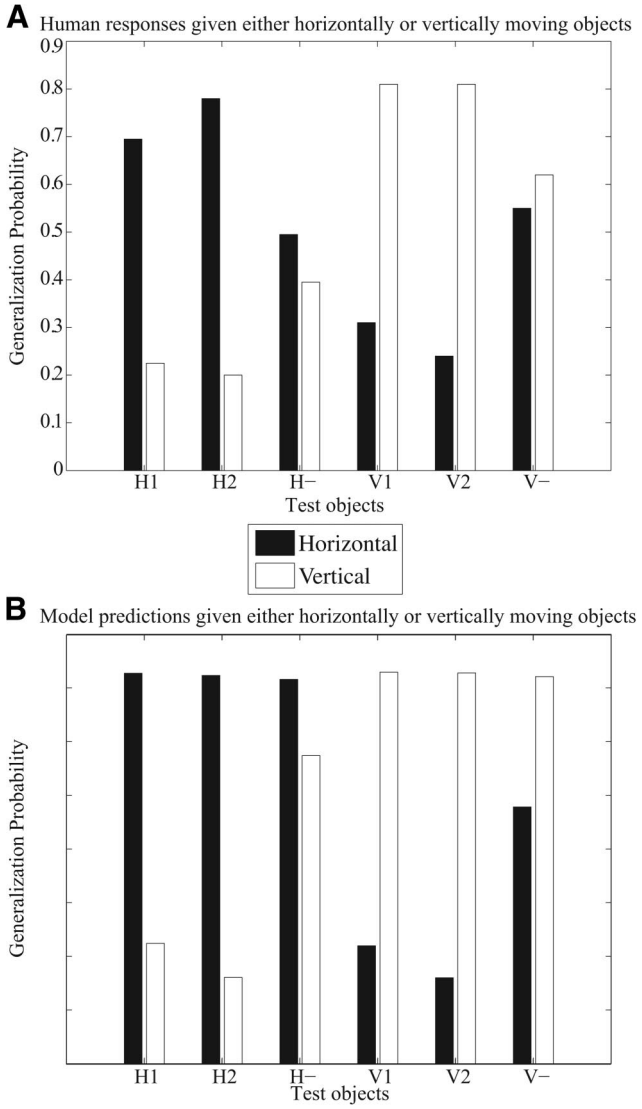

Figure 12. Sharing information learned about one type of transformation (translations) with another type of transformation (scalings). Human responses from Experiment 3 of Smith (2005; A) and predictions from the transformed Indian buffet process model (B). H = horizontal; V = vertical.

Learning Which Transformations Apply

An important question, once transformational invariance is introduced, is what kinds are transformations a feature can undergo while maintaining its identity. A given transformation (e.g., rotation) may or may not be permissible, depending on the object. For example, a 45° rotated image of the digit 5 will still be recognized as a 5, but a 45° rotated image of a square takes on a new identity as a diamond (Mach, 1897/1959). Squares and diamonds do not preserve their identity when rotated, whereas other shapes do, which suggests that people learn which transformations are allowed for particular shapes and features.

This phenomenon raises the question of whether people can infer which transformations are permissible for a particular feature. We addressed this question by testing whether participants' generalization behavior changed depending on feature transformations that were present in the observed set of objects. Using the same cover story as in the One or Two Features Experiment, we recruited two groups of 20 online participants from Amazon Mechanical Turk, presented them with either the rotation (see Figure 13A) or size set (Figure 13B), and paid $0.50 for their participation. Participants were then asked to rate (0–6 scale) how strongly five new objects appeared to belong to the observed set: *Same Both* (the object that is observed by both groups of participants), *Same Rot* (the last object of the rotation set), *Same Size* (the last object of the size set), *New Rot*, and *New Size* (see Figure 13C).

As expected, participants in the rotation condition generalized more to the *New Rot* object than those in the size condition, $t(38) = 4.44$, $p < .001$, and vice versa for the *New Size* object, $t(38) = 5.34$, $p < .001$. Supporting our hypothesis, people infer the appropriate set of transformations (a subset of all transformations) that features are allowed to use for a class of objects.

In its present form, the tIBP allows all transformations to be applied to all features because the transformations are assumed to be independent and identically distributed. To capture the experimental results supporting that people learn the types of transformation a feature is allowed to undergo, we relax the assumption that transformations are independent. We do this by extending the tIBP to infer the appropriate set of transformations by introducing latent variables for each feature that indicate which transformations the feature is allowed to use.[16] Given the observed set of objects, the extended tIBP infers the set of allowed transformations, along with the transformations and appropriate feature representation (see Appendix D for more details).

This extension to the tIBP predicts the *New Rot* object when given the rotation set and predicts the *New Size* object when given the size set—effectively learning the appropriate type of invariance for a given object class. Figure 14B shows the model pre-

[16] Note that the model is given general scale and rotation transformations, which return scaled or rotated images when given an arbitrary image. It learns which type of transformations can be applied, but not a general scale or rotation transformation.

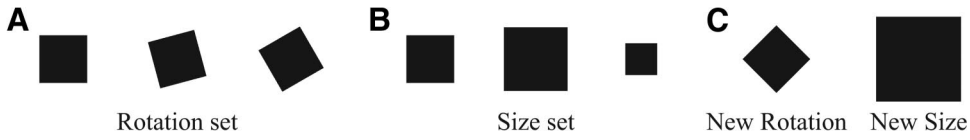

Figure 13. Stimuli for investigating how different types of invariance are learned for different object classes. (A) The rotation training images. (B) The size training images. (C) The two images used to test whether people inferred that rotation transformations were allowed (new rotation) or that size transformations were allowed (new sizes).

dictions (see Appendix D for simulation details). The model exhibits nearly the same behavior as the participants, with a Pearson's product–moment correlation of .98 (Spearman's rank order correlation of .88).

One interpretation of these results is that we have learned that a square is not orientation invariant because it is called a new name when rotated 45° (a diamond). This provides a novel potential explanation for why some shapes are orientation invariant and others are not: A shape is assumed to be orientation invariant unless there is reason from some observation that its identity changes when it is rotated (e.g., it is called a new name). Thus, the image of a rotated square is not a square because we have observed others call it a diamond.[17]

Using Categories to Infer Features

In this section we explore Criterion 6 (category diagnosticity), which is that the categorization of an object set should be able to affect the features people use to represent the objects (Pevtzow & Goldstone, 1994; Schyns & Murphy, 1994). In Pevtzow and Goldstone (1994), participants learned to categorize four objects, $\mathbf{x}_1$–$\mathbf{x}_4$, as shown in Figure 15. In the horizontal condition, participants learned the "horizontal" categorization scheme, where objects $\mathbf{x}_1$ and $\mathbf{x}_2$ formed one category and objects $\mathbf{x}_3$ and $\mathbf{x}_4$ formed another category. Conversely, in the vertical condition, participants learned the "vertical" categorization scheme, where objects $\mathbf{x}_1$ and $\mathbf{x}_3$ formed one category and objects $\mathbf{x}_2$ and $\mathbf{x}_4$ formed another category. After participants finished the category learning phase of the experiment, they showed enhanced processing[18] for the part that was diagnostic for the learned categorization scheme (specific to their condition), but there was no enhanced processing for the part that was not diagnostic for categorization. For example, participants who learned the horizontal categorization scheme showed enhanced processing for the two diagnostic parts (shown in the right of Figure 15), but not for the nondiagnostic parts (shown in the bottom of Figure 15), and vice versa for the participants in the vertical categorization scheme.

Note that participants in both conditions observed each object an equal number of times, and so the distribution of parts over images is the same for both groups of participants. As such, information about how parts are distributed over the images cannot explain why participants inferred different features (as in Austerweil & Griffiths, 2011) because the only difference between the two conditions is how the participants categorized the objects. Thus, the model used by Austerweil and Griffiths (2011) would infer the same features in both conditions, and so it must be modified so that categorization affects the features inferred by the model.

We now investigate two potential methods for incorporating category information into the feature learning framework: (a) appending the category label to the sensory information as a sort of observable property (similar to how category information is encoded by the rational categorization model; Anderson, 1990) and (b) assuming each category has its own feature representation (or IBP), but that they are coupled such that the features come from some common source (similar to how clusters are shared across categories in a hierarchical form of the rational categorization model; Griffiths et al., 2011). When the first technique is used (see Appendix E for simulation details), the model learns features to encode the diagnostic parts for the given categorization scheme, as shown in Figures 16A and 16B.[19] Although the model successfully infers different features depending on the categorization scheme (satisfying Criterion 6), it can only do so when it knows the number of categories a priori. However, to infer features for a newly learned category, new bits would have to be appended to the end of every observed image. This is unsatisfying because it should be simple to add new categories to the framework without altering every observed image.

Another approach for learning features that incorporates categorization information is to have a separate feature ownership and image matrix for each category. Unfortunately, a naive approach where feature ownership and image matrices for the images of category's objects are inferred independently does not work. Although it will infer diagnostic features to represent the images in each category, it will rarely share feature images between categories because the representations inferred to represent categories are completely independent of each other. This is undesirable because feature images tend to be shared between categories (e.g., all mammals have similar-looking legs). How do we define a model in this framework that has two feature ownership and image matrices for each category, and infers diagnostic features, but not at the expense of not sharing information between categories?

[17] Note that we are not suggesting that squares and diamonds are observed from specific viewpoints more often (than, e.g., a pentagon). Nor are we suggesting that pentagons are given different labels when observed from different viewpoints, but that squares and diamonds are given different labels when observed from different viewpoints. Although we do not contend that the differences between squares and diamonds may be due to other, more perceptual factors (this may be where the difference in labeling came from in the first place), we are interested in whether people can learn which transformations apply based on how labels are applied to transformed images of the feature.

[18] It was enhanced in the sense that participants were able to recognize a diagnostic part of a category in an object faster than a nondiagnostic part.

[19] When we only present a single set of features for a condition, the posterior probability of the feature set is usually orders of magnitude larger than the second largest feature set a posteriori. This is true of the feature sets shown in Figures 16 and 18.

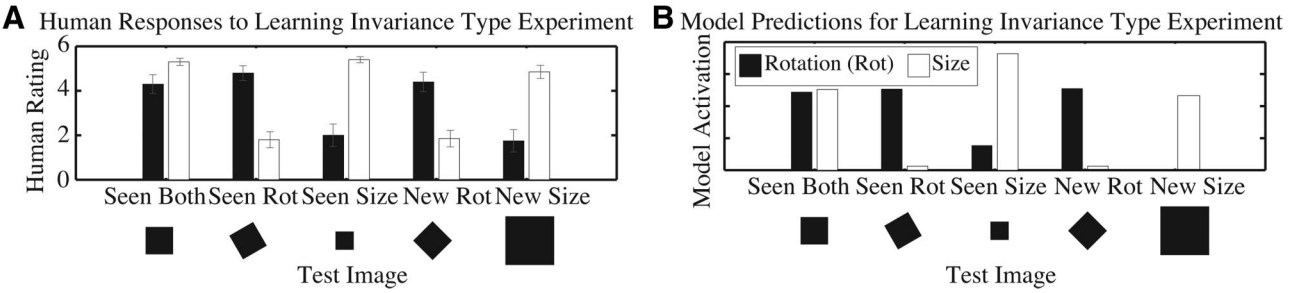

Figure 14. Learning the type of invariance. (A) Responses of human participants. (B) Model results.

One simple idea for coupling the features inferred to represent each category's images is to assume that they are generated from a common repository of features. In other words, the repository of feature images that are associated with each of the features of each IBP is the same.[20] As there is an infinite number of features to represent the objects in each category, the repository of feature images should be infinite while promoting feature images to be probable to the extent that they have been previously observed. To couple the feature images between the categories, the new feature images of each category's IBP are generated from a shared CRP (Pitman, 2002). The CRP is a commonly used nonparametric Bayesian model for categorizing data (Teh & Jordan, 2010) and seems to capture how people infer coherent clusters within categories and use that information when inferring clusters in other categories (Griffiths, Sanborn, Canini, & Navarro, 2008). Thus, it is apt for capturing how people assume feature images should be coupled between categories.

