Cognitive Psychology 63 (2011) 173–209

Contents lists available at SciVerse ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

A rational model of the effects of distributional information on feature learning

Joseph L. Austerweil *, Thomas L. Griffiths

Department of Psychology, University of California, Berkeley, United States

ARTICLE INFO

Article history:
Accepted 22 August 2011
Available online 20 September 2011

Keywords:
Representational change
Features
Rational analysis
Bayesian modeling
Nonparametric Bayesian statistics

ABSTRACT

Most psychological theories treat the features of objects as being fixed and immediately available to observers. However, novel objects have an infinite array of properties that could potentially be encoded as features, raising the question of how people learn which features to use in representing those objects. We focus on the effects of distributional information on feature learning, considering how a rational agent should use statistical information about the properties of objects in identifying features. Inspired by previous behavioral results on human feature learning, we present an ideal observer model based on nonparametric Bayesian statistics. This model balances the idea that objects have potentially infinitely many features with the goal of using a relatively small number of features to represent any finite set of objects. We then explore the predictions of this ideal observer model. In particular, we investigate whether people are sensitive to how parts co-vary over objects they observe. In a series of four behavioral experiments (three using visual stimuli, one using conceptual stimuli), we demonstrate that people infer different features to represent the same four objects depending on the distribution of parts over the objects they observe. Additionally in all four experiments, the features people infer have consequences for how they generalize properties to novel objects. We also show that simple models that use the raw sensory data as inputs and standard dimensionality reduction techniques (principal component analysis and independent component analysis) are insufficient to explain our results.

* Corresponding author. Address: Department of Psychology, University of California, Berkeley, 3210 Tolman Hall # 1650, Berkeley, CA 94720-1650, United States. Fax: +1 510 642 5293.
E-mail address: Joseph.Austerweil@gmail.com (J.L. Austerweil).

0010-0285/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogpsych.2011.08.002

1. Introduction

> Does the stimulus have no intrinsic structure of its own, or does it simply exist to provide structure for the sets in which it exists? The answer to this question is that the properties of the single stimulus cannot be specified except in relation to the properties of the set within which it exists. (Garner, 1974, p. 9).

A fundamental problem faced by any learner is the formation of the basic units used to represent observed stimuli and support generalizations. Although psychologists typically choose a single representation for each observed stimulus, Garner (1974) identified the basic issue with this approach: The representation of a stimulus is not fixed or unchanging regardless of context, but rather, it is defined with respect to the context in which it appears. This idea coincides with the classic Gestalt analysis of visual perception, which argued through introspection that the interpretation of a whole set of stimuli is different than the sum of its parts (Wertheimer, 1938). Although the investigation of Gestaltist principles has led to a fruitful body of research on how context affects the perception of single objects (e.g., Palmer, 2003), there currently does not exist a computational account of how the set of objects people observe influences how they represent those objects without assuming the number of features is known ahead of time. In this paper, we present a computational model of how feature representations should be inferred from a set of observed objects without assuming the number of features is known ahead of time, and compare the predictions of this model against human behavior.

In discussing feature learning, it is valuable to distinguish between two ways in which an object can be described: as an observable stimulus, and as an internal representation. We will try to maintain this distinction by talking about the "properties" or "parts" of a stimulus, reserving "features" for the components of its internal representation. The key problem of feature learning is thus explaining how people come to establish a relationship between observable properties and internally represented features, deciding which combinations of properties constitute features. There are many factors that influence the features people infer to represent objects, such as the changes in concavity of its contour (Hoffman & Richards, 1985), the usefulness of the potential feature for explaining categorization of objects (Pevtzow & Goldstone, 1994; Schyns & Murphy, 1994), background knowledge of the function of objects (Lin & Murphy, 1997), and prior knowledge of what types of features have been useful in the past (e.g., Gestalt principles; Palmer, 1977). However, we will focus on one particular factor: the distribution of properties over a set of objects.

Distributional information is one example of how context can affect the features formed to represent a set of objects. The same objects in two different contexts (presented with two different sets of other objects) can have different feature representations depending on how properties are distributed across the other objects. For example, imagine a set of objects that are each composed of several parts. If the parts co-vary perfectly in each of the objects, then there is no distributional information to suggest that these parts are separable, and they should be combined into a single feature. Conversely, if the parts occur independently over all of the observed objects, then each of the parts can be identified based on this statistical information, and those parts could be used as features for representing the objects. This captures the intuition that a feature representation is useful if knowing an unknown object has a feature gives you information as to which object it is. In the first case (co-varying parts), individuating each part does not provide useful information as to object membership; however, in the second case, individuating the parts helps differentiate the objects.

A large body of previous research has demonstrated that people are sensitive to distributional information. Expectations about the distributions of different quantities in our environment influence how people perceive the world (Ernst & Banks, 2002; Koerding & Wolpert, 2004; Weiss, Simoncelli, & Adelson, 2002). Studies of human language acquisition show that infants can exploit statistical patterns in sequences of speech sounds to segment a continuous speech stream into words (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996). A similar capacity seems to exist in the visual system, with people being able to pick out visual stimuli that often appear in sequence (Kirkham, Slemmer, & Johnson, 2002), or appear together in a two-dimensional array (Fiser & Aslin, 2001). We complement these previous results by performing a rational analysis of the role of distributional information in feature learning. In the spirit of Anderson (1990) and Marr (1982), this analysis considers

the nature of the underlying computational problem and how that problem might be solved by a rational agent. In particular, we analyze the effect that distributional information should have on the representations people form, and demonstrate that people use these statistical cues when they infer feature representations for novel objects.

Our rational model uses an approach from nonparametric Bayesian statistics, which allows objects to be represented using potentially infinitely many features. This infinite capacity is consistent with the idea that potentially any combination of properties might be considered a feature. The model creates features to reproduce the objects it observes, but is penalized for each feature it produces. This results in a "simplicity bias," similar to that seen in other domains (Chater, 1999; Chater & Vitanyi, 2003; Lombrozo, 2007). Thus, the model seeks to infer the number of features necessary to represent the objects it observes. To the best of our knowledge, it is the only model of feature inference that infers the number of features and their identity directly from the observable properties of objects.

Our rational analysis predicts that distributional information should affect the features people learn, which we show is consistent with previous behavioral results, and go on to confirm in a series of new experiments. We focus on how distributional information is used in either uniting a set of parts into a single feature, or differentiating those parts into different features. If the parts that compose objects co-vary over objects, an observer should infer the objects themselves as features. On the other hand, if the parts that compose objects vary independently over objects, then an observer should infer the parts as features. We demonstrate that this effect appears with more than one type of visual objects and potentially extends to conceptual domains as well.

The plan of the paper is as follows. First, we present a computational level framework for inferring feature representations from the raw sensory impressions of observed objects. Next, we explore some properties of the framework and relate them to previous behavioral results on human feature learning. Our computational framework predicts that how parts co-vary over objects should affect how an observer infers features to represent the objects. In a series of experiments, we confirm this prediction and demonstrate that the pattern of results cannot be explained by classic exemplar and prototype models or more traditional dimensionality reduction techniques in machine learning (principal component analysis and independent component analysis). Finally, we conclude with a discussion of the generality of the framework, the limitations of our account, and future directions.

2. A rational analysis of feature learning

Rational analysis is a technique for understanding a cognitive process by comparing it to the optimal solution to an underlying computational problem, with the goal of understanding how this solution relates to human behavior (Anderson, 1990; Oaksford & Chater, 1998). Using the distinction introduced by Marr (1982), this analysis focuses on the abstract "computational" level of formulating the problem and its ideal solution, rather than the "algorithmic" level of the cognitive processes that might execute this solution. By formally analyzing the problem of inferring feature representations from the properties of a set of objects, we can determine how distributional information should influence the features used to represent those objects, giving us a set of predictions against which we can compare human judgments.

2.1. Identifying the computational problem

Part of what makes determining the features of a novel object challenging is that these features are not intrinsic to the object, but depend on context. One of the major criticisms of accounts of similarity that are based on features (such as Tversky, 1977) is that it is possible to identify infinitely many possible features for any particular object (Goldmeier, 1936/1972; Goodman, 1972; Medin, Goldstone, & Gentner, 1993; Murphy & Medin, 1985). For example, a mug has a handle and is roughly cylindrical, but it is also heavier than a piece of paper, larger than an ant, smaller than a breadbox, and so forth. Which of these features are relevant will depend on the context. This flexibility led Murphy and Medin (1985) to conclude that feature-based accounts of similarity provide a poor explanation of many

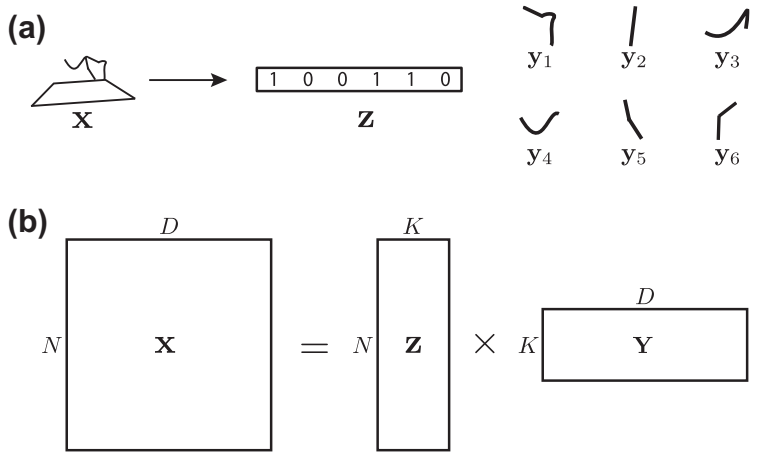

Fig. 1. Feature representation formulation. (a) Each object **x** is represented by a binary feature ownership vector **z** encoding which features **x** contains and the feature images $\mathbf{y}_1, \ldots, \mathbf{y}_6$. The object is recreated by superimposing the feature images that it contains (given by **z**). (b) The problem of learning the features of a set of objects reduces to one of factorizing the matrix **X**, whose rows are the objects **x**, into the product of a matrix **Z** whose rows are the corresponding feature ownership vectors, and a matrix **Y** containing the feature images.

aspects of cognition, since "the explanatory work is on the level of determining which attributes will be selected" (p. 296).

The computational problem that we want to solve is thus not one of determining the intrinsic features of an object, but picking out which of the infinite array of features of that object should be used to represent it in a given context. We will focus on a special case of this problem, where the context is provided by the other objects that appear in a set, and where the features are binary (ie. each object either possesses or does not possess that feature). These features need to be inferred from the observable properties of the set of objects. Features will be context-dependent to the extent that different sets of objects result in different representations.

Formally, we describe the observable properties of the set of objects with a matrix **X**, where the rows of the matrix correspond to objects and the columns to properties.[1] Each object **x** is a row of the matrix **X**. In this article, the observable properties of an object **x** are an array of pixels representing the intensity of light over the retina. The two-dimensional array is converted into a one-dimensional vector. Thus the matrix **X** has N rows, one for each object, and D columns, one for each pixel in the image of an object. The properties of objects can be expressed either as binary or continuous values describing how the objects vary along a specific observable dimension.

The computational problem is to find a feature representation for **X** of size K. Each row **x** will correspond to a vector **z** of size K, whose elements indicate which features are possessed by that object (with 1 indicating the object has the feature, and 0 that it does not). We also need to identify how these features are expressed in the original D dimensions used to encode **x**, so each of the K features will be associated with an "image" indicating which properties correspond to that feature. Fig. 1a illustrates how a stimulus **x** might be encoded in terms of a feature ownership vector **z** and a set of images associated with those features. **x** is recreated using **z** and $\mathbf{y}_1, \ldots, \mathbf{y}_6$ by superimposing all the images of the features it has (given by **z**) on top of one another.[2]

[1] We will use a convention where matrices are represented by uppercase boldface letters, such as **X**, vectors are lowercase boldface, such as **x**, and scalars are lowercase italic, such as x. Constants are uppercase italic, such as K, and parameters are Greek letters such as α. $P(\cdot)$ is used for probability mass functions (ie. distributions on discrete quantities, summing to 1) and $p(\cdot)$ is used for probability density functions (ie. distributions on continuous quantities, integrating to 1).

[2] In this example, the observable properties of **x** are binary pixels. However, our method is general to other observable properties. For example, in Experiment 3, we represent the observable properties of **x** are continuous-valued pixels, and then the feature images will also be continuous.

This problem of finding the features for a set of objects can be cast mathematically as a problem of matrix factorization. Intuitively, our goal is to be able to reconstruct $\mathbf{X}$ as the product of two matrices: (1) a $N \times K$ binary matrix feature ownership matrix $\mathbf{Z}$ (whose rows are the feature ownership vectors $\mathbf{z}$) and (2) a $K \times D$ feature "image" matrix $\mathbf{Y}$ that encodes the consequence of an object having each feature. This is illustrated in Fig. 1b. However, to complete this formulation of the problem we need to specify the number of features, K. As discussed above, this is a key part of the problem, as objects can have an arbitrary number of features, with different features becoming salient in different contexts. Thus, we want to allow K to be arbitrarily large, but expect that for a set of N objects we will only use a finite number of features. In the rest of the section, we discuss how this problem can be solved using tools from nonparametric Bayesian statistics, introduce the particular probability model used, and then show how this approach can be used to find the features that represent a set of objects.

2.2. Feature learning as nonparametric Bayesian inference

Setting aside for the moment the problem of determining the number of features, the computational problem presented in the previous section reduces to identifying the most probable feature representation for a set of objects given the observable properties of those objects. That is, we want to infer the most probable feature ownership matrix $\mathbf{Z}$ and feature image matrix $\mathbf{Y}$ given the observed properties $\mathbf{X}$. However, the values of $\mathbf{Z}$ and $\mathbf{Y}$ are underdetermined by $\mathbf{X}$, with many different factorizations being possible. The matrix factorization equation $\mathbf{X} = \mathbf{ZY}$ is underdetermined in exactly the same way as $2 = ab$, where any pair of values for a and b that multiple together to yield 2 is possible. To solve a problem of this kind, we need to combine the information provided by $\mathbf{X}$ with some expectations about the values of $\mathbf{Z}$ and $\mathbf{Y}$.

Viewing the problem in these terms makes it clear that it is an inductive problem, and the rational solution to problems of this kind is provided by Bayesian inference (Griffiths, Kemp, & Tenenbaum, 2008). Here, the properties of the objects $\mathbf{X}$ constitute some observed data, and the values of $\mathbf{Z}$ and $\mathbf{Y}$ are hypotheses that might explain those data. Finding the most likely values of $\mathbf{Z}$ and $\mathbf{Y}$ requires calculating $P(\mathbf{Z},\mathbf{Y}|\mathbf{X})$, the posterior probability of $\mathbf{Z}$ and $\mathbf{Y}$ given that we have observed $\mathbf{X}$. This can be done by applying Bayes' rule, with

$$P(\mathbf{Z},\mathbf{Y}|\mathbf{X}) = \frac{P(\mathbf{X}|\mathbf{Y},\mathbf{Z})P(\mathbf{Z})P(\mathbf{Y})}{\sum_{\mathbf{Z}',\mathbf{Y}'} P(\mathbf{X}|\mathbf{Y}',\mathbf{Z}')P(\mathbf{Z}')P(\mathbf{Y}')} \tag{1}$$

$$\propto P(\mathbf{X}|\mathbf{Z},\mathbf{Y})P(\mathbf{Z})P(\mathbf{Y}) \tag{2}$$

where $P(\mathbf{X}|\mathbf{Z},\mathbf{Y})$, known as the "likelihood", expresses the probability of observing $\mathbf{X}$ if a particular value of $\mathbf{Z}$ and $\mathbf{Y}$ generated it, and $P(\mathbf{Z})$ and $P(\mathbf{Y})$ indicate the "prior" probability of $\mathbf{Z}$ and $\mathbf{Y}$ respectively, expressing the expectations of the learner about the values of these matrices.

Formulating the problem of identifying a set of features as one of Bayesian inference breaks this problem into two subproblems: finding a representation that generates the observed properties of the objects with high probability, as captured by the likelihood $P(\mathbf{X}|\mathbf{Z},\mathbf{Y})$, and finding a representation that is in general probable, as expressed by the priors $P(\mathbf{Z})$ and $P(\mathbf{Y})$. The representation that is formed should be a compromise between these two factors, providing a good match to the objects but also being consistent with the expectations of the learner. This decomposition also provides insight into the problem of determining the number of features that a set of objects possess.

The likelihood $P(\mathbf{X}|\mathbf{Z},\mathbf{Y})$ should measure how well $\mathbf{X}$ is approximated by the product of $\mathbf{Z}$ and $\mathbf{Y}$. Informally, we might imagine comparing each object $\mathbf{x}$ with the superposition of the images in $\mathbf{Y}$ corresponding to the features possessed by that object, as indicated by its feature ownership vector $\mathbf{z}$. Thus, the only features that have a consequence are those which are used by at least one object, or in other words those columns of $\mathbf{Z}$ that have at least one element that is one (we will refer to these as "non-zero" columns). If we allow $\mathbf{Z}$ to have an infinite number of columns, but allow only a finite number of non-zero columns, then we can learn feature representations for objects with an arbitrary number of features. For convenience, we can order the columns such that the non-zero columns appear first in the matrix, being followed by an arbitrary number of zeros. The non-zero columns thus determine the effective dimensionality of $\mathbf{Z}$.

This analysis locates the solution to the problem of having infinitely many possible features of which only a finite subset are used to represent any set of objects in the prior $P(\mathbf{Z})$. We need to define a prior distribution that generates matrices with an infinite number of columns, of which a finite number are non-zero. By applying Bayes' rule, we can then infer the matrix that best accounts for the structure of the observed data. In the remainder of this section we introduce such a prior distribution and then consider how it can be combined with different likelihood functions to infer the features of sets of objects with different types of observable properties.