It is important to note that unlike with the IBP, the features are the customers of the CRP process that generates their images and they sit at tables, which represent their corresponding images. So, according to the CRP, customers (features) enter a restaurant with an infinite number of tables (each of which can hold an infinite number of customers) sequentially and each sits at a table, which serves a single dish (the image of features sitting at the table).[21] When the first customer enters the restaurant, all of the tables are empty. So, she sits at the first table and orders a dish for the table. As more customers enter the restaurant, they each sit at a table according to the following rule: Assume there are K tables with at least one customer and the number of customers sitting at table k is n_k, the ith customer sits at table $k \leq K$ with probability $\frac{n_k}{\beta + i - 1}$ and starts a new table with probability $\frac{\beta}{\beta + i - 1}$ (and order a new dish for the table), where β is a parameter that controls the probability of creating a new table (i.e., how likely are two features to have the same image?). Unlike in the IBP where each customer can take multiple dishes, customers in the CRP only take the one dish that is being served at their table.

In fact, this is a novel model, which we call the Indian buffet franchise (IBF), due to the analogous manner that parameters are shared across categories in a model known as the Chinese restaurant franchise (Teh, Jordan, Beal, & Blei, 2004). According to the IBF, each category has its own feature representation, and so the feature representation of category a is the combination of feature ownership matrix $\mathbf{Z}^{(a)}$ and feature image matrix $\mathbf{Y}^{(a)}$. Although the feature ownership matrices are generated independent of each other from the standard IBP ($\mathbf{Z}^{(a)} \sim \text{IBP}(\alpha)$), the feature images are generated from a common source $\mathbf{Y}^{(0)}$, according to the rules of the CRP ($\mathbf{Y}^{(a)}|\mathbf{Y}^{(0)} \sim \text{CRP}(\beta)$). The new dishes in $\mathbf{Y}^{(0)}$ are generated from the feature image prior, which has been an independent Bernoulli prior in this article.

Figure 17 illustrates how this model infers features after observing the first two objects of each category that have been categorized according to the horizontal scheme of Pevtzow and Goldstone (1994).[22] Let $\mathbf{x}_1^{(a)}$ denote the first object in category a, $\mathbf{y}_k^{(a)}$, to be the kth feature image of restaurant a (where $a = 0$ denotes the shared CRP and $a > 0$ denotes category a's IBP) and $n_k^{(a)}$ to be the number of customers who have taken dish k in restaurant a. To start, the participant observes the first object in Category 1, $\mathbf{x}_1^{(1)}$, enters Category 1's empty IBP and draws Poisson(α) new dishes (which happens to be two for this example). The two new dishes enter the shared CRP sequentially (in any order), which is empty to start. The first customer, $\mathbf{y}_1^{(1)}$, creates a new table and orders a feature image, $\mathbf{y}_1^{(0)}$, which is sampled from the feature image prior (the product of independent coin flips as before). This is the first dish that trickles back down to Category 1's IBP, and so $\mathbf{y}_1^{(1)} = \mathbf{y}_1^{(0)}$, which is now served in Category 1's IBP. Next, the second customer entering the shared CRP, $\mathbf{y}_2^{(1)}$, sits at the first table with probability $\frac{1}{1 + \beta}$ or starts a new table with probability $\frac{\beta}{1 + \beta}$. In this example, she happens to start a new table, order its dish $\mathbf{y}_2^{(0)}$, and this trickles back down to Category 1's IBP, and so $\mathbf{y}_2^{(1)} = \mathbf{y}_2^{(0)}$. Now the participant observes the next object, which happens to be the first object of Category 2. As this is the first object of Category 2, it enters an empty IBP, and draws Poisson(α) new

[20] It does not make sense to share information between the feature ownership matrices because this would amount to how features are assigned to objects in one category influencing the assignment of features to objects in the other category.

[21] Technically, in the statistics literature, customers sitting at the same table are only required to be in the same category under the CRP and do not necessarily share a parameter (or feature image in our case). When customers at the same table are assigned a parameter as well, it would be more precise to call the resulting culinary metaphor a form of the Pólya urn. However, much of the human and machine learning literature gloss over this technicality, and so we call this process the CRP for consistency with previous work in the area.

[22] For simplicity, we only list the prior probabilities in this example and not the likelihood terms (that relate the actual visual image to the feature image) when deciding to take each feature and sampling new feature images. See Appendix E for technical details regarding inference.

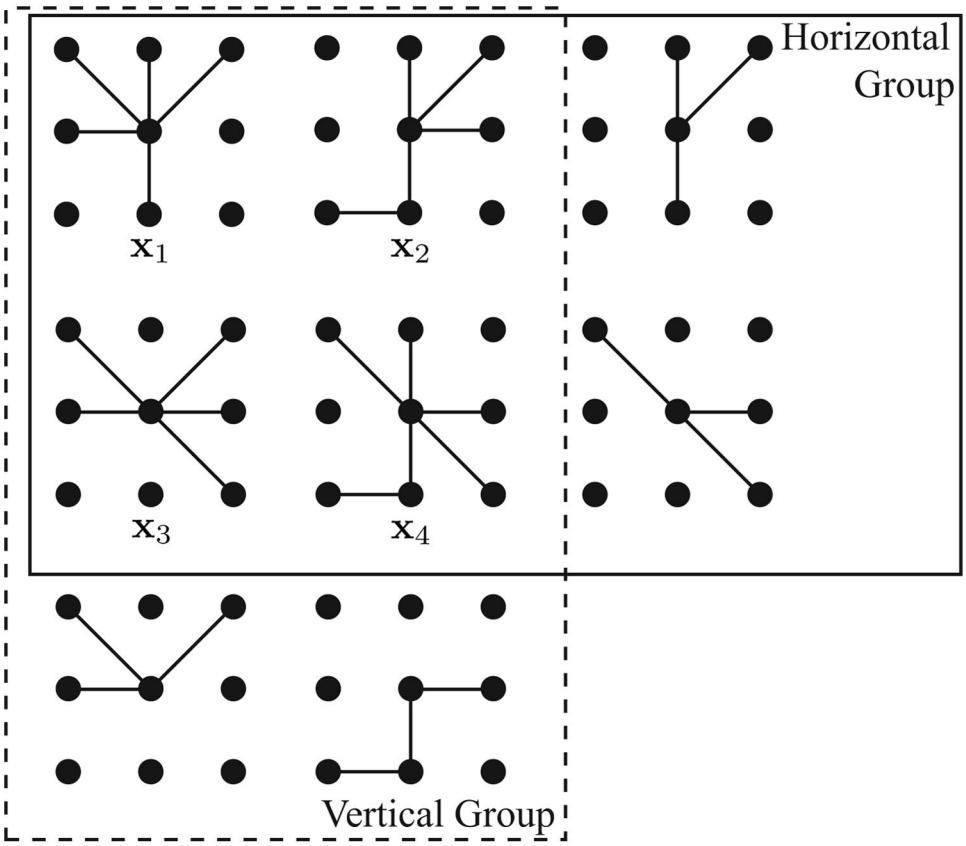

Figure 15. Effect of category learning on feature representations. Participants taught the horizontal categorization rule ($\mathbf{x}_1$ and $\mathbf{x}_2$ in one category, $\mathbf{x}_3$ and $\mathbf{x}_4$ in another) learn the features on the right, whereas those taught the vertical categorization rule ($\mathbf{x}_1$ and $\mathbf{x}_3$ in one category, $\mathbf{x}_2$ and $\mathbf{x}_4$ in another) learn the features on the bottom. Adapted from "Learning to Perceive While Perceiving to Learn," by R. L. Goldstone, in *Perceptual Organization in Vision: Behavioral and Neural Perspectives* (p. 256), edited by R. Kimchi, M. Behrmann, and C. R. Olson, 2003, Mahwah, NJ: Erlbaum. Copyright 2003 by Lawrence Erlbaum Associates, Inc.

dishes (which happens to be two for this example). The two new features for Category 2 become customers of the shared CRP and enter it to sample their feature images. When $\mathbf{y}_1^{(2)}$ enters the CRP, there are two tables with one customer, and so it sits at Table 1 with probability $\frac{1}{2+\beta}$ and Table 2 with probability $\frac{1}{2+\beta}$, and starts a new table with probability $\frac{\beta}{2+\beta}$. It happens to sit at the second table, and so $\mathbf{y}_1^{(2)} = \mathbf{y}_2^{(0)}$. Next, the second new feature of Category 2's IBP enters the CRP, and sits at Table 1, Table 2, or a new table with probabilities $\frac{1}{3+\beta}$, $\frac{2}{3+\beta}$, and $\frac{\beta}{3+\beta}$, respectively. It happens to create a new table, orders new dish $\mathbf{y}_3^{(0)}$ for the table, and brings it back to Category 2's IBP (meaning that $\mathbf{y}_2^{(2)} = \mathbf{y}_3^{(0)}$). This process continues as more objects are observed for each category, and Figure 17 shows the result after observing two new objects (one that is in Category 1 and another that is in Category 2). We refer the reader interested in the simulation details to Appendix E.

Figures 16C and 16D show that this model also represents each object with the same diagnostic and nondiagnostic parts as features. The features diagnostic for categorization are not shared between each category's IBP, and so it would be easy for an object recognition process to distinguish them. This allows a simple object recognition process to account for the result reported by Pevtzow and Goldstone (1994): A recognition process that encodes features specific to each category first (or weights them higher so they are likely to be encoded faster) and then encodes features that are in both categories would show enhanced processing for category-diagnostic features. Which features are shared across the IBP of each category depends on the categorization scheme of the object set given to the model. As categorization affects the features inferred by the model, Criterion 6 is still satisfied. Additionally, this model has the added benefit that it is simple to learn features for new categories. To add a new category, a new IBP is constructed, and it shares feature images with the previous categories using the global CRP in the same manner. Unlike adding categorization bits to the observable properties of an image, the previously observed images are not affected. In future work, it would be interesting to explore whether the two methods for introducing categorization information into the framework make diverging predictions and, if so, to distinguish between them using behavioral experiments.

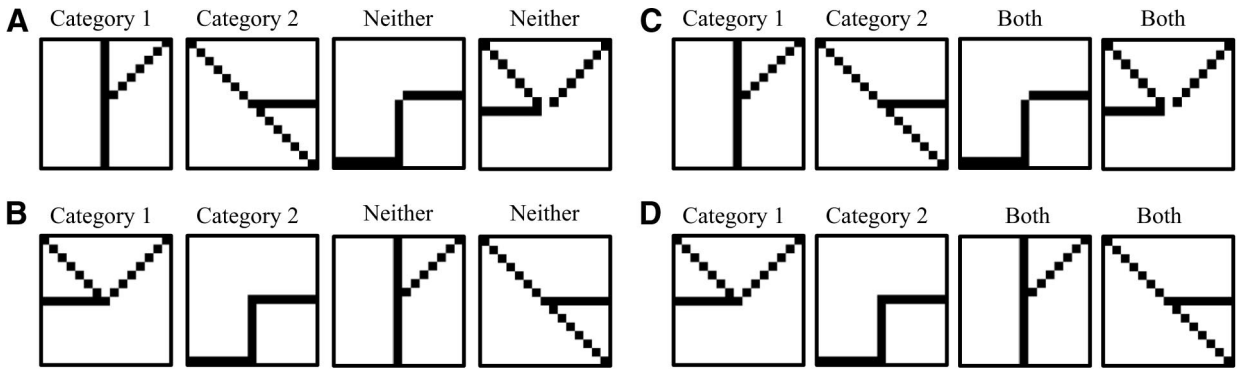

Figure 16. The effect of category information on the inferred feature representations given the objects from Pevtzow and Goldstone (1994), while varying the categorization scheme. (A–B) Features learned by the model, which encodes category information by appending it to the sensory data, using the horizontal and vertical categorization schemes, respectively. (C–D) Features learned by the extended model that has a separate Indian buffet process for each category, but couples the source of new features, using the horizontal and vertical categorization schemes, respectively. The title of each feature refers to whether or not the model associates the feature with one of the categories (but not the other), both of the categories, or neither of the categories.