2.2.1. A prior on feature ownership matrices

The problem of defining distributions on matrices with infinitely many columns has been explored in nonparametric Bayesian statistics, a branch of Bayesian statistics that focuses on models that have potentially infinite complexity. By defining a prior that allows infinite complexity but penalizes more complex models, the nonparametric Bayesian approach trades off keeping the amount of structure in the model as small as possible with the model's ability to explain the data. Thus, nonparametric Bayesian models find the simplest representation for a set of objects without limiting the space of possible representations.

We will use a particular distribution on binary matrices, which has been used as a prior in a number of nonparametric Bayesian models. This distribution, known as the Indian Buffet Process (IBP; Griffiths & Ghahramani, 2006), has several nice properties: it allows for multiple features per object, the probabilities of possessing different features are independent of each other,[3] and it generates binary matrices of arbitrarily large dimensionality. These properties make it a good starting point for an investigation of people's expectations about the features of objects, although we imagine that future research might explore priors that make different assumptions (see the General Discussion for further details).

The IBP defines a distribution over binary matrices with a fixed number of rows and an infinite number of columns, of which only a finite number are expected to have non-zero elements. However, an intuitive understanding for this distribution can be obtained by starting with the problem of generating a binary matrix with K columns. The simplest way we might imagine doing this is by flipping a coin for each entry in the matrix, putting a 1 in that entry for "heads" and 0 for "tails." A slightly more complex model might use a different coin for each column, with a parameter π_k indicating the probability of heads for the coin used for column k. Now, consider what happens as K, the number of columns, becomes larger. Provided π_k becomes smaller at a corresponding rate, the total number of ones in the matrix (and the number of non-zero columns) will remain finite. This is exactly how the IBP was constructed, by assuming each π_k is drawn from a distribution that favors values closer to 0 as $K \to \infty$ (for details, see Griffiths & Ghahramani, 2006).

The probability distribution over binary matrices that results from this limiting construction is equivalent to that provided by a simple stochastic process, which is typically described via a metaphor in which objects are customers and features are dishes in an Indian buffet. If we imagine a buffet with infinitely many dishes that is visited by a fixed number of customers, we can generate a binary matrix with a potentially infinite number of columns but a fixed number of rows by recording the dishes that are tasted by each customer. The IBP is equivalent to the distribution induced by the following process (see Griffiths & Ghahramani, 2006 for details). The first customer samples a number of dishes that is drawn from a Poisson(α) distribution. The next customer tastes each of these dishes with probability 1/2 (as each of these have been sampled once and this is the second customer), and samples Poisson($\alpha/2$) new dishes. This process continues, with the ith customer tasting the kth dish with probability m_k/i, where m_k is the number of people who previously sampled the dish, and sampling Poisson(α/i) new dishes.

If we define a binary matrix $\mathbf{Z}$ such that $z_{ik} = 1$ if the ith customer samples the kth dish and is 0 otherwise, the IBP specifies a distribution on binary matrices that can be used as a prior on feature ownership matrices. If we continue the culinary metaphor for N customers, imagining that the binary

[3] Under the IBP prior, the probabilities of possessing different features are independent. However, they are dependent in the posterior because if an object has one feature that explains some of its observable properties, it is less likely to take another feature that explains the same observable properties.

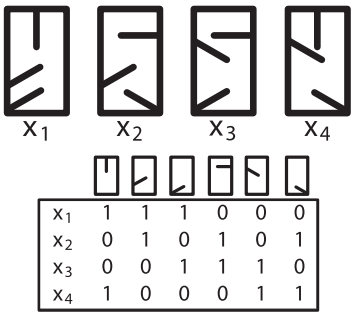

Fig. 2. Inferring representations for objects. Stimuli and feature ownership matrix from Shiffrin and Lightfoot (1997).

matrix represents how each customer (object) came into the restaurant and choose dishes (features), the probability of getting a matrix with a particular set of columns is

$$P(\mathbf{Z}) = \frac{\alpha^{K_+}}{\prod_{h=1}^{2^N-1} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!} \tag{3}$$

where N is the number of objects, K_h is the number of features with history h (the history is the column of the feature interpreted as a binary number), K_+ is the number of columns with non-zero entries, H_N is the Nth harmonic number $\left(H_i = \sum_{j=1}^{i} j^{-1}\right)$, and m_k is the number of objects that have feature k.[4] For example, the history for the first feature in Fig. 2 is 9 ((1,0,0,1) is the binary expansion of $9 = 2^0 + 2^3$), the history of the third feature is 5 ((1,0,1,0) is the binary expansion of $5 = 2^0 + 2^2$), and K_h for all of the non-zero features in the matrix of Fig. 2 is 1. Note that Eq. (3) does not depend on the ordering of the rows or columns of $\mathbf{Z}$; meaning that under the IBP, the order that objects enter the restaurant does not matter. This means that our model is insensitive to the order in which the objects are presented, a point that we return to in the General Discussion.

The distribution defined by the IBP has the properties that we want in a prior on feature ownership matrices. First, it allows for potentially infinitely many features. The matrix $\mathbf{Z}$ can be thought of as having infinitely many columns, although most of those columns contain zeros. As N grows to infinity, the number of non-zero columns K also grows to infinity, corresponding to the idea that we discover more features for objects as more objects are observed. Second, this distribution can favor feature representations with a fewer number of features. By choosing α such that $\frac{\alpha}{N} < 1$, the IBP implements this bias because the $\left(\frac{\alpha}{N}\right)^K$ term decreases when the number of features in the representation, K, grows. Using this distribution as a prior, we can infer the number of features required to represent a set of objects, combining a bias towards a simpler representation with the information provided by the observed properties of the objects.

2.2.2. Two likelihood functions

To define the likelihood, $P(\mathbf{X}|\mathbf{Y},\mathbf{Z})$, we assume N objects with D observed properties (e.g., pixels in an image) are grouped in a matrix $\mathbf{X}$ ($\mathbf{X} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T]$, where $\mathbf{x}_i \in \mathbb{R}^D$ or $\mathbf{x}_i \in \{0,1\}^D$). The feature ownership matrix $\mathbf{Z}$ marks the commonalities and contrasts between these objects, and the likelihood $P(\mathbf{X}|\mathbf{Y},\mathbf{Z})$ expresses how these relationships influence their observed properties. The likelihood links the feature ownership matrix to the observable properties. We do this by defining a hidden $K \times D$ feature image matrix $\mathbf{Y}$ that encodes the consequence of having each feature. The appropriate likelihood depends on the form of the observed properties in $\mathbf{X}$. When the observed properties are continuous, we use the linear-Gaussian likelihood (Griffiths & Ghahramani, 2006), and when they are binary, we use the noisy-OR likelihood (Wood, Griffiths, & Ghahramani, 2006).

[4] Technically, this probability is obtained by summing over all matrices that have the same set of columns, regardless of the order of those columns.

The linear-Gaussian likelihood assumes that $\mathbf{x}_i$ is drawn from a Gaussian distribution with mean $\mathbf{z}_i\mathbf{Y}$ and covariance matrix $\mathbf{\Sigma}_X = \sigma_X^2\mathbf{I}$, where $\mathbf{z}_i$ is the binary vector defining the features of object $\mathbf{x}_i$ and $\mathbf{Y}$ is a matrix of the weights of each element of D properties for each feature k, having continuous values. This gives the likelihood function

$$p(\mathbf{X}|\mathbf{Z},\mathbf{Y},\sigma_X) = \frac{1}{(2\pi\sigma_X^2)^{ND/2}}\exp\left\{-\frac{1}{2\sigma_X^2}\mathrm{tr}((\mathbf{X}-\mathbf{ZY})^T(\mathbf{X}-\mathbf{ZY}))\right\}, \tag{4}$$

where $\mathrm{tr}(\cdot)$ is the trace function, which returns the sum of the elements on the main diagonal of a matrix. This likelihood is the rational choice when the values in $\mathbf{X}$ are continuous and the metric of success is the sum of the squared errors (SSE) between $\mathbf{ZY}$ (the objects reconstructed using the inferred feature representation) and $\mathbf{X}$. The trace term in Eq. (4) is simply the SSE, meaning that values of $\mathbf{Z}$ and $\mathbf{Y}$ that result in a larger SSE will have a lower likelihood. In addition, we assume that each element of $\mathbf{Y}$ is generated from a Gaussian distribution with mean 0 and variance σ_Y^2. The standard deviations σ_X and σ_Y are parameters of the model, and determine how the model trades off errors in reconstructing $\mathbf{X}$ and large values in $\mathbf{Y}$.