Capturing Incremental Feature Learning

In this section we tackle Criterion 7 (incremental learning), which states that a computational model of human feature learning should be sensitive to the order that objects are presented. A common assumption (and criticism; see, e.g., Jones & Love, 2011; Kruschke, 2006; Trueblood & Busemeyer, 2011) of Bayesian models is that the probability of a sequence of objects is *exchangeable*, meaning that the model assigns the same probability to any order of the objects.[23] As human category and feature learning are sensitive to the order that information is presented (Anderson, 1990; Schyns & Rodet, 1997), order effects are a challenge for Bayesian models that assume exchangeability. This includes the models in our framework, which also assume exchangeability and are thus unable to produce order effects.

So far, we have focused on understanding how people infer features to represent objects from the computational level. We have not discussed the actual processes that might implement the computational solution (an algorithmic-level explanation). Inferring the features to represent objects according to a model in our computational framework is intractable (impossible to solve for explicitly), and so even machine learning methods can only find an approximate solution. Because Bayesian inference is intractable for most Bayesian models, machine learning researchers have developed many different techniques to approximate Bayesian models. Although most techniques used by machine learning researchers for approximating Bayesian inference yield the optimal solution under ideal conditions (i.e., with infinite memory and time), they may not in practice. As different approximation techniques deviate from the optimal solution in different manners, perhaps the order effects reported by Schyns and Rodet (1997) can be understood in our framework as a signature of the type of approximation technique used by people. This is the idea behind *rational process* models, which explain deviations from optimal behavior according to a Bayesian model as an artifact of using a statistically motivated approximation to a Bayesian model (Sanborn et al., 2010).

In this section, we explore two approximation techniques that have been proposed as rational process models: Gibbs sampling (Geman & Geman, 1984) and particle filtering (Gordon, Salmond, & Smith, 1993). The model predictions we have presented so far were computed with Gibbs sampling. Gibbs sampling works by starting with an initial random feature representation and then sequentially updating the feature representation by cycling through each element of the feature ownership and image matrices and resampling its value conditioned on the other values. As the number of updates goes to infinity, each update is a sample from the true posterior distribution over feature representations. Although Gibbs sampling is psychologically plausible in some situations, it requires all objects to be observed before approximation commences. Thus, it must start from scratch whenever it is given a new object, and it is not an *incremental* learner (Sanborn et al., 2010), which is a desirable property for an approximation technique to be psychologically valid (Anderson, 1991). Furthermore, the features learned by a Gibbs sampler are the same regardless of the object presentation order; however, people learn different features depending on the object presentation order.

[23] This is a valid criticism for Bayesian models that assume that the order in which information is observed does not matter. Although this is true of many Bayesian models of cognition, this assumption is not necessary to formulate Bayesian models. In fact, it is usually made for pragmatic reasons (computational ease) and not due to a strong theoretical commitment that agents should be indifferent to the order of observing information. We thank Michael Lee for pointing this out. Additionally, this also does not mean that Bayesian models assume that objects are independent and identically distributed. Exchangeability is a weaker assumption, which is related to the common assumption that a sequence of objects is independent and identically distributed. An independent and identically distributed sequence of objects is exchangeable, but an exchangeable sequence of objects is not necessarily independent and identically distributed. For example, the probability of a sequence of objects in our computational framework is exchangeable, but it is only independent and identically distributed given the feature representation of the objects. See Bernardo and Smith (1994) for more details on the distinction between objects being exchangeable and independent and identically distributed.

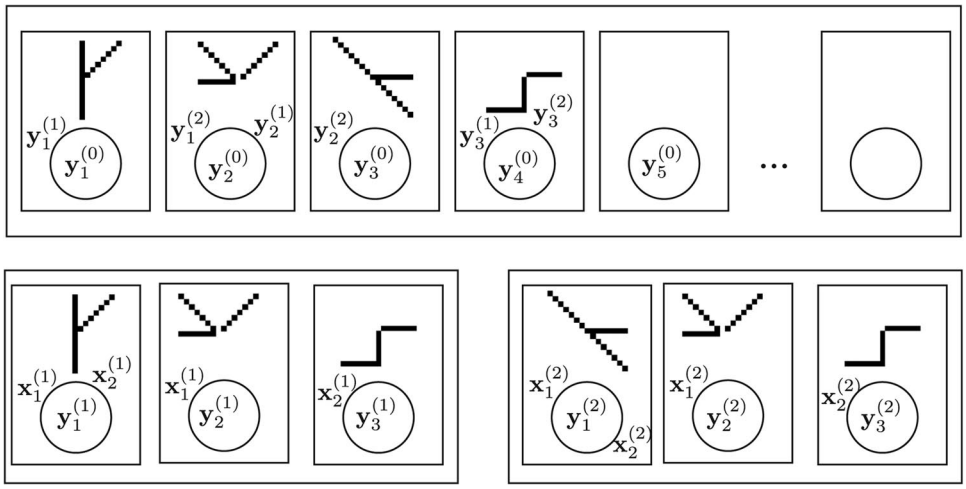

Figure 17. The Indian buffet franchise after observing two objects from the two categories under the horizontal categorization scheme of Pevtzow and Goldstone (1994).

A particle filter is an incremental learner and an alternative rational process model for models in our computational framework. As Sanborn et al. (2010) used a particle filter to explain order effects in categorization, it is plausible that a rational process model for the IBP can explain the feature learning order effects found by Schyns and Rodet (1997). The particle filter for models in the computational framework uses a number of "particles" for approximation, where each particle is a possible feature representation for the currently observed objects. The probability of a feature representation is given by the proportion of particles containing that feature representation. Each time a new object is given to the particle filter, the representations in each particle are updated to account for the new object. The feature representation in each particle is inferred sequentially as each object is observed using the previously inferred feature representations. Importantly, the feature representation for an object within a particle does not change once features have been inferred for it.[24] For capturing the results of Schyns and Rodet, this property is the most important difference between the previously discussed Gibbs sampler and the particle filter: Each particle in the particle filter infers features for an object once, only when it is first observed. So, the only information that influences the features inferred by the model to represent an object is the object itself and the previously observed objects. The objects observed in a particle filter after the current object have no effect on the features inferred to represent the current object. This is not the case though for the Gibbs sampler.

Because each particle in the particle filter only infers features for objects once, it is weakly dependent on the object presentation order (especially when only a small number of particles are used). This allows it to account for the order effects observed by Schyns and Rodet (1997). Consider the condition where participants learn *AB*, *A*, and then *B*. When a particle first observes the *AB* objects, it infers a single feature, *ab*, because it can encode all of the observed objects. Next, the particle encounters the *A* objects. As the feature it currently possesses, *ab*, cannot explain all of the objects in *A* (they do not have the *b* part), the particle captures the objects by inferring the *a* part as a feature. By an analogous argument, the particle infers the *b* part as a feature to explain the objects in *B* that it observes next. As the particle represents the *AB* objects with a single feature that is the conjunction of the *ab* parts in a particular spatial arrangement, it cannot represent the test objects that have *a* and *b* in new spatial arrangements with its *ab* feature. Thus, in this case, it would not extend the *AB* category membership to these objects.

Conversely, consider the condition where participants learn *A*, *B*, and then *AB*. When a particle first encounters the *A* and *B* objects, it infers two features, *a* and *b*, to represent the two categories. Next, the crucial difference between the previous condition occurs. When the particle observes the *AB* objects, it can represent them as containing two features, *a* and *b*. So, any object containing those two features (regardless of spatial arrangement) can be represented in the same way as the particle represents the *AB* objects in this case. Thus, it would represent the test objects (that have the *a* and *b* parts in a new spatial arrangement) in the same way as the *AB* objects and, accordingly, extend the *AB* category membership to these objects.

By extending the particle filter for the IBP by Wood and Griffiths (2007) to the tIBP (the tIBP is necessary because *a*, *b*, and *ab* have independent translations for each object they are in), we reproduce the order effects reported by Schyns and Rodet (1997).[25] Figures 18B and 18D show that the features inferred and predictions made by the tIBP using a Gibbs sampler (Gibbs) and a particle filter are the same as those made by participants in Schyns

[24] Though the feature representation of an object within a given particle does not change, the distribution over feature representations can change over time (if the proportion of particles with each feature representation changes). For the simulations reported in this article, all of the particles typically contain the same feature representation. In general, this may not happen. See Appendix F for more details.

[25] Although participants in Schyns and Rodet (1997) learned categories as part of the experiment, no category information is necessary for our model to capture participant judgments. This suggests that the category training portion of the experiment may have been unnecessary and mere exposure could have elicited the same ordering effects on feature learning. The exact nature of how category learning affects feature learning is unclear and demands future research.

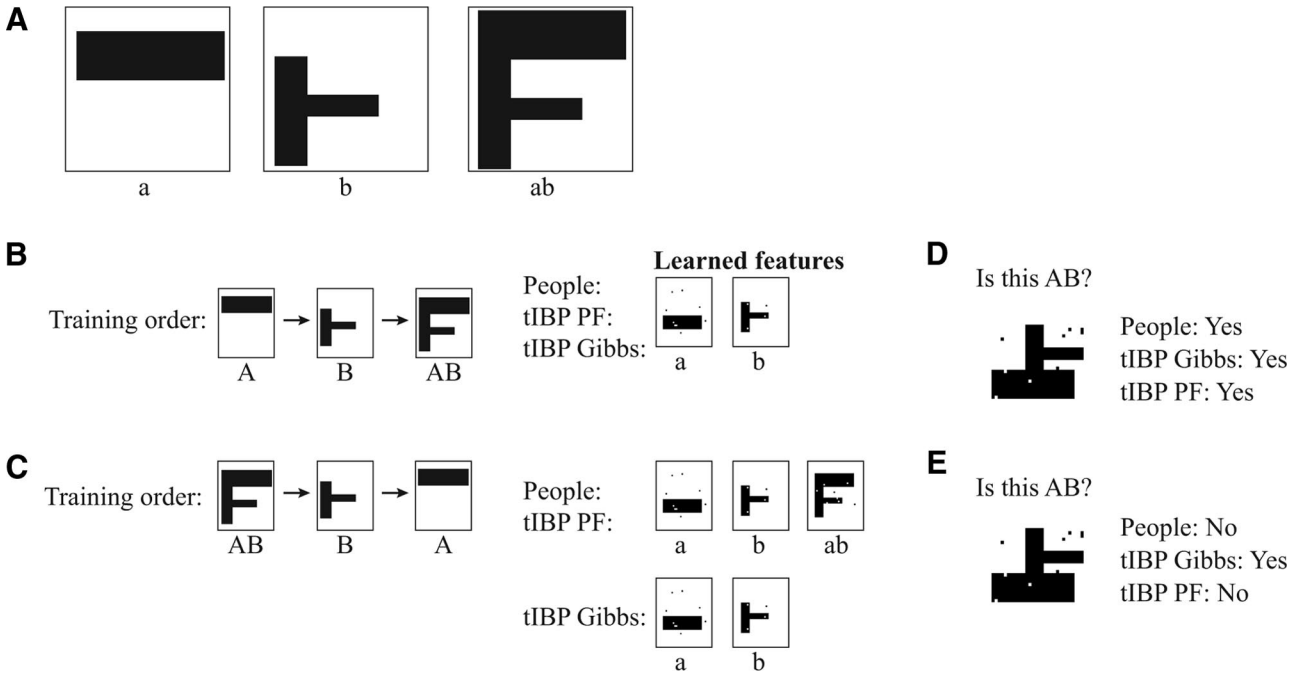

Figure 18. Effects of order on feature learning. The order that categories are observed from categories affects the features inferred and generalization judgments of participants in the experiments of Schyns and Rodet (1997) and the transformed Indian buffet process (tIBP) using particle filtering (PF), but not the tIBP using Gibbs sampling (Gibbs). (A) Three parts used to construct the stimuli of Schyns and Rodet. (B–C) The features learned after observing objects from categories in the *AB* last order and the *AB* first order, respectively. (D–E) Categorizing objects with parts *a* and *b* in a new spatial arrangement as *AB* after observing objects from categories in the *AB* last order and the *AB* first order, respectively.

and Rodet. However, as shown in Figures 18C and 18E, the tIBP using Gibbs sampling is insensitive to the presentation order and incorrectly predicts that only *a* and *b* are learned as features and thus predicts that participants in both conditions should generalize *AB* to objects with *a* and *b* in a new spatial arrangement. In line with our previous discussion, the features learned by the particle filter depend on the order that objects are presented, and so it makes the same predictions as people (see Appendix F for more details). Using particle filtering to perform inference rather than Gibbs sampling results in a model that learns features incrementally, and so it satisfies Criterion 7.