For the noisy-OR model (Wood et al., 2006), the raw visual data are reduced to binary pixel values. This model assumes that the observable properties $\mathbf{X}$ are generated from a noisy-OR distribution, where $\mathbf{Z}$ defines the features that each object has and $\mathbf{Y}$ defines which properties are associated with each feature. The noisy-OR distribution is used in research on human causal learning, where it seems to capture the assumptions that people make about how multiple causes combine to influence an effect (Cheng, 1997; Griffiths & Tenenbaum, 2005). In this case, each feature is a potential cause of a particular pixel turning on. The total number of potential causes for object i having pixel d turned on is given by the inner product of the feature vector for that object with the vector of $\mathbf{Y}$ indicating which features are associated with that pixel, $\mathbf{z}_i\mathbf{y}_d$. The likelihood function results from assuming each $x_{i,d}$ is sampled independently, with the probability that it takes the value 1 given by

$$P(x_{i,d} = 1|\mathbf{Z},\mathbf{Y},\lambda,\epsilon) = 1-(1-\lambda)^{\mathbf{z}_i\mathbf{y}_d}(1-\epsilon) \tag{5}$$

where ϵ and λ are parameters representing the probability a property is present without a cause and the probability a feature causes an object to possess a property respectively. One interpretation of Eq. (5) is to assume that each pixel is off *a priori*, but each feature the object has with that pixel on turns it on with probability λ or that the pixel turns on its own with probability ϵ. In addition, $\mathbf{Y}$ is assumed to have a Bernoulli prior with parameter ϕ representing the probability that an entry of $\mathbf{Y}$ is one, with $P(\mathbf{Y}) = \prod_{k,d}\phi^{y_{k,d}}(1-\phi)^{1-y_{k,d}}$. The values of ϵ, λ, and ϕ determine how the model trades off errors in reconstructing $\mathbf{X}$ with the sparsity of feature images in $\mathbf{Y}$.

The assumptions made about the prior distribution on $\mathbf{Y}$ in both of these variants of the model ignore spatial factors. That is, the pixels that are used by each feature are assumed to be independent. This simplifying assumption is made because our primary interest is how people form basic units regardless of domain (as Experiment 4 uses conceptual objects), but it is not necessary. In the General Discussion we explore a prior that imposes a proximity constraint, in which neighboring pixels are constrained to have similar values.

3. Modeling the formation of features through unitization

Having discussed theoretical issues of feature learning and defined a model for inferring features, we now show how one basic phenomenon of feature learning can be understood from the perspective of this rational model. One line of investigation of human feature learning concerns the phenomena of unitization and differentiation. *Unitization* occurs when two or more features that were previously perceived as distinct features merge into one feature. In a visual search experiment by Shiffrin and Lightfoot (1997), after learning that the parts generating the observed objects co-vary in particular ways, participants represented each object as its own feature instead of as three separate features. In contrast, *differentiation* is when a fused feature splits into new features. For example, color novices cannot distinguish between a color's saturation and brightness; however, people can be trained to make these distinctions (Goldstone, 1994).

Unitization and differentiation are typically discussed as the result of perceptual learning (Goldstone, 1998). Instead of interpreting all cases of unitization and differentiation as the result of perceptual learning, we explore how unitization and differentiation occur in our rational model depending on how parts co-vary over the set of observed objects. As the rational model does not contain any perceptual constraints, this can be thought of as a domain general approach that connects perceptual unitization and chunking (Hall, 1991). Although general conditions for when differentiation or unitization occur have been outlined, there is no formal account for why and when these processes take place. Here, we will focus on the phenomenon of unitization, showing how this falls out of our rational model.

In Experiment 1 of Shiffrin and Lightfoot (1997), participants were trained to find one of the objects shown in Fig. 2 in a scene where the other three objects were present as distractors. Each object is composed of three parts (single line segments) inside a rectangle. The objects can thus be represented by the feature ownership matrix shown in Fig. 2, with $z_{ik} = 1$ if object i has feature k. After prolonged practice, human performance drastically improved, and this advantage did not transfer to other unseen objects created from the same feature set. Shiffrin and Lightfoot concluded that the human perceptual system had come to represent each object holistically, rather than as being composed of its more primitive features. In this case, the fact that the features tended to co-occur only in the configurations corresponding to the four objects provides a strong cue that they may not be the best way to represent these stimuli.

When should whole objects or line segments be inferred as features? It is clear which features should be inferred when all of the line segments occur independently and when the line segments in each object always occur together (the line segments and the objects respectively). In fact, Medin, Altom, Edelson, and Freko (1982) have shown that for conceptual stimuli people act in accordance with this principle. However, in the intermediate cases of non-perfect co-occurrence (as with these stimuli), what should be inferred? Without a formal account of feature learning, there is no basis for determining when object "wholes" or "parts" should be inferred as features. Our rational model provides an answer – when there is enough statistical evidence for the individual line segments to be features, then each line segment should be differentiated into features. Otherwise, the collection of line segments should be inferred as one unitized feature.

The stimuli constructed by Shiffrin and Lightfoot (1997) constitute one of the intermediate cases between the extremes of total independence and perfect correlation, and are thus a case in which formal modeling can be informative. Fig. 3 presents the features learned by applying the model with a noisy-OR likelihood to this object set. Although there is imperfect co-occurrence between the features in each object, there is not enough statistical evidence to warrant representing the object as a combination of features. The rational model thus favors representing each object as a single unit. These results were obtained with an object set consisting of five copies of each of the four objects with added noise that flips a pixel's value with probability $\frac{1}{75}$ (see Appendix for simulation details).

Though the learned features match the representation formed by people in this experiment, their psychological plausibility is weakened by the "speckled holes" in the features. In addition to domain

Fig. 3. Inferring feature representations using distributional information from Shiffrin and Lightfoot (1997). The four objects are learned as features; however, the features inferred by the IBP with the independent feature image prior contain discontinuities where the objects overlap (see Section 9 for how to include perceptual constraints into the model, which allow the model to infer more psychologically valid features). The rational model justifies the human perceptual system's unitization of the objects as features

general statistical cues, people use perceptual constraints to infer features, such as proximity information (Goldstone, 2000), which are not included in our model. In the General Discussion, we show how perceptual constraints can be incorporated in our model using a feature image prior with a proximity bias and how this model infers more psychologically plausible features.

4. Wholes and parts: inferring features using distributional information

Our rational model predicts unitization will occur in Shiffrin and Lightfoot's (1997) first experiment because of the statistical structure of the stimuli: The co-occurrence of parts creates a pattern of correlation that is best explained by postulating the objects themselves as features, resulting in a holistic representation of the objects. To demonstrate that distributional information drives this prediction, we need to show that we obtain the opposite result – differentiation of features – when the stimuli have a different statistical structure.

We conducted a simulation to demonstrate that the statistical structure of the input affects the parts of objects identified as features. Fig. 4a shows the base (on left) and the set of six parts used in the simulations. Fig. 4b is an artificially generated set of observed objects for which there is not enough statistical evidence to warrant differentiation. The feature membership matrix corresponding to this set of observed objects is the same as that used in Experiment 1 of Shiffrin and Lightfoot (1997), except that there are four copies of each object. We refer to this as the *correlated* set. The sixteen images in the set were made by adding noise to images of the four objects, and then copying the resulting images identically four times. The noise was implemented by flipping each pixel with probability $\frac{1}{75}$. The four copies thus had identical noise, although similar results are obtained with independent noise.

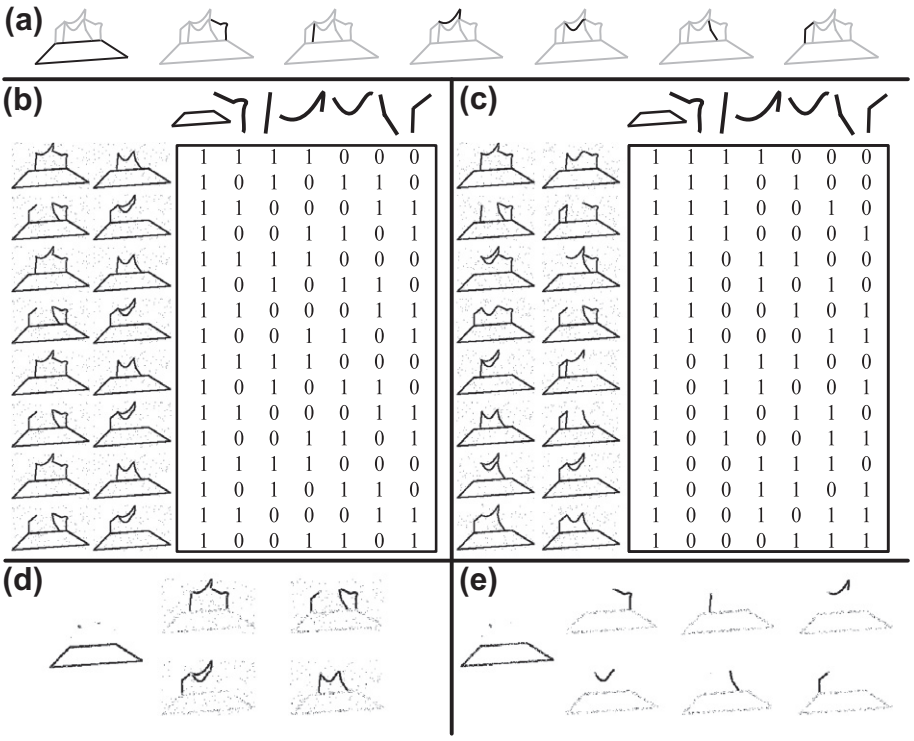

Fig. 4. Inferring different feature representations depending on the distributional information. (a) The base (on left) and the six features used to generate both object sets. (b) and (c) The feature membership matrices and objects for the (b) *correlated* and (c) *independent* sets respectively. (d) and (e) The feature representations inferred by model for the (d) *correlated* and (e) *independent* sets respectively.

Fig. 4c is an artificially generated object set in which the observed objects should be differentiated. Here, the parts used to generate the objects occur (nearly) independently of each other. The underlying feature membership matrix used to generate the observed objects in this set is the four objects from the *correlated* set and twelve of the other possible objects. We refer to this as the *independent* set.[5] Each image in the set is given independent noise, which was implemented by flipping each pixel with probability $\frac{1}{75}$. This leaves four remaining objects, which form the *unseen* set (in total there are $\binom{6}{3} = 20$ objects constructed by combinations of three parts out of a set of six).

We applied the rational model to the *correlated* and *independent* object sets, using a noisy-OR likelihood (see Appendix). Fig. 4d and e show the results. When the parts strongly co-occur (the *correlated* set), the model forms a representation which consists simply of the objects themselves. When the underlying parts occur strongly independently of one another (the *independent* set), the model uses the parts as features. The pixelation of the inferred features could be removed by averaging over multiple runs of the model.

Importantly, the two different feature representations make different predictions on the *unseen* objects. When the whole objects are the features, the *unseen* objects are unexpected because they cannot be represented using the objects as features and thus the model should differentiate between them and the objects it observed. When the parts are features, the *unseen* objects are expected because they can be represented using the parts as features and thus, in this case, the model should not differentiate between them and the objects it observed. These simulations demonstrate that even when the same underlying parts create two object sets and the same four objects are in both sets, different representations should be inferred depending on the distributional information contained in the context of the other objects presented with the four original objects. This suggests that distributional information can be a powerful driving force behind unitization and differentiation. Furthermore, the different representation imply behavioral consequences as different objects should be expected and generalized to depending on which representation is used. In the remainder of the article, we examine whether this prediction holds for human feature learning.

5. Experiment 1: Feature learning from binary images

The simulations in the previous section lead to the question of whether people use distributional information to infer the features of objects, as our rational model predicts. In Experiment 1, we test whether the way that parts are distributed over objects (*correlated* or *independent*) affects how people form generalizations. Based on our model, the prediction is that people who see correlations will form features consisting of the whole objects, and thus generalize less to unseen objects made of the same parts. However, people who have the parts as features should not differentiate between the objects they observed and the unseen objects.

The stimuli were the same as those used in the simulations in the previous section, with people being trained on the *correlated* or *independent* sets, and then being tested with a set of stimuli that includes the *unseen* objects. The training and test sets were carefully constructed to ensure that: (1) the variance at each pixel was equal for all training sets, as was the number of times each part appeared, and (2) the average similarity (in terms of pixel overlap) between any training set and any test set was equal.

5.1. Methods

5.1.1. Participants

A total of 28 undergraduates from University of California, Berkeley, participated in the experiment in exchange for course credit. There were 14 participants in each of the *correlated* and *independent* conditions, with *test order* counterbalanced.

[5] The amount of independence between parts increases as novel objects are added to the set. The correlated set has four of the possible objects and the independent set has 16 of the possible objects. Neither set is perfectly correlated or independent, so a computational model is needed to decide which features should be used.

5.1.2. Stimuli

Fig. 4a shows the images of the parts (combinations of lines and curves) and base (a trapezoid shared by all objects) that combine to form the objects shown to participants. Each object was formed by the union of three parts and the base. Thus, there were twenty possible objects, corresponding to all possible ways of choosing three parts from a set of six. The parts and base were designed such that any combination of three parts with the base formed a connected object with roughly equal *a priori* "goodness." For simplicity, the images were black and white (binary).

The main manipulation of this experiment, *distribution type* had two levels: *correlated* and *independent*. The parts of the *correlated* sets strongly co-varied over the objects, but did not so perfectly. Thus, finding out that a particular part was in an object in a *correlated* set provided information about which other parts were in the object, but not perfect certainty. The parts of the objects in the *correlated* sets had the same amount of correlation as those in Shiffrin and Lightfoot (1997). The correlated stimuli are shown in Fig. 5a. Each set consisted of four identical copies of four objects that were perturbed by random noise. The *independent* set consisted of 16 of the 20 possible objects and is shown in Fig. 5b. The set of four objects missing from the *independent* set were the same as the four objects of the *correlated* set. This method of generating stimuli guaranteed that each part in the *correlated* set and the *independent* set appeared with the same frequency, allowing us to control for familiarity

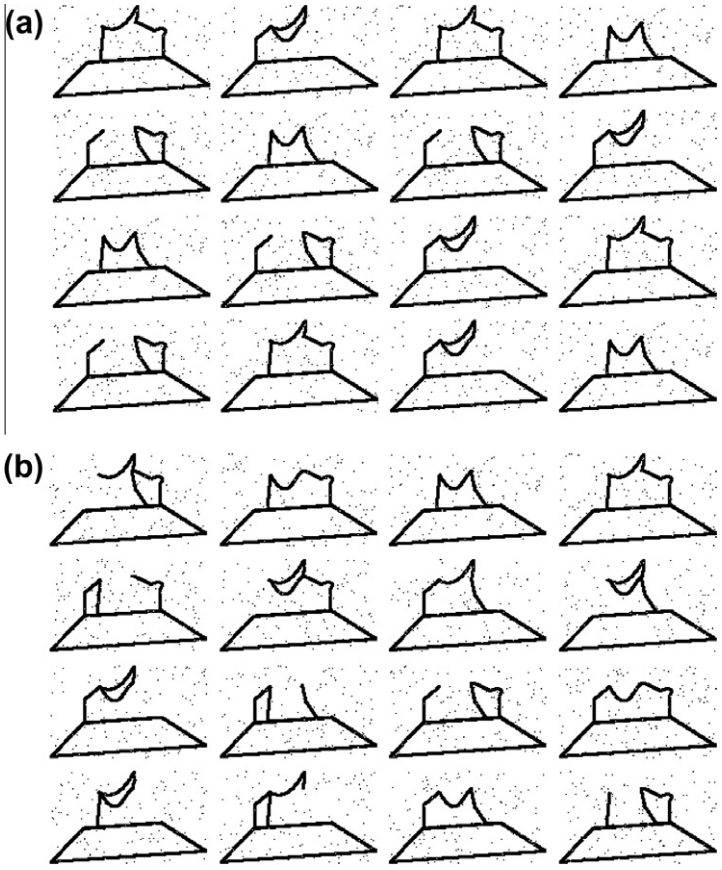

Fig. 5. Object sets used for Experiment 1. (a) The *correlated* training set, consisting of four copies of four images of different objects. (b) The *independent* training set. These two sets share four objects. The four objects missing from the *independent* training set forms the *unseen* objects used in testing.

and raw frequency effects. Finally, noise was added to each object by flipping each pixel in the image with probability $\frac{1}{75}$.

Each participant was shown either the *correlated* or *independent* training set. Participants were given their objects on printed cards (described in more detail below). The same test set of twelve objects was given to all participants in one of two random orderings (*test order*). Fig. 6a–c show how the twelve test objects group into three *test types*: four objects seen by the participant already (*seen*), four objects that had not been seen already by the participant that were composed of the same parts (*unseen*), and four objects created by deconstructing the images into different parts (*shuffled parts*) that still maintain the gross statistical properties of the objects (equal pixel variance and average pixel similarity to all other training and test sets). The *shuffled parts* stimuli were created by first taking the image formed by the union of all six parts and segmenting it into six different parts. We picked four of the 20 possible images we could make by combining three of the six resulting parts.

5.1.3. Procedure

Participants were given the 16 images appropriate to their conditions on business cards (width 3.5 in. by height 2.5 in.) randomly in front of them and given the following cover story:

> Recently a Mars rover found a cave with a collection of different images on its walls. A team of scientists believe the images could have been left by an alien civilization. The scientists are hoping to understand the images so they can find out about the civilization.

They were asked to alert the experimenter after "investigating the images" by "laying all the cards out on the table and organizing them in any way you think might help you learn about the images" and told that "no longer than 5–10 min is necessary." After they finished investigating the images, they were given the following test instructions:

> It looks like there are many more images on the cave wall that the rover has not yet had a chance to record. If the rover explored the cave wall further, which images do you think it would be likely to see?
> Your task is to rate how likely you believe it is that the rover sees each image as it explores further through the cave.
> In the booklet in front of you are twelve images, each on its own page. After you are finished rating each image, turn the page to the next image. Once you have turned to the next image, please DO NOT TURN BACK to any previous images.