General Discussion

People represent the same object differently depending on the other objects in its context. This is a challenge for many models of cognition, as they typically treat the representation of an object as an immutable property that is intrinsic to the object. In this article, we have presented a computational framework for constructing Bayesian models that infer feature representations for a set of objects. Like people, these models are context sensitive, meaning that they infer different features to represent an object depending on the other objects it is presented with.

Furthermore, we have demonstrated that the proposed computational framework can satisfy the seven criteria we established for a computational explanation of human feature learning. First, we showed that the Criterion 1 (sensory primitives) and Criterion 2 (unlimited features) are satisfied by setting up the computational problem as a problem of matrix factorization and using the IBP as our prior on the feature ownership matrix. Second, we reviewed the results of Austerweil and Griffiths (2011), who used the simplest model in the framework and showed that it infers different features for an image depending on the other images presented in its context (satisfying Criterion 3: context sensitivity). Using a result from Austerweil and Griffiths, we showed that prior expectations of simplicity (representations with fewer features are better) and proximity (adjacent pixels are likely to share the same value) can be included into the model, satisfying Criterion 4 (prior expectations).

Although we showed that previous models developed by Austerweil and Griffiths (2011) that fit in our computational framework satisfy the first four criteria, they did not satisfy the other criteria. We thus defined new models in the computational framework capable of satisfying these criteria. First, we defined a novel model, the tIBP, which inferred transformation-invariant features and incorporated prior knowledge as to how transformations are dependent on each other to explain novel and previous experimental results (satisfying Criterion 5: transformation invariance). Next, we explored two methods for including categorization information into the framework, and defined a novel process, the IBF, to properly account for previous human feature learning results (satisfying Criterion 6: category diagnosticity). Finally, we defined an incremental learner (as a method of inferring the features for the

tIBP), which learned different features depending on the order that objects were presented to it in the same manner as people (satisfying Criterion 7: incremental learning). To the best of our knowledge, this is the first computational framework capable of satisfying all seven criteria.

For the remainder of the article, we address some theoretical questions raised by explaining how people construct representations using Bayesian models, the implications of our results for perceptual object recognition theories, limitations of our approach, and future directions. The explanatory status of Bayesian models for investigating psychological representations is controversial (e.g., Jones & Love, 2011), and so we first directly address the issue of how to interpret the representations inferred by the Bayesian models. Afterward, we discuss how models in our framework fit into classic theories of object recognition. Finally, we consider limitations and directions for future research.

Marr's Levels and the Interpretation of Representations in Bayesian Models

Trying to understand the nature of multiple explanations of the same behavioral phenomenon, Marr (1982) proposed a framework for understanding models of information processing systems at three levels of analysis: the computational level, the algorithmic level, and the implementational level. The computational level explores solutions to the problem the cognitive system is trying to solve and compares this solution to the behavior of the cognitive system. The algorithmic level explores how well an actual process solves the problem and compares the behavior of that process to how people behave. Finally, the implementational level explores how the process described at the algorithmic level is actually instantiated in the world.

The algorithmic level is also called the representational level because the objects manipulated by the hypothesized process are supposed to correspond to psychological representations. Although the representations posited at the algorithmic level are given psychological importance, the representations used at the computational level traditionally are not. This is because computational level solutions are typically used to explain why an agent well adapted to his or her environment should act in a particular way, regardless of the actual algorithm the agent uses. Our analysis of human feature learning gives psychological importance to the representations learned by our computational-level models. What justification do we have in giving psychological importance to these representations?

One cannot specify a problem or formulate a solution at the computational level without first choosing a representation for the problem. The representation chosen at the computational level affects the solution identified by the model and, thus, the behavior of an agent (and in turn, also the algorithmic- and implementational-level explanations). To see this, consider the computational problem of generalization (which underlies many cognitive problems, such as word learning, categorization, and property induction). As argued by Shepard (1987), Bayesian inference solves the problem of generalization at the computational level and matches human (and animal) generalization behavior in a wide range of perceptual domains. However, the solution provided by Bayesian inference depends on how properties in the domain are represented. For example, consider generalization over two dimensions (such as circles varying in the orientation of their radius and size) where properties are assumed to be intervals of the dimensions (contiguous rectangular regions). The ideal solution depends on whether the rectangular regions are aligned with the axes or indifferent to the axes (any orientation of rectangular region is allowed).

Given an appropriate representation of how properties are distributed in a domain, the Bayesian generalization model has also been used to explain human behavior in many conceptual domains (Kemp & Tenenbaum, 2009; Tenenbaum & Griffiths, 2001; Xu & Tenenbaum, 2007). For example, Tenenbaum (2000) found that a Bayesian model endowed with intuitive numerical concepts (such as powers of 2 and numbers between 1 and 10) predicts which other numbers people believe satisfy an unknown mathematical property given one or more numbers that satisfy the property. Importantly, the procedure for predicting generalization behavior in all of the domains (conceptual and perceptual) is the same. The only difference between the models is how hypothetical properties are represented by the model. Thus, the representation used in each particular Bayesian model has psychological importance (different representations explain the differing generalization behaviors in different domains) despite the Bayesian model being at a computational level of analysis.

These examples suggest that Bayesian models can be relevant for understanding how people infer context-sensitive feature representations. For example, consider the debate between fixed and flexible feature representations discussed earlier. As our work is a computational-level analysis, it does not provide a definitive answer to the debate. This is because it cannot distinguish between the processes used to infer different features for the same object (whether people are truly learning new features or simply reweighting preexisting features), as both are valid cognitive processes that might approximate the computational-level solution. However, this does not imply that the analysis leaves us where we started. We believe that our results weakly favor the feature learning interpretation. This is because people must be reweighting the features used to represent an object remarkably often to capture the changes in generalization behavior from the same object in different contexts, if the feature reweighting process is the correct process-level description. Regardless of whether our results are ultimately due to a feature learning or reweighting process, the key principles governing feature learning that we have identified through our approach establish important criteria for evaluating different algorithmic-level accounts. For example, we compared two algorithmic-level accounts, Gibbs sampling and particle filtering, and found that only particle filtering could explain the ordering effects reported by Schyns and Rodet (1997). Clearly more future work is needed to investigate further how feature representations are inferred at the algorithmic level.

Despite this argument, the representations in Bayesian models may not be what many psychologists have typically thought of as psychological representations, since it is unclear whether they are being explicitly manipulated at the process level. In what sense, then, are the representations inferred by the model encoding the observed stimuli? Only further experimenting and more nuanced process-level models will allow us to explore this important question. Regardless of the answer, the proposed computational framework will be a source of useful hypotheses to test, and as a result further our understanding of how the brain and mind *really* construct representations.

The disconnect between representations in Bayesian models and traditional process models is one source of the recent controversy about the utility of Bayesian models (Danks, 2008; Jones & Love, 2011), especially because it is always possible to create a Bayesian model that describes any pattern of human behavior, provided the behavior satisfies basic properties, such as consistency. To define Bayesian models that fit the results of experiments described in this article, we may have needed to include many additional, highly complex assumptions. Remarkably, this was not the case. Many different qualitative behaviors and principles fell out of our framework without the need for superfluous assumptions or ad hoc fixes. Although we did add different modeling components to capture different phenomena, the components were added as a natural by-product of solving different computational problems. When the computational problem changes, the result of a rational analysis changes. Each change to the computational problem that we explored brought us closer to the true problem that people solve, and thus, they were not arbitrary changes to "overfit" the rational analysis. We are not claiming that people are necessarily Bayesian when they form representations and agree that the wide range of behaviors possible to describe with all possible Bayesian models make this kind of argument hazardous. Instead, our goal has been to examine how well we can explain the representations that people form in terms of the statistical structure of the data that they observe. Using Bayesian models has allowed us to perform this analysis, as the models that we defined indicated the representations that follow from the data when combined with transparent assumptions about the quality of different hypothetical representations.

Our perspective is that the representations identified by computational models should be interpreted as those that are the most useful for determining generalization behavior in a particular situation (which can be simple line segments or more complex configurations depending on the context). This is theoretically similar to the "intermediate representations" posited by Biederman and colleagues (Biederman, 1987; Biederman & Cooper, 1991) or, to borrow a term from the categorization literature, "basic-level features" (Pomerantz, 2003). However, further work is needed to explore precisely in what way representations in Bayesian models are explanatory and how their representations relate to the representations typically posited by psychologists.

Connection to Perceptual Theories of Object Representation

In terms of evolutionary adaptability, one of the most important roles of perception is to identify the function of the different objects in the environment. Arguably, the most influential and agreed-upon theory is that the function of an object is inferred through its category membership (Palmer, 1999). The major debates about object representation then concern the nature of the internal representation of object categories and the processes that act on them. Three major proposals are templates, features, and structural descriptions.

Template models assume that object categories are represented by prototypical spatial images. The category with a template that has the maximum overlap with the retinal image is chosen as the currently viewed object's category. Feature theories propose that people represent objects using a set of features. Typically, the features are invariant over different viewpoints and transformations, and so the category of an object can be identified in many situations. Structural description models build on feature models by proposing that object representations are not merely sets of the features, but contain more structured information, such as how those features relate to one another. All three proposals have strengths and weaknesses (see Palmer, 1999, for more details).

In some sense, different aspects of our model bridge the three perceptual proposals. Fundamentally, the model infers features from the raw retinal input and it forms a template of how each feature is instantiated in the image. This is inevitable because only template theories work on the retinal image and not some already prefiltered format. Because our model infers the templates from the images it observed and the templates can be transformed, the type of template model that it is most similar to is a flexible template model. However, it does not use critical points to align the templates to the observed image, and thus, it does not suffer from some of the major criticisms of template models.

On the other hand, our model also decomposes objects into multiple parts (called features), and so it is a feature model as well. This is a desirable property, as there is strong phenomenal (Hoffman & Richards, 1984; Palmer, 1977) and empirical (Biederman & Cooper, 1991; Braunstein et al., 1989; Palmer, 1977) evidence that people represent objects using features. Although our proposal does not yet include relations between features or feature hierarchies (though we are pursuing future work in this direction), we have already demonstrated how more structure can be included into our object representations (beyond just a set of features). We demonstrated how models in the framework can include structural and relational information about how features transform. This incorporates realistic prior expectations of how the features of objects are transformed into a model in our framework, and thus, our framework already has some of the characteristics of structural description representations. Thus, our framework integrates some of the strengths of three of the most popular proposals for object representation in perception.

Limitations and Future Directions

As feature representations for images are one of the most simple representations posited by psychologists, the presented work has several limitations and directions for future research. Although we have focused on inferring feature representations for objects given visual images, and so one might believe our approach is limited to vision, the approach should extend to inferring representations for other modalities given their raw sensory data gracefully. For example, Yildirim and Jacobs (2012) explored an interesting application of our approach to inferring representations with multimodal features. They demonstrated that a simple extension of the IBP can infer a set of features from visual and auditory sensory data. Furthermore, Austerweil and Griffiths (2011) successfully applied the same model used on images to animal concepts as well. Thus, it is clear that our approach is not limited to the visual domain, but rather, we have been using the visual images as a nice test bed for exploring more powerful models in our framework.

In the future, it would be interesting to explore how to extend our framework to incorporate prior expectations about the goodness of different feature images (Gestalt constraints such as good continuation), more detailed category information; to incorporate relations between objects and their features; and to explore what kinds of

features are learned from natural scenes. Including Gestalt constraints into our framework amounts to defining a more elaborate prior distribution on the types of feature images allowed, $P(\mathbf{Y})$. Austerweil and Griffiths (2011) demonstrated one method that we can use to learn more psychologically valid features: defining a prior on feature images that favors adjacent pixels to have the same value. Although this probability distribution does not fully capture human perceptual expectations, it is straightforward to improve the psychological validity of the inferred features by incorporating more complex probability distributions used in computer vision (e.g., Sudderth & Jordan, 2009). In fact, our framework provides a natural method for evaluating different computational proposals of perceptual expectations and to investigate the precise nature of these expectations. Furthermore, the framework predicts that there should be a trade-off between inferring statistically coherent parts and parts satisfying Gestalt constraints as features.

Another future direction is to explore the sorts of features inferred by the model from natural scenes. In fact, this direction is partially motivated from a limitation of our approach. Unfortunately, learning features from natural scenes is not computationally feasible with the inference algorithms described in this article. However, there have been exciting recent developments in machine learning that improve the efficiency of feature inference for the tIBP to allow it to work on images taken from video games (e.g., Mario Bros.), surveillance cameras, and the Microsoft Kinect sensor (Hu, Zhai, Williamson, & Boyd-Graber, 2012; Joho, Tipaldi, Engelhard, Stachniss, & Burgard, 2013). Potentially, these new algorithms could be used to see what kinds of features are learned by the tIBP when given a database of natural scenes. If inference is improved to be able to infer features from natural scenes, we could see how the inferred representations change as the model receives more scenes and compare the changing representations in the model to developmental changes. Furthermore, using more sophisticated inference algorithms could allow us to explore another direction for future research, including compositions of transformations (e.g., features that are first scaled and then translated) and more complex transformations with higher dimensional parameterizations in the model (e.g., affine transformation). Including compositions and more complex transformations increases the number of possible transformations exponentially in the number of transformations and parameters, and so even more sophisticated inference techniques may prove to be necessary.

One last limitation of our work that motivates future work is that it only infers binary features that are subsets of observed images. Fortunately, inferring more complex representations can be addressed with recent advances in machine learning. Machine learning research has already extended the IBP to deal with relations between the observed objects (Miller, Griffiths, & Jordan, 2009), and it is likely that similar techniques could be used to incorporate relations between features (e.g., ON TOP OF). Also, dynamic features that evolve over time could be learned with a time series form of the IBP (Fox, Sudderth, Jordan, & Willsky, 2009; Williamson, Orbanz, & Ghahramani, 2010). Other, more complex features, such as qualitative features (e.g., dashed or dotted lines), may be more difficult to capture using these methods. The most obvious method for developing a model in our framework that can infer more qualitative features would be to use the tIBP where the transformations are defined to relate feature images to each possible quality. For example, if there were image transformations that converted solid contours or regions into different textures, including those in the tIBP would allow it to infer a feature that can be "solid," "dashed," or "dotted." Although this demonstrates the potential of handling qualitative features with the framework, we agree that fleshing out the details is nontrivial, and we do not know of an obvious technique that can be adopted from machine learning to address this limitation. Finally, we are interested in learning features that can occur more than once in an image (where each instantiation has its own transformation parameter). For example, the two vertical bar features in the separate feature representation are treated as completely different features, rather than two instances of the same feature with different transformations. This could be accomplished by extending the tIBP in the same vein as the Poisson-gamma model, an extension of the IBP where that there can be multiple instances of a feature in the same object (Titsias, 2008). It is plausible that a simple extension to this model could allow the IBP to learn the expected number of times a feature should be instantiated in an object and, therefore, be able to learn that there should be two features per object given the separate test objects.

Conclusions

Our capability to form complex representations of our environment and adapt the representations to changes in our environment enables us to navigate and manipulate our environment successfully. As the best artificial systems for solving most cognitive problems still pale in comparison to human ability, our success remains a puzzle. In this article, we explored a computational framework, based on nonparametric Bayesian statistics, that flexibly constructs representations to explain its observations depending on their context. Like people, models in this framework learn features using category and distributional information, and can even infer features invariant over transformations. Although we are still a long way from explaining how art theorists learn to represent Jackson Pollock paintings, our framework provides an important step toward the larger goal of explaining how our brain learns to construct complex representations based on its experiences in the world.

References

Abdi, H., Valentin, D., & Edelman, B. G. (1998). Eigenfeatures as intermediate-level representations: The case for PCA models. *Brain and Behavioral Sciences, 21,* 17–18. doi:10.1017/S0140525X98220103

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85,* 249–277. doi:10.1037/0033-295X.85.4.249

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429. doi:10.1037/0033-295X.98.3.409

Austerweil, J. L., & Griffiths, T. L. (2009). Analyzing human feature learning as nonparametric Bayesian inference. In D. Koller, Y. Bengio, D. Schuurmans, L. Bottou, & A. Curlotta (Eds.), *Advances in neural information processing systems 21* (pp. 97–104). Cambridge, MA: MIT Press.

Austerweil, J. L., & Griffiths, T. L. (2010). Learning invariant features using the transformed Indian buffet process. In J. Lafferty, C. Williams, J. Shawne-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems 23* (pp. 82–90). Cambridge, MA: MIT 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. doi:10.1016/j.cogpsych.2011.08.002

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 11, 629–654. doi:10.1037/0278-7393.11.1-4.629

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York, NY: Wiley. doi:10.1002/9780470316870

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94,* 115–147. doi:10.1037/0033-295X.94.2.115

Biederman, I., & Cooper, E. E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology, 23,* 393–419. doi:10.1016/0010-0285(91)90014-F

Braunstein, M. L., Hoffman, D. D., & Saidpour, A. (1989). Parts of visual objects: An experimental test of the minima rule. *Perception, 18,* 817–826. doi:10.1068/p180817

Carreira-Perpiñán, M. Á. (1997). *A review of dimension reduction techniques* (Tech. Rep. CS-96-09). Sheffield, England: Department of Computer Science, University of Sheffield.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science. *Trends in Cognitive Sciences, 7,* 19–22. doi:10.1016/S1364-6613(02)00005-0

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405. doi:10.1037/0033-295X.104.2.367

Chomsky, N. (1959). [Review of the book *Verbal behavior*, by B. F. Skinner]. *Language, 35,* 26–58. doi:10.2307/411334

Cummins, R. (1989). *Meaning and mental representations*. Cambridge, MA: MIT Press.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 59–75). Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199216093.003.0003

Edelman, S. E. (1999). *Representation and recognition in vision*. Cambridge, MA: MIT Press.

Edelman, S. E., & Intrator, N. (1997). Learning as extraction of low-dimensional representations. In R. L. Goldstone, D. L. Medin, & P. G. Schyns (Eds.), *The Psychology of Learning and Motivation: Vol. 36. Perceptual learning* (pp. 353–380). San Diego, CA: Academic Press. doi:10.1016/S0079-7421(08)60288-1

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2* (pp. 524–532). Los Altos, CA: Morgan Kaufmann.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 499–504. doi:10.1111/1467-9280.00392

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: Statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General, 134,* 521–537. doi:10.1037/0096-3445.134.4.521

Fox, E., Sudderth, E., Jordan, M. I., & Willsky, A. (2009). Sharing features among dynamical systems with beta processes. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 549–557). Cambridge, MA: MIT Press.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers, C-23,* 881–890. doi:10.1109/T-C.1974.224051

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Geisler, W. S. (2003). Ideal observer analysis. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 825–837). Boston, MA: MIT Press.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias–variance dilemma. *Neural Computation, 4,* 1–58. doi:10.1162/neco.1992.4.1.1

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6,* 721–741. doi:10.1109/TPAMI.1984.4767596

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology, 56,* 1–12. doi:10.1016/j.jmp.2011.08.004

Ghahramani, Z. (1995). Factorial learning and the EM algorithm. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems 7* (pp. 617–624). Cambridge, MA: MIT Press.

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York, NY: Appleton-Century-Crofts.

Gluck, M. A., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27,* 166–195. doi:10.1016/0749-596X(88)90072-1

Goldmeier, E. (1972). Similarity in visually perceived forms. *Psychological Issues, 8*(1). (Original work published 1936)

Goldstone, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 86–112. doi:10.1037/0096-1523.26.1.86

Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In R. Kimchi, M. Behrmann, & C. R. Olson (Eds.), *Perceptual organization in vision: Behavioral and neural perspectives* (pp. 233–278). Mahwah, NJ: Erlbaum.

Goldstone, R. L., Gerganov, A., Landy, D., & Roberts, M. E. (2008). Learning to see and conceive. In L. Tommasi, M. A. Peterson, & L. Nadel (Eds.), *Cognitive biology: Evolutionary and developmental perspectives on mind, brain, and behavior* (pp. 163–188). Cambridge, MA: MIT Press.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General, 130,* 116–139. doi:10.1037/0096-3445.130.1.116

Goodman, N. (1972). *Problems and projects*. New York, NY: Bobbs-Merrill.

Gordon, N. J., Salmond, D. J., & Smith, A. F. M. (1993). A novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings on Radar and Signal Processing, 140,* 107–113.

Griffiths, T. L. (2010). Bayesian models as tools for exploring inductive biases. In M. T. Banich & D. Caccamise (Eds.), *Generalization of knowledge: Multidisciplinary perspectives* (pp. 135–156). New York, NY: Psychology Press.

Griffiths, T. L., & Ghahramani, Z. (2005). *Infinite latent feature models and the Indian buffet process* (Tech. Rep. No. 2005-001). London, England: Gatsby Computational Neuroscience Unit, University College London.

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

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 59–100). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511816772.006

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 303–328). Oxford, England: Oxford University Press. doi:10.1093/acprof:oso/9780199216093.003.0014

Griffiths, T. L., Sanborn, A. N., Canini, K. R., Navarro, D. J., & Tenenbaum, J. B. (2011). Nonparametric Bayesian models of category learning. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 173–198). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511921322.008

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334–384. doi:10.1016/j.cogpsych.2005.05.004

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11,* 23–63. doi:10.1111/j.1551-6708.1987.tb00862.x

Hansen, L. K., Ahrendt, P., & Larsen, J. (2005). Towards cognitive component analysis. In T. Honkela, V. Könönen, M. Pöllä, & O. Simula (Eds.), *Proceedings of AKKR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning* (pp. 148–153). Helsinki, Finland: Laboratory of Computer and Information Science, Helsinki University of Technology.

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figural "goodness". *Journal of Experimental Psychology, 46,* 361–364. doi:10.1037/h0055809

Hoffman, D. D., & Richards, W. A. (1984). Parts in recognition. *Cognition, 18,* 65–96. doi:10.1016/0010-0277(84)90022-2

Hu, Y., Zhai, K., Williamson, S., & Boyd-Graber, J. (2012). Modeling images using transformed Indian buffet processes. In J. Langford & J. Pineau (Eds.), *Proceedings of the 29th International Conference on Machine Learning* (pp. 1511–1518). New York, NY: Omnipress.

Joho, D., Tipaldi, G. D., Engelhard, N., Stachniss, C., & Burgard, W. (2013). Unsupervised scene analysis and reconstruction using nonparametric Bayesian models. In N. Roy, P. Newman, & S. Srinivasa (Eds.), *Robotics: Science and systems VIII* (pp. 161–168). Cambridge, MA: MIT Press.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34,* 169–188. doi:10.1017/S0140525X10003134

Jordan, M. I. (2010). Bayesian nonparametric learning: Expressive priors for intelligent systems. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Heuristics, probability and causality: A tribute to Judea Pearl* (pp. 167–186). London, England: College Publications.

Kanizsa, G. (1979). *Organization in vision: Essays on Gestalt perception*. New York, NY: Praeger.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116,* 20–58. doi:10.1037/a0014282

Krumhansl, C. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review, 85,* 445–463. doi:10.1037/0033-295X.85.5.445

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44. doi:10.1037/0033-295X.99.1.22

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review, 113,* 677–699. doi:10.1037/0033-295X.113.4.677

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review, 9,* 43–58. doi:10.3758/BF03196256

Lin, E. L., & Murphy, G. L. (1997). Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 1153–1169. doi:10.1037/0096-1523.23.4.1153

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55,* 232–257. doi:10.1016/j.cogpsych.2006.09.006

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111,* 309–332. doi:10.1037/0033-295X.111.2.309

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: Wiley.