To minimize memory effects, the images from the training set were not taken away from the participants. Each image was shown on a single page and participants were asked to generalize to the test

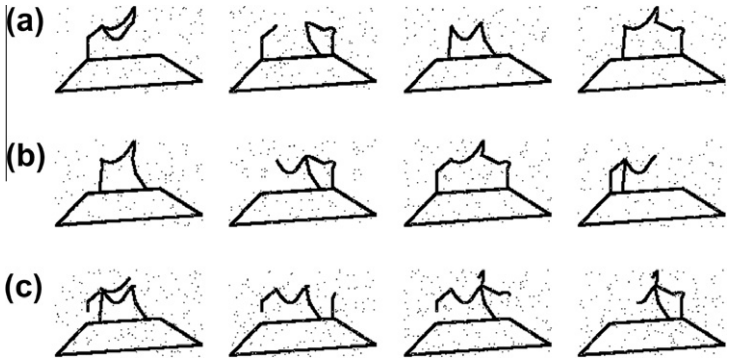

Fig. 6. The three sets of test images. (a) *Seen* for Experiment 1 and *unseen* for Experiment 2. (b) *Unseen* for Experiment 1 and *seen* for Experiment 2. The *seen* and *unseen* objects are switched for Experiments 1 and 2, with the *unseen* objects for Experiment 1 being the objects that appear in the *correlated* set of Experiment 2. (c) *Shuffled parts* for Experiments 1 and 2.

set (“Rate from 0-10 how likely you believe the rover is to see this image on another part of the cave wall”).

5.2. Results and discussion

Fig. 7 shows the responses of participants for each test object in the experiment and the predictions of three models (the IBP, an exemplar model, and a prototype model). Participant responses for the same object were averaged. We discuss the participant responses first. Subjects in *both distribution type* groups rated the three types of test objects differently ($F(2,48) = 40.72, p < 0.001$).[6] Importantly, participants in *each distribution type* group rated the *test types* differently as is shown by a two-way interaction ($F(2,48) = 8.77, p < 0.005$). There were no other main effects or interactions (all $F < 2$). As there were no major effects of *test order*, we collapsed over this variable in the subsequent pre-planned analyses.

Confirming our hypothesis, there was no difference between the *seen* and *unseen* image ratings for participants in the *independent* condition ($t(13) = -0.09, p = 0.93$), but there was for those in the *correlated* condition ($t(13) = 8.32, p < 0.001$). Participants in the *correlated* condition were more likely to generalize to the *seen* images than those in the *independent* condition ($t(26) = 2.06, p < 0.05$). Additionally, participants in the *independent* group were more likely to generalize to the unseen images than those in the *correlated* condition ($t(26) = 3.84, p < 0.001$). There was no difference between participants in the *independent* and *correlated* conditions on the *shuffled parts* images ($t(26) = 0.41, p = 0.69$). Finally participants in the *correlated* condition were not more likely to generalize to the *unseen* images than the *shuffled parts* images ($t(13) = 1.33, p = 0.21$). However, participants in the *independent* condition were more likely to generalize to the *unseen* images than the *shuffled parts* images ($t(13) = 5.39, p < 0.001$).

The second plot from the top of Fig. 7b shows the predictions of our model. The results from our experiment are qualitatively the same as the predictions by our model. The input representation of each object given to the model is a binary vector of the pixel values in its image. The predictions of the rational model are based on the probability of the new images given the images from either the *independent* or *correlated* set, and then averaged in the same way as the human data. The model predictions are computed by approximating the full posterior predictive distribution with the probability of the new images using the most likely features as determined by a Markov chain Monte Carlo simulation (see the Appendix for details). As there is a large difference in the probabilities of different types of test images for the IBP, we use an exponentiated Luce choice rule (Kruschke, 1992; Luce, 1959) to model the relationship between log probabilities and judgments (see the Appendix for details). In this case, this corresponds to raising the probabilities to the power γ that minimized the squared difference between the model and human responses and renormalizing. In this case, $\gamma = 1.5677 \times 10^{-3}$, which was fit by minimizing the mean squared error between the average ratings of participants and the IBP model for the *test images* in Experiments 1 and 2 (grouped and averaged by *test image* type). The quantitative fit of the model with the best-fitting γ value to human responses on the twelve *test images* in the *independent* and *correlated* conditions are $r = 0.93$ and $r = 0.98$, respectively (Pearson’s product-moment correlation coefficient).

As our task is essentially a categorization task, requiring people to generalize to the other members of an observed category, we need to demonstrate that our results cannot be explained using pre-existing psychological models of categorization. To rule out this alternative explanation, we investigated how an exemplar model and a prototype model would generalize to the test objects given each set of objects observed by participants as belonging to a category. We used an exemplar model similar to the popular Generalized Context Model (GCM; Nosofsky, 1986), taking each pixel as an input dimension and assuming that the attention weights for all dimensions were equal. The GCM is designed to be applied to dimensions that are derived from applying multidimensional scaling to the stimuli, but in this case our key question is what representation people form of these stimuli, making this inappropriate. Our goal in using an exemplar model here is simply to demonstrate that similarity

[6] All statistics concern participant ratings grouped into the three *test types* (*seen*, *unseen*, and *shuffled parts*) and then averaged.

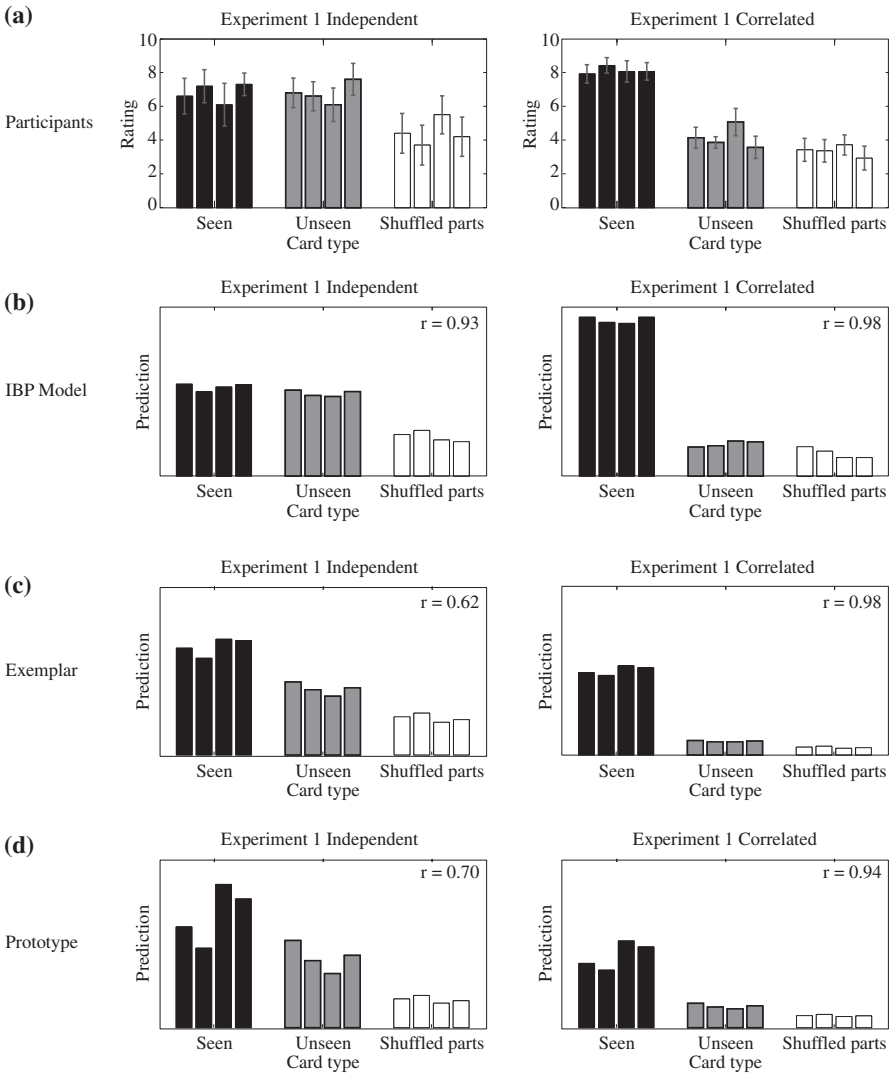

Fig. 7. Results of Experiment 1 for *test images* in left to right order given by Fig. 6. (a) The mean ratings participants made for test items as a function of training condition. Error bars show one standard error. (b) Predictions made by the IBP model as a function of training set. Notice the close qualitative correspondence to human performance. (c) Predictions made by an exemplar model. It incorrectly predicts different ratings of the *seen* and *unseen* images for the *independent* set. (d) Predictions made by a prototype model. It also incorrectly predicts different ratings of the *seen* and *unseen* images for the *independent* set.

to exemplars, expressed in the original raw pixel space, cannot account for the effects we see in the experiment. As the raw sensory data are not intended to be inputs to the GCM (only features in psychological space), this is not intended as a critique of the GCM.[7] For the prototype model, we predicted generalization based on the distance to the average of the seen objects (as in Reed, 1972).

[7] In fact, one interesting avenue for future research would be to use the features inferred by our model as the inputs to the GCM. This unified model would allow the GCM to be applied to stimuli without first eliciting human similarity judgments and forming a psychological space through multidimensional scaling.