Mach, E. (1959). *The analysis of sensations*. Chicago, IL: Open Court. (Original work published 1897).

Markman, A. B. (1998). *Knowledge representation*. Hillsdale, NJ: Erlbaum.

Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B: Biological Sciences, 200,* 269–294. doi:10.1098/rspb.1978.0020

McCarthy, J. D., Cordeiro, D., & Caplovitz, G. P. (2012). Local form–motion interactions influence global form perception. *Attention, Perception, & Psychophysics, 74,* 816–823. doi:10.3758/s13414-012-0307-y

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254–278. doi:10.1037/0033-295X.100.2.254

Miller, K. T., Griffiths, T. L., & Jordan, M. I. (2009). Nonparametric latent feature models for link prediction. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 1276–1284). Cambridge, MA: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316. doi:10.1037/0033-295X.92.3.289

Navarro, D. J., & Perfors, A. F. (2010). Similarity, feature discovery, and the size principle. *Acta Psychologica, 133,* 256–268. doi:10.1016/j.actpsy.2009.10.008

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics, 9,* 249–265.

Neisser, U. (1967). *Cognitive psychology*. New York, NY: Appleton-Century-Crofts.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114. doi:10.1037/0278-7393.10.1.104

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57. doi:10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 3–27. doi:10.1037/0096-1523.17.1.3

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 18–39). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511921322.002

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition*. Oxford, England: Oxford University Press.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 2745–2750. doi:10.1073/pnas.0708424105

Palmer, S. E. (1975). Visual perception and world knowledge: Notes on a model of sensory–cognitive interaction. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition* (pp. 279–307). San Francisco, CA: Freeman.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology, 9,* 441–474. doi:10.1016/0010-0285(77)90016-0

Palmer, S. E. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 250–303). Hillsdale, NJ: Erlbaum.

Palmer, S. E. (1983). The psychology of perceptual organization: A transformational approach. In J. Beck, B. Hope, & A. Rosenfeld (Eds.), *Human and machine vision* (pp. 269–339). New York, NY: Academic Press.

Palmer, S. E. (1989). Reference frames in the perception of shape and

orientation. In B. E. Shepp & S. Ballesteros (Eds.), *Object perception: Structure and process* (pp. 121–163). Hillsdale, NJ: Erlbaum.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Pevtzow, R., & Goldstone, R. L. (1994). Categorization and the parsing of objects. In A. Ram & K. Eiselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 717–722). Hillsdale, NJ: Erlbaum.

Pitman, J. (2002). *Combinatorial stochastic processes* (Tech. Rep. No. 621). Berkeley: Department of Statistics, University of California.

Pitt, D. (2008). Mental representation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from http://plato.stanford.edu/archives/fall2008/entries/mental-representation/

Pomerantz, J. R. (2003). Wholes, holes, and basic features in vision. *Trends in Cognitive Sciences, 7,* 471–473. doi:10.1016/j.tics.2003.09.007

Pothos, E. M., & Bailey, T. M. (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the generalized context model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1062–1080. doi:10.1037/a0015903

Rock, I. (1973). *Orientation and form*. New York, NY: Macmillan.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science, 9,* 75–112. doi:10.1207/s15516709cog0901_5

Rust, N. C., & Stocker, A. A. (2010). Ambiguity and invariance: Two fundamental challenges for visual processing. *Current Opinion in Neurobiology, 20,* 382–388. doi:10.1016/j.conb.2010.04.013

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117,* 1144–1167. doi:10.1037/a0020511

Schyns, P. G., Goldstone, R. L., & Thibaut, J. (1998). Development of features in object concepts. *Behavioral and Brain Sciences, 21,* 1–17. doi:10.1017/S0140525X98000107

Schyns, P. G., & Murphy, G. (1994). The ontogeny of part representation in object concepts. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 31, p.305–349). San Diego, CA: Academic Press. doi:10.1016/S0079-7421(08)60413-2

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 681–696. doi:10.1037/0278-7393.23.3.681

Selfridge, O. G. (1955). Pattern recognition and modern computers. In *Proceedings of the March 1–3, 1955, Western Joint Computer Conference* (pp. 91–93). New York, NY: Association for Computing Machinery. doi:10.1145/1455292.1455310

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science, 237,* 1317–1323. doi:10.1126/science.3629243

Shiffrin, R. M., & Lightfoot, N. (1997). Perceptual learning of alphanumeric-like characters. In R. L. Goldstone, D. L. Medin, & P. G. Schyns (Eds.), *The Psychology of Learning and Motivation: Vol. 36. Perceptual learning* (pp. 45–81). San Diego, CA: Academic Press. doi:10.1016/S0079-7421(08)60281-9

Shultz, T. R., & Sirois, S. (2008). Computational models of developmental psychology. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 451–476). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511816772.020

Smith, L. B. (2005). Shape: A developmental product. In L. Carlson & E. van der Zee (Eds.), *Functional features in language and space: Insights from perception, categorization, and development* (pp. 235–255). Oxford, England: Oxford University Press. doi:10.1093/acprof:oso/9780199264339.003.0016

Spratling, M. W. (2006). Learning image components for object recognition. *Journal of Machine Learning Research, 7,* 793–815.

Sudderth, E., & Jordan, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman–Yor processes. In D. Koller, Y. Bengio, D. Schuurmans, L. Bottou, & A. Curlotta (Eds.), *Advances in neural information processing systems 21* (pp. 1585–1592). Cambridge, MA: MIT Press.

Teh, Y. W., & Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In N. L. Hjort, C. Holmes, P. Müller, & S. G. Walker (Eds.), *Bayesian nonparametrics* (pp. 158–207). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511802478.006

Teh, Y. W., Jordan, M. I., Beal, M., & Blei, D. (2004). Sharing clusters among related groups: Hierarchical Dirichlet processes. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems 17* (pp. 1385–1392). Cambridge, MA: MIT Press.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Advances in neural information processing systems 12* (pp. 59–65). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–640. doi:10.1017/S0140525X01000061

Titsias, M. (2008). The infinite gamma–Poisson feature model. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems 20* (pp. 1513–1520). Cambridge, MA: MIT Press.

Trueblood, J. S., & Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. *Cognitive Science, 35,* 1518–1552. doi:10.1111/j.1551-6709.2011.01197.x

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352. doi:10.1037/0033-295X.84.4.327

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review, 19,* 1047–1056. doi:10.3758/s13423-012-0300-4

Williamson, S., Orbanz, P., & Ghahramani, Z. (2010). Dependent Indian buffet processes. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* (pp. 924–931). Retrieved from http://machinelearning.wustl.edu/mlpapers/paper_files/AISTATS2010_WilliamsonOG10.pdf

Wood, F., & Griffiths, T. L. (2007). Particle filtering for nonparametric Bayesian matrix factorization. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (pp. 1513–1520). Cambridge, MA: MIT Press.

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 Twenty-Second Conference on Uncertainty in Artificial Intelligence* (pp. 536–543). Arlington, VA: AUAI Press.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114,* 245–272. doi:10.1037/0033-295X.114.2.245

Yildirim, I., & Jacobs, R. A. (2012). A rational analysis of the acquisition of multisensory representations. *Cognitive Science, 36,* 305–332. doi:10.1111/j.1551-6709.2011.01216.x

Zeigenfuse, M. D., & Lee, M. D. (2010). Finding the features that represent stimuli. *Acta Psychologica, 133,* 283–295. doi:10.1016/j.actpsy.2009.07.014

Appendix A

Inference Using Gibbs Sampling

Throughout the article, we have presented models by describing a step-by-step probabilistic process, which first creates the latent variables (e.g., a feature ownership matrix) and results in generating the set of images observed by the model. Inference is the inverse of this process, where we start with the set of images and work backward to compute the latent variables (in our case, feature representations) that were likely to have produced the observations. We can formulate new predictions by seeing which observations are likely to be generated from the feature representation that we inferred.

For the simplest model in the framework, we represent objects using two matrices: a feature ownership matrix $\mathbf{Z}$, which encodes each object's features, and a feature image matrix, which encodes the image associated with each feature. Given a set of images $\mathbf{X}$, we use Bayes' theorem to infer the most probable feature representation to have produced the images:

$$P(\mathbf{Z}, \mathbf{Y} \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z})P(\mathbf{Y})P(\mathbf{Z})}{\sum_{\mathbf{Z}', \mathbf{Y}'} P(\mathbf{X}' \mid \mathbf{Y}', \mathbf{Z}')P(\mathbf{Y}')P(\mathbf{Z}')}. \quad \text{(A1)}$$

Unfortunately, there is no known solution to the sum in the denominator. So, we approximate the probability distribution over feature representations. As the most probable a posteriori feature representation is usually much more likely than other possible feature representations, we typically only report it (as the other potential feature representations have negligible probability). One exception is the simulations in Appendix C, where we marginalize over samples due to the substantial uncertainty in the covariance matrix generating translation and scaling transformations.

To infer the feature representation with maximum posterior probability, we use a combination of Gibbs sampling and simulated annealing (Geman & Geman, 1984). Gibbs sampling proceeds by first randomly initializing all of the unobserved variables,[26] cycling through the elements of each matrix (e.g., z_{nk}) and sampling new values for each element conditioned on the values of all of the other variables (e.g., the elements of $\mathbf{Z}$ that are not z_{nk}, $\mathbf{Y}$, and $\mathbf{X}$). For our applications, we first resample the feature ownership matrix; we cycle through the objects 1 to N and, for each object n, resample its feature ownership assignments for features that are owned by at least one object without counting object n (resample all z_{nk} for all k such that for at least one object $i \neq n$, $z_{ik} = 1$). The Gibbs sampler for elements of the feature ownership matrix is given by

$$P(z_{nk} = 1 \mid \mathbf{Z}_{-(nk)}, \mathbf{Y}, \mathbf{X}) \propto P(z_{nk} \mid \mathbf{Z}_{-(nk)})P(\mathbf{x}_n \mid \mathbf{z}_n, \mathbf{Y}) \quad \text{(A2)}$$

$$= \frac{m_{-(nk)}}{N} P(\mathbf{x}_n \mid \mathbf{z}_n, \mathbf{Y}), \quad \text{(A3)}$$

where $\mathbf{Z}_{-(nk)}$ is the matrix of current feature ownership assignments without z_{nk}; $m_{-(nk)}$ is the number of objects that own feature k without considering object n; and $p(\mathbf{X}|\mathbf{Z}, \mathbf{Y})$ is the likelihood of the observed objects under the feature representation that includes object n owning feature k. The first term is derived by assuming object n to be the last customer in the Indian buffet process (IBP) and using the culinary metaphor, which is valid by the exchangeability of the IBP. After cycling through the previously owned features, we sample K_{new} features from

$$P(K_{\text{new}} \mid \mathbf{x}_n, \mathbf{z}_n, \mathbf{Y}) \propto P(\mathbf{x}_n \mid \mathbf{z}_n, \mathbf{Y}, K_{\text{new}})P(K_{\text{new}}) \quad \text{(A4)}$$

$$\propto P(K_{\text{new}}) \prod_{d=1}^{D} \left[1 - (1 - \varepsilon)(1 - \lambda)^{\mathbf{z}_n \mathbf{y}_d}(1 - \lambda\theta)^{K_{\text{new}}}\right], \quad \text{(A5)}$$

where $P(K_{\text{new}})$ is Poisson distributed with parameter $\frac{\alpha}{N}$, ε is the probability that a pixel is on by chance, λ is the probability that an object's feature with the pixel on in its image fails to turn on that pixel in the object's image, and θ is the prior probability that a pixel is on in a feature image. We approximate this distribution by calculating it explicitly for each K_{new} from 0 to a cutoff at 10, which is a good approximation because $P(K_{\text{new}})$ decays to 0 quickly. When new features are sampled ($K_{\text{new}} > 0$), K_{new} column vectors are appended to $\mathbf{Z}$. The column vectors contain all zeros except for a 1 at index n. We sample the images of the new features jointly by defining an auxiliary variable h_d, which denotes the number of the new features that have pixel d on and sample it from

$$P(h_d \mid x_{n,d}, \mathbf{z}_n, \mathbf{Y}, K_{new}) \propto (h_d \mid K_{new})P(x_{n,d} \mid \mathbf{z}_n, \mathbf{Y}, h_d) \quad \text{(A6)}$$

$$= \begin{cases} \binom{K_{\text{new}}}{h_d} \theta^{h_d}(1 - \theta)^{K_{\text{new}} - h_d}(1 - \varepsilon)(1 - \lambda)^{\mathbf{z}_n \mathbf{Y} h_d} & \text{if } x_{nd} = 0 \\ \binom{K_{\text{new}}}{h_d} \theta^{h_d}(1 - \theta)^{K_{\text{new}} - h_d}\left(1 - (1 - \varepsilon)(1 - \lambda)^{\mathbf{z}_n \mathbf{Y} h_d}\right) & \text{if } x_{nd} = 1 \end{cases} \quad \text{(A7)}$$

where $\mathbf{z}_n$ is the vector of feature ownership for object n before adding these new features. The Gibbs sampler for binary feature

[26] Although in theory the initialized values should not matter, in practice some initializations can result in the algorithm getting "stuck" in local minima. Thus, we recommend trying a number of random initializations in practice to ensure that the results of any particular run are not due to the initialization.