Fig. 7c and d show the predictions of the exemplar and the prototype model, respectively. The exemplar models' predictions were made by calculating the category similarity with the pixels as features, based on the Euclidean distance between each test image and each input image. The prototype model predictions were made by calculating the category similarity with the pixels as features based on the Euclidean distance between each test image and the mean of all input images.[8] Like the predictive distribution for the IBP, there was a large difference in the category similarity of different types of test images for the exemplar and prototype models. We also used the exponentiated Luce choice rule to produce the values in the plot, applied directly to the category similarity (with $\gamma = 1.5 \times 10^{-3}$ for the exemplar model and $\gamma = 0.38$ for the prototype model, fit in the same manner as for the IBP model).

The exemplar and prototype models both have trouble explaining in judgments made by the human participants in the two conditions. The exemplar wrongly predicts a difference between the *seen* and *unseen* images for the *independent* images. This is one of the main problems with explaining the behavior of our participants with an exemplar model: a previously observed stimulus is typically rated higher than an unobserved stimulus. The quantitative fits of the exemplar model with the best-fitting γ value to the human responses for the *independent* and *correlated* conditions are $r = 0.62$ and $r = 0.98$, respectively (Pearson's product-moment correlation coefficient). Like the exemplar model, the prototype model predicts a difference between the *seen* and *unseen* images for the *independent* condition, which was not observed in participant responses. The quantitative fits of the prototype model with the best-fitting γ value to the human responses for the *independent* and *correlated* conditions are $r = 0.70$ and $r = 0.94$, respectively (Pearson's product-moment correlation coefficient). Though it is plausible that the exemplar and prototype models could explain human judgments using a representation inferred via multidimensional scaling or the IBP model, they are insufficient on their own to explain the results in this experiment.

6. Experiment 2: Feature learning from binary images with independent noise

Experiment 1 demonstrated that people infer features using distributional information as our model predicts. However, there were two possible confounds in the experiment. First, the observed differences could be due to the particular parts we correlated together, rather than distributional information. Second, each copy of the four *correlated* objects had identical noise, whereas each object in the *independent* condition had independent noise. In this experiment, we rule out these confounds by replicating the same effect using *correlated* images created by correlating a different set of parts than those used in Experiment 1 and by adding independent noise to each image in each set.

6.1. Methods

6.1.1. Participants

A total of 28 undergraduates from University of California, Berkeley, participated in the experiment in exchange for course credit. There were 14 participants in each of the *correlated* and *independent* conditions, with *test order* counterbalanced.

6.1.2. Stimuli

Like Experiment 1, Fig. 4a shows the images of parts and base that combine to form the objects shown to participants. There are two main differences between the stimuli in this experiment and those in Experiment 1: whereas in Experiment 1, the copies of the four objects in the *correlated* set had identical noise, each image in the *correlated* set of Experiment 2 was perturbed by independent noise (each pixel's value was flipped with probability $\frac{1}{75}$), and different parts were correlated together, which resulted in the *correlated* and *independent* sets shown in Fig. 8a and b. As the *correlated* set of this experiment is the same as the *unseen* test set used in Experiment 1, Fig. 6b, a, and c form the *seen*,

[8] For both models, the negative Euclidean distance was multiplied by a specificity parameter κ and then exponentiated. The best-fitting value for the exemplar model was $\kappa = 6.5273 \times 10^{-4}$ and for the prototype model was $\kappa = 0.37$, found in the same manner as γ.

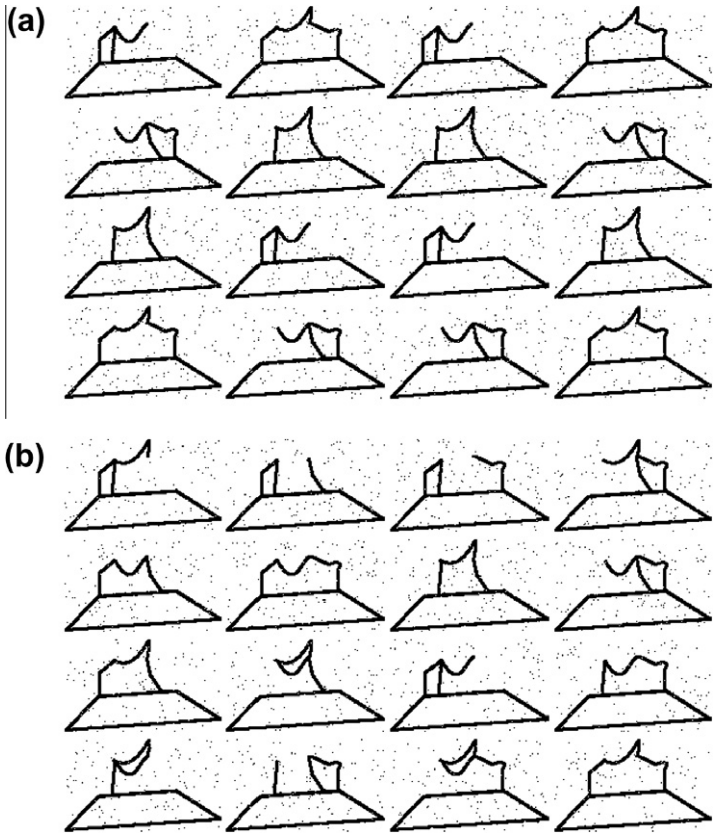

Fig. 8. Object sets used for Experiment 2. (a) The *correlated* training set, consisting of sixteen images of four objects, each with independent noise. (b) The *independent* training set. These two sets share four objects. The four objects missing from the *independent* training set forms the *unseen* objects used in testing.

unseen, and *shuffled parts* test sets for Experiment 2, respectively. Otherwise the stimuli were identical to Experiment 1.

6.1.3. Procedure

The procedure was identical to Experiment 1.

6.2. Results and discussion

As shown in Fig. 9, the results of Experiment 2 for participants and predictions of the three models replicate the results of Experiment 1. Like before, we first analyze the participant responses and then the predictions of the three models in order. Participants in both *distribution type* groups rated the three types of test objects differently ($F(2,48) = 21.38, p < 0.001$). Importantly, participants in each *distribution type* group rated the *test types* differently as is shown by a two-way interaction ($F(2,48) = 4.14, p < 0.05$). There were no other main effects or two-way interactions (all $F < 2$). There was a three-way interaction of *test type*, *test order*, and *distribution type* ($F(2,48) = 4.25, p < 0.05$) However, the effect is irrelevant to the question of whether people use distributional information as it is caused by participants in the first *test order*, *independent* condition rating the seen images higher than those in the second *test order*, *independent* condition (and was not replicated in any other experiment).

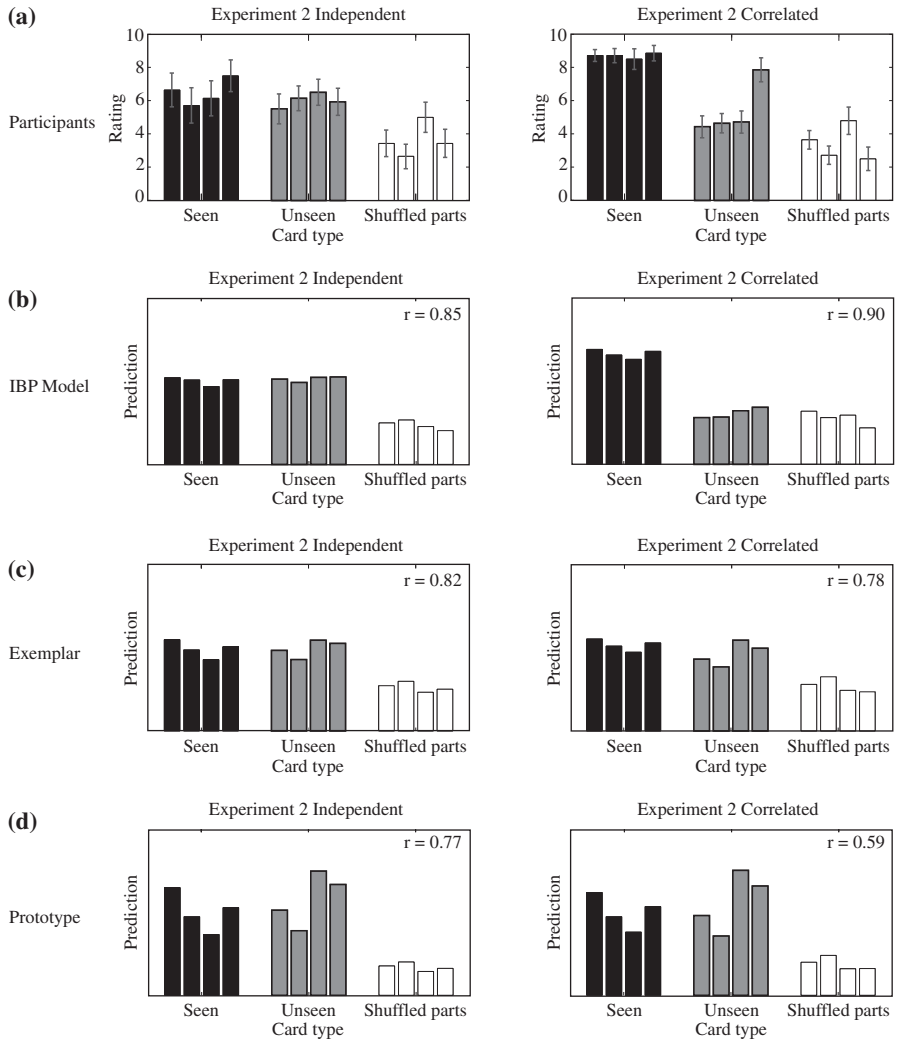

Fig. 9. Experiment 2 results for *test images* in left to right order given by Fig. 6. (a) The mean ratings participants made for test items as a function of training condition. Error bars show one standard error. (b) Predictions made by the IBP model as a function of training set. Notice the close qualitative correspondence to human performance. (c) Predictions made by an exemplar model. It incorrectly predicts little effect of *distribution type* on human ratings. (d) Predictions made by a prototype model. It also incorrectly predicts little effect of *distribution type* on human ratings.