(*Appendices continue*)

images using the noisy-OR likelihood and the knowledge-less feature image prior is

$$P\left(y_{kd}\middle|\mathbf{Z},\mathbf{Y}_{-(kd)},\mathbf{X}\right)\propto\theta^{y_{kd}}(1-\theta)^{y_{kd}}P\left(\mathbf{X}\middle|\mathbf{Z},\mathbf{Y}\right). \quad \text{(A8)}$$

We refer the reader to Griffiths and Ghahramani (2011) for a general derivation of the Gibbs sampler for the IBP and Wood et al. (2006) for the derivation specific to noisy-OR likelihood function.

After a number of iterations of resampling the elements of each matrix with the above Gibbs sampler, the distribution over resampled elements approximates the true posterior distribution. Following Austerweil and Griffiths (2011), we use simulated annealing to aid inference, which samples from $P(\mathbf{Z}|\mathbf{X},\mathbf{Y})^{1/t}$, and use a logarithmic cooling schedule ($t = \min(1, \log(i + 1)/T)$, where i is the iteration number and T depends on the dimensionality of $\mathbf{X}$). This helps the Gibbs sampler get out of bad local minima early on by sampling more uniformly at first and results in $P(\mathbf{Z}|\mathbf{X},\mathbf{Y})^{1/t}$ still converging to $P(\mathbf{Z}|\mathbf{X},\mathbf{Y})$ as $t \to 1$.

Given an estimate of the most probable feature representation from an inference procedure, $\hat{\mathbf{Z}}$ and $\hat{\mathbf{Y}}$, the model predictions for what other images, $\mathbf{x}_{\text{test}}$, are probable is given by

$$P(\mathbf{x}_{\text{test}}|\mathbf{X}) \approx P(\mathbf{x}_{\text{test}}|\hat{\mathbf{z}}_{\text{test}},\ \hat{\mathbf{Y}}), \quad \text{(A9)}$$

where $\hat{\mathbf{z}}_{\text{test}} = \arg\max_{\mathbf{z}} P\left(\mathbf{x}_{\text{test}}\middle|\mathbf{z},\hat{\mathbf{Y}}\right)$ denotes the most probable features for representing $\mathbf{x}_{\text{test}}$ according to the model. We find $\hat{\mathbf{z}}_{\text{test}}$ by sampling it under the constraint that only previously observed features can be used (new features are not allowed to be constructed). Finally, we use the exponentiated Luce choice rule, with response probabilities proportional to $\left|P\left(\mathbf{x}_{\text{test}}\middle|\mathbf{z}_{\text{test}},\hat{\mathbf{Y}}\right)\right|^{\phi}$ (Kruschke, 1992; Luce, 1959) to put the unnormalized probabilities on a scale similar to the results of the participants in our experiments. ϕ is fit by minimizing the square distance between human and model responses, and then normalizing the distribution by the different test objects. Although the models contain several parameters, this is the only one that is ever fit to participant responses.

Appendix B

The Transformed Indian Buffet Process

The transformed Indian buffet process (tIBP) is defined by the following generative process:

$$\begin{aligned}\mathbf{Z}|\alpha &\sim \text{IBP}(\alpha) & r_{nk}|\eta &\overset{iid}{\sim} \Phi(\eta)\\ \mathbf{Y}|\theta &\sim g(\theta) & \mathbf{x}_n|\mathbf{r}_n,\ \mathbf{z}_n,\mathbf{Y},\ \gamma &\sim f(\mathbf{x}_n|\mathbf{r}_n(\mathbf{Y}),\ \mathbf{z}_n,\ \gamma),\end{aligned}$$

where $\Phi(\eta)$ is a distribution over the set of transformations (parameterized by η), $g(\theta)$ is the feature image prior (in this article, independent coin flips with bias θ as in the standard IBP), $\boldsymbol{r}_n$ is the vector of transformations for object n's features, $f(\mathbf{x}_n|\mathbf{r}_n(\mathbf{Y}),\mathbf{z}_n,\gamma)$ is the distribution generating the images given the feature representation (the noisy-OR distribution for this article), where $\mathbf{r}_n(\mathbf{Y})$ is the matrix resulting from applying transformation r_{nk} to each feature k and γ is the set of parameters for that distribution (e.g., $\gamma = [\varepsilon, \lambda]$ for the noisy-OR), and the other variables are defined as in the standard Indian buffet process (IBP). Inference using Gibbs sampling is computed in the same manner as the standard IBP, except that the transformation applied to feature k specific to object n, r_{nk}, needs to be inferred as well. Additionally, the current transformations are applied to the feature images when inferring $\mathbf{Z}$ and $\mathbf{Y}$. The model predictions are given by

$$P(\mathbf{x}_{\text{test}}|\mathbf{X}) \approx P(\mathbf{x}_{\text{test}}|\hat{\mathbf{z}}_{\text{test}},\ \hat{\mathbf{r}}_{\text{test}}(\hat{\mathbf{Y}})), \quad \text{(B1)}$$

where $(\hat{\mathbf{z}}_{\text{test}}, \hat{\mathbf{r}}_{\text{test}}) = \arg\max_{(\mathbf{z},\mathbf{r})} P\left(\mathbf{x}_{\text{test}}\middle|\mathbf{z},\mathbf{r}(\hat{\mathbf{Y}})\right)$. As before, model predictions are compared to experiment results by transforming them through the exponentiated Luce choice rule, and ϕ is fit by minimizing the squared distance between the human results and the transformed model predictions.

Except for the simulations reported in the beyond transformation independence and learning which transformations apply subsections, we used a uniform distribution on all possible translations as our prior on feature transformations ($r_{nk} \sim U(\{1, \ldots, D_1\} \times \{0, \ldots, D_2\})$, where D_1 and D_2 are the size of the dimensions of the image and r_{nk} is a parameter of the transformation specifying how much to translate the image right and down). We assumed the image was a torus, so any parts of the image that are transformed to go beyond the dimensions are wrapped back to the beginning. Unless otherwise stated, we use Gibbs sampling for inference, drawing samples from $P(\mathbf{Z},\mathbf{Y},\mathbf{R}|\mathbf{X})$.

For the tIBP, Gibbs sampling proceeds in the following manner. We first resample z_{nk} for all currently used features ($m_k > 0$ without counting z_{nk}) by summing over all the possible transformations (thus avoiding having to get "lucky" in randomly sampling the appropriate transformation). This can be done explicitly, with

$$p(z_{nk}|\mathbf{Z}_{-(nk)},\mathbf{R}_{-(nk)},\ \mathbf{Y},\mathbf{X}) = \sum_{r_{nk}} p(z_{nk}|\mathbf{Z}_{-(nk)},\ \mathbf{R},\mathbf{Y},\mathbf{X})p(r_{nk}) \quad \text{(B2)}$$

$$\propto \sum_{r_{nk}} p(\mathbf{x}_n|\mathbf{z}_n,\ \mathbf{r}_n(\mathbf{Y}))p(z_{nk}|\mathbf{Z}_{-(nk)})p(r_{nk}), \quad \text{(B3)}$$

(*Appendices continue*)

where the first term in the sum is the given by the noisy-OR likelihood, the second term is given by the IBP culinary metaphor, and the last term is the prior over transformations. If we draw $z_{nk} = 1$, then we also sample draw r_{nk} from

$$p(r_{nk}|z_{nk} = 1, \mathbf{Z}_{-(nk)}, \mathbf{R}_{-(nk)}, \mathbf{Y}, \mathbf{X}) \propto p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{Y}, \mathbf{R})p(r_{nk}). \quad \text{(B4)}$$

Note that the probabilities involving transformations over images can be stored in memory for efficiency, and thus, they only need to be computed once per (n,k) pair per each iteration of the Gibbs sampler.

Second, we sample the number of new features. To derive this sampler, we adapt the technique described above and derived in Wood et al. (2006), which yields

$$p(K_n^{\text{new}}|\mathbf{x}_n, \mathbf{Z}_{n,1:(K+K_n^{\text{new}})}, \mathbf{Y}, \mathbf{R}) \propto p(\mathbf{x}_n|\mathbf{Z}^{\text{new}}, \mathbf{Y}, K_n^{\text{new}})\ P(K_n^{\text{new}}), \quad \text{(B5)}$$

where $\mathbf{Z}^{\text{new}}$ includes the K_n^{new} extra columns for the new features (which is 1 for row n, but 0 otherwise). To compute the first term, $p(\mathbf{x}_n|\mathbf{Z}^{\text{new}}, \mathbf{Y}, K_n^{\text{new}})$, we need to sum over possible feature images and transformations. To simplify this calculation, we assume that the first time a feature is sampled, it is not transformed (this is justified because the parameters of one of the transformations of a feature and its image are statistically unidentifiable). Without transformations, this is equivalent to summing over the possible feature images in the noisy-OR IBP, which was derived by Wood et al. to be

$$p(K_n^{\text{new}}|\mathbf{x}_n, \mathbf{Z}_{n,1:(K+K_n^{\text{new}})}, \mathbf{Y}, \mathbf{R}) \propto \frac{\alpha^{K_n^{\text{new}}} e^{-\alpha}}{K_n^{\text{new}}!} (1 - (1 - \varepsilon)(1 - \lambda)^{\mathbf{z}_n \mathbf{r}_n(\mathbf{y}_d)}(1 - \theta\lambda)^{K_n^{\text{new}}}), \quad \text{(B6)}$$

where $\mathbf{r}_n(\mathbf{y}_d)$ is the value of pixel d for all the features after object n's transformations are applied. Finally, $\mathbf{Y}$ is sampled as before, except the feature images for each object are transformed by $\mathbf{R}$.

Given T samples $(\mathbf{Z}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{R}^{(t)})^T_{t=1}$ from $P(\mathbf{Z}, \mathbf{Y}, \mathbf{R}|\mathbf{X})$, an approximation to the posterior probability of new images is given by

$$P(\mathbf{x}_{N+1}|\mathbf{X}) \approx \frac{1}{T}\sum_{t=1}^{T} P(\mathbf{x}_{N+1}|\mathbf{Z}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{R}^{(t)}). \quad \text{(B7)}$$

To compute the term of the sum at every t, we marginalize over $\mathbf{z}_{N+1}$ and $\mathbf{r}_{N+1}$, the features of $\mathbf{x}_{N+1}$ and their transformations.

For the simulations described in the Models Learning Invariant Versus "Variant" Features, Learning Spatially Invariant Features, and One or Two Features? sections, we approximate the predictive distribution, $P(\mathbf{x}_{N+1}|\mathbf{X})$, given by Equation B7 with a single sample from the Gibbs sampler, the sample whose feature representation has the largest posterior probability. The Gibbs sampler was run for 1,000 iterations. For the simulations described in the Models Learning Invariant Versus "Variant" Features section, the parameter values were set to the following values: $\alpha = .8$, $\varepsilon = 0.05$, $\lambda = 0.99$, and $\theta = 0.4$. Both the tIBP and IBP use these same parameter values for the simulations reported in the Models Learning Invariant Versus "Variant" Features section, and the images were 5×5 pixels. For the simulations described in the One or Two Features? section, the parameter values were set to the following values: $\alpha = .8$, $\varepsilon = 0.01$, $\lambda = 0.99$, and $\theta = 0.175$. To relate $P(\mathbf{x}_{N+1}|\mathbf{X})$ to the experimental results in the One or Two Features? section, we transform it using the exponentiated Luce choice rule with parameter $\phi = 0.05$.

For the simulations described in the Learning Spatially Invariant Features section, the following parameter values were used: $\alpha = .8$, $\varepsilon = 0.01$, $\lambda = 0.999$, and $\theta = 0.15$. The starting temperature for simulated annealing, T was initialized to 5. Due to the complexity of inference when there are hundreds of scenes, we interleaved a split–merge step after every five steps of the Gibbs sampler. In a split–merge step, a new feature representation was proposed that either merged two random features together with probability .5 or split a feature into two new features with probability .5. A merge step was performed by first picking two features at random, combining their feature ownership values (by taking their union), and then sampling a single feature image from scratch. A split step was performed by first picking a random feature to split into two new features, then randomly distributing the objects that owned the original feature to one of the two new features and finally sampling the feature image from scratch. The proposed feature representation was automatically accepted if it had greater likelihood than the likelihood of the current feature representation. Otherwise, it was accepted with probability proportional to the ratio of the proposed feature representation's log-likelihood to the log-likelihood of the current feature representation with a fixed annealing temperature of 1/150. Because this split–merge technique does not provide proper samples from the posterior, this amounts to performing stochastic search for the feature representation with maximum probability rather than approximating the posterior probability over a set of samples.

(*Appendices continue*)

Appendix C

How Variance in Translations Affects Variance in Scalings

To capture participant responses in Experiment 3 of Smith (2005), we couple the variance of translations and scalings by putting a prior on each transformation type that shares the same covariance matrix Σ. The resulting generative process is

$$\begin{aligned} \Sigma &\sim IW(2.5,\ 5\mathbf{I}) \\ \mu_T &\sim N\left(\mu_{T0}, \Sigma / k_0\right) \\ w_{nk} &\overset{iid}{\sim} \text{Bernoulli}(1/2) \\ r_{nk} \,\big|\, w_{nk} = 0 &\sim N\left(\mu_S,\ \Sigma\right) \\ r_{nk} \,\big|\, w_{nk} = 1 &\sim N\left(\mu_T,\ \Sigma\right), \end{aligned}$$

where $\mathbf{I}$ is the identity matrix and IW is the inverse-Wishart distribution. Features and transformations are sampled as before with Gibbs sampling. The parameters for the transformations (μ_T and Σ) are inferred using the conjugate updates for the normal inverse-Wishart model (Bernardo & Smith, 1994). μ_S is not inferred because there are no scalings in the training set to base the inference on. The images were 25 × 25 pixels. The set of training images given to the model consisted of a centered 7 × 7 pixel square, which occurred in every possible horizontal or vertical (depending on the condition) translation. The test images were centered rectangles with one side of length 7 and the other side had length 9, 11, or 5 (corresponding to 1, 2, or −). So, the Horizontal 1 test object (H1) had a width of length 9 and height of length 7.

The parameters α, ε, λ, and θ were set to 2, 0.005, 0.995, and 0.25, respectively, and the Gibbs sampler was run for 1,000 iterations. k_0 was set to 40 and μ_S was set to $[1.2, 1.2]^T$. μ_{T0} was set to $[25, 25]^T$. The posterior distribution over r_{nk} and Σ is fairly broad. So rather than use the maximum a posteriori estimate of the feature representation to form our predictions (as we do for most other simulations), the predictions were made by averaging over the samples using a burn-in of 200 and taking every one out of 10 samples (a procedure called *thinning*). The exponentiated Luce rule with parameter $\phi = 0.031$ was used, except the unnormalized probabilities were not renormalized after exponentiating. Simulated annealing was not used for these simulations.

Appendix D

Learning Invariance Type Using the Transformed Indian Buffet Process

To learn the ways a feature can be transformed, we add a latent indicator for each feature that denotes the types of transformations it is allowed to undergo. Let t_k be a binary indicator for feature k, where $t_k = 1$ indicates the feature is rotated a random number of degrees (uniform between 0° and 45° in steps of 15°) and $t_k = 0$ indicates the feature is scaled by a random amount (uniformly drawn from {3/8, 7/8, 3/5, 5/7, 1, 7/5, 11/7, 5/3, 11/5, 7/3, 11/3}). Finally, for the purpose of inference, we assume that $t_k \overset{iid}{\sim} \text{Bernoulli}(0.5)$.

Inference and prediction are performed in the same manner as the normal transformed Indian buffet process except that t_k needs to be sampled as well and predictions are made conditioned on the type of transformation the feature is allowed to undergo. A Gibbs sampler for t_k is given by

$$p(t_k|\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{R}_{-\mathbf{k}}, \mathbf{t}_{-k}) \propto \sum_{n=1}^{N} p(\mathbf{x}_n|r_{nk}, t_k, \mathbf{Y}, \mathbf{Z}, \mathbf{R}_{-k}, \mathbf{t}_{-k}) p(\mathbf{r}_k|t_k) p(t_k). \tag{D1}$$

The parameters were set to $\alpha = 2$, $\varepsilon = 0.01$, $\lambda = 0.99$, $\theta = 0.5$, and $\pi = 0.5$ (with $\phi = 0.005$ for the exponentiated Luce choice rule). The predictions shown were made after running the Gibbs sampler for 1,000 iterations, and the images were 38 × 38 pixels each. Simulated annealing was not used for these simulations.

Appendix E

Incorporating Category Information

In the article, we presented two methods for incorporating categorization information into our framework: as part of the observable properties of an object and representing it directly through the Indian buffet franchise (IBF). For the first method, we appended $c = 35$ bits per category to each image. So, each image had $2c = 70$ extra bits. To encode that an image was in Category 1, the first c extra bits were set to 1 and the last c bits were set to 0 (and vice versa for Category 2). Except for setting $\theta = 0.25$, we used the same parameter values as used by Austerweil and Griffiths (2011): $\alpha = 2$, $\varepsilon = 0.01$, and $\lambda = 0.99$. The features with largest posterior probability were found with a Gibbs sampler (including simulated annealing with $T = 5$ and a split–merge step

(*Appendices continue*)

every six iterations) and are shown in Figures 16A and 16B. The sampler was run for 1,000 iterations, and it was given eight examples from each category ($N = 32$).

Inference for the IBF proceeds as described in Appendix A, except that when the IBP creates new features, it draws the images for those features from the shared Chinese restaurant process (CRP). We refer the reader to Neal (2000) for a thorough description of inference for the CRP. Additionally, Gibbs updates for $y^{(0)}{}_{kd}$ are performed as described by Wood et al. (2006), except that all images (regardless of their category) that have a feature whose image is assigned to $\mathbf{y}^{(0)}{}_{k}$ should be included in the product over likelihood terms. Note that when a pixel of $y^{(0)}{}_{kd}$ is resampled, the corresponding $y^{(a)}{}_{k'd}$ should be set to the sampled value of $y^{(0)}{}_{kd}$ for the image of any category image k' that is assigned to feature image k of the CRP. The feature images of the individual categories otherwise do not change from iteration to iteration of inference (i.e., they are deterministically set to the appropriate image in the shared feature image matrix $\mathbf{Y}^{(0)}$). The results described in the article use parameter values: $\alpha = 2$, $\varepsilon = 0.025$, $\lambda = 0.975$, $\theta = 0.15$, and $\beta = 0.0001$. The sampler was run for 1,000 iterations, and simulated annealing was used ($T = 20$). A split–merge step was completed every six iterations, which included within-category split and merge proposals, and between-category proposals, where a random feature of one category takes the feature image from another random feature.

Appendix F

Learning Features for the Transformed Indian Buffet Process Incrementally

To explain the feature learning order effects found by Schyns and Rodet (1997), we approximate features under the transformed Indian buffet process (tIBP) model using an incremental learning algorithm, the particle filter. The particle filter approximates the posterior distribution over features (ownership matrices $\mathbf{Z}$, images $\mathbf{Y}$, and transformations $\mathbf{R}$) by forming P "particles" (samples that store the inferred feature representations of objects observed so far) and updating the particles sequentially as more objects are observed. Rather than infer the features for all of the objects from scratch each time a new object is observed (as the Gibbs sampler would if it were given objects sequentially), the particle filtering algorithm infers features for the new object (potentially updating the feature images and transformations used by previous objects), but keeps the assignments for the previous objects fixed. When the number of particles is large, this yields the same posterior distribution as Gibbs sampling (in fact, both converge to the true distribution in the limit), but for a small number of particles the inferred posterior distribution depends on the object presentation order.

Formally, let $\mathbf{X}^{(n)}$ be objects 1 to n and $\mathbf{Z}^{(n)}$ be the feature ownership matrix inferred after observing n objects (the feature images and transformations are still updated using Gibbs sampling). The posterior distribution after observing n objects can be decomposed recursively with the features inferred after observing $n - 1$ objects as shown in Equation F1, with

$$P(\mathbf{Z}^{(n)}, \mathbf{Y}, \mathbf{R} | \mathbf{X}^{(n)}) = \sum_{\mathbf{Z}^{(n-1)}} P(\mathbf{x}_n | \mathbf{Z}^{(n)}, \mathbf{Y}, \mathbf{R}) P(\mathbf{Z}^{(n)} | \mathbf{Z}^{(n-1)}) P(\mathbf{Z}^{(n-1)}, \mathbf{Y}, \mathbf{R} | \mathbf{X}^{(n-1)}). \quad \text{(F1)}$$

Intuitively, we have decomposed the posterior into three terms (described from left to right): the likelihood of $\mathbf{x}_n$ given the proposed feature representation (which is given by Equation 3), the probability of the proposed feature ownership matrix given the feature ownership matrix we had after $n - 1$ objects (which is simply the restaurant sampling scheme for generating the Indian buffet process), and the posterior distribution of the feature representation after observing $n - 1$ objects. The last term is the posterior distribution over feature representations, but for $n - 1$ objects instead of n objects, and thus we can sequentially update our posterior distribution given the next object using the posterior distribution given one fewer object.

Of course we cannot sum over all possible previous feature ownership matrices $\mathbf{Z}^{(n-1)}$ (as it is infinite for even one object). Instead, we approximate the distribution by storing P particles, which are samples of the feature representation for the currently observed objects. We first initialize the particles in the same manner as the Gibbs sampler is initialized. Then we infer the feature representation after observing n objects by updating each particle (the inferred feature ownership matrix after $n - 1$ objects) using the restaurant sampling scheme, sampling new $\mathbf{Y}$ and $\mathbf{R}$ using Gibbs sampling with the updated $\mathbf{Z}^{(n)}$ and weighting the particles by how well they reconstruct the most recently observed object $\mathbf{x}_n$ using the updated features.

For the simulations reported in the article, $P = 100$ and the parameters α, ε, λ, and θ were initialized to 2, 0.001, 0.999, and 0.4 for both the Gibbs sampler (run for 1,000 iterations) and particle filter, and the model was given five examples of each type of object. Each image was 30×30 pixels. Simulated annealing ($T = 25$) was applied for the Gibbs sampler, which converged quickly to the two feature solution in both object orders. After observing all 15 objects, all particles contained the same feature representation.

Received November 9, 2012
Revision received July 17, 2013
Accepted July 17, 2013 ■