As there were no major effects of *test order*, we collapsed over this condition in the subsequent pre-planned analyses.

Confirming our hypothesis, there was no difference between the *seen* and *unseen* image ratings for participants in the *independent* condition ($t(13) = 0.60, p = 0.56$), but there was for those in the *correlated* condition ($t(13) = 5.29, p < 0.001$). Participants in the *correlated* condition were more likely to generalize to the *seen* images than those in the *independent* condition ($t(26) = 2.16, p < 0.05$). However, we did not detect a difference between participants in the *independent* and *correlated* groups in how likely they were to generalize to the *unseen* images ($t(26) = 0.80, p = 0.22$). As can be seen in Fig. 9a, this is due to the high rating to the fourth *unseen* image and the effect was found in all other

experiments (this is probably due to the unanticipated high similarity between the fourth *unseen image* and second image in the *correlated* set). There was no difference between participants in the *independent* and *correlated* conditions on the *shuffled parts* images ($t(26) = 0.2, p = 0.80$). Finally participants in both the *independent* and *correlated* conditions were more likely to generalize to the *unseen* images than the *shuffled parts* images ($t(13) = 2.46, p < 0.05$ and $t(13) = 3.51, p < 0.005$, respectively).

Fig. 9b shows the predictions of our model, which were generated in the same manner as Experiment 1, but using the images from Experiment 2. The results from our experiment are qualitatively the same as the predictions by our model. The quantitative fit of the model with the same γ value as before to human responses on the twelve *test images* in the *independent* and *correlated* conditions are $r = 0.85$ and $r = 0.90$, respectively (Pearson's product-moment correlation coefficient).

Fig. 9c and d show the predictions of the exemplar and the prototype model, respectively. The exemplar and prototype models' predictions were made in the same manner as Experiment 1, but using the images from Experiment 2. Contrary to participant responses, the exemplar and prototype models both do not predict any major effect of *distribution type*. The sensitivity of participant responses to *distribution type* is a challenging result for prototype and exemplar models (particularly challenging is insensitivity of participants to the weak part correlations present in the *independent* conditions). The quantitative fits of the exemplar model using the same γ value as before to the human responses for the *independent* and *correlated* conditions are $r = 0.82$ and $r = 0.78$, respectively (Pearson's product-moment correlation coefficient). The quantitative fits of the prototype model using the same γ value as before to the human responses for the *independent* and *correlated* conditions are $r = 0.77$ and $r = 0.59$, respectively (Pearson's product-moment correlation coefficient).

7. Experiment 3: Feature learning from grayscale images

Although Experiments 1 and 2 established that people use statistical cues to infer feature representations as our model predicts, the images were very impoverished, using pixels that were either black or white. In this experiment, we demonstrate the same effect using computer-rendered 3-dimensional grayscale images. This provides a more stringent test of our hypothesis by replicating the result with more complex stimuli. Additionally, it showcases the power of our rational model, which can work with grayscale images as input representations.

7.1. Methods

7.1.1. Participants

A total of 98 undergraduates from University of California, Berkeley, participated in the experiment in exchange for course credit. There were 48 participants in the *correlated* condition and 50 participants in the *independent* condition.

7.1.2. Stimuli

Fig. 10 shows the objects used as stimuli, and location of the six parts used in the experiment. The part is a extrusion on the surface of the sphere (motivated by the stimuli used in Schyns & Murphy, 1994). The parts are equidistant from the focal point and single artificial light source of the image. The sphere has other extrusions on it and is connected on the bottom to a cylinder, which is the same in every image. To generate the images, we used the computer graphics renderer Blender (http://www.blender.org/). Fig. 11 shows four of the images used in the experiment.

This experiment had the same main manipulation as Experiments 1 and 2, *distribution type* had two levels: *correlated* and *independent*. As the parts we chose to correlate together did not have an effect in the previous experiments (the results of Experiments 1 and 2 were identical), we did not include a *training set* condition in this experiment. Additionally, there were no *shuffled part* images, so there were only two levels to the *test type* manipulation. Gaussian noise was added to each image, by adding to each pixel a random draw from a Gaussian distribution with a mean of zero and standard deviation

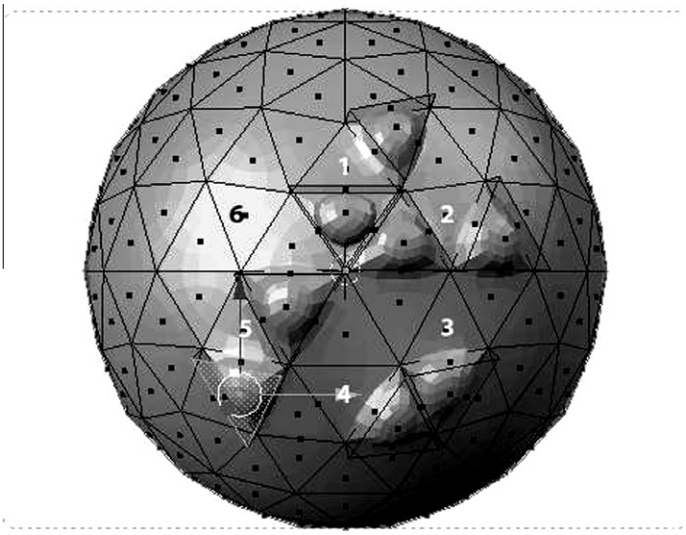

Fig. 10. Images used as stimuli in Experiment 3. Each number corresponds to the location of a part – a small extrusion on the surface of the sphere. Each object has three parts, which were assigned in the same manner as Experiments 1 and 2.

Fig. 11. Four of the Martian tools from the *independent* set. From left to right, the objects have parts 1, 2, and 6, parts 1, 2, and 5, parts 1, 2, and 3, and parts 1, 3, and 4.

of two. Otherwise, the design of this experiment was identical to Experiments 1 and 2, with the combinations of parts being used to generate each training set being the same as those in the *correlated* and *independent* conditions of the previous experiment.

7.1.3. Procedure

Participants were given the sixteen images appropriate to their conditions on cards (width 4.25 in. by height 5.5 in.) randomly in front of them and given the following cover story (which was similar to but slightly different from Experiments 1 and 2):

> Recently a Mars rover found a cave with a collection of Martian artifacts. A team of scientists believes the artifacts were used by an alien civilization as tools. The scientists are hoping to understand the artifacts so they can find out about the civilization.

They were asked to alert the experimenter after "investigating the images" by "laying all the cards out on the table and organizing them in any way you think might help you learn about the images" and told that "no longer than 5–10 min is necessary." Additionally, they were told that "although some cards may be the same, every card does not have the same image on it." After they finished investigating the images, they were given the following test instructions:

It looks like there are many more artifacts in the cave that the rover has not yet had a chance to record. If the rover explored the cave wall further, which artifacts do you think it would be likely to see?

Your task is to rate how likely you believe it is that the rover sees each artifact as it explores further through the cave. While rating the artifacts, please feel free to refer back to the cards that you explored.

In the booklet in front of you are eight images, each on its own page. After you are finished rating each image, turn the page to the next image. Once you have turned to the next image, please DO NOT TURN BACK to any previous images.

To minimize memory effects, the images from the training set were not taken away from the participants. Each image was shown on a single page and participants were asked to generalize to the test set ("Rate from 0-10 how likely you believe the rover is to see this object in another part of the cave").

7.2. Results and discussion

Fig. 12a shows the mean responses of participants in the experiment. Participants responses were grouped into the two *test types* (*seen* and *unseen*) and then averaged. The results replicate the main findings of Experiments 1 and 2: Participants who observe parts that occur independently over objects do not differentiate between old and new combinations of the parts ($t(48) = 1.16, p = 0.25$), but participants who observe the parts co-vary over objects do ($t(46) = 3.55, p < 0.001$). Fig. 12b gives the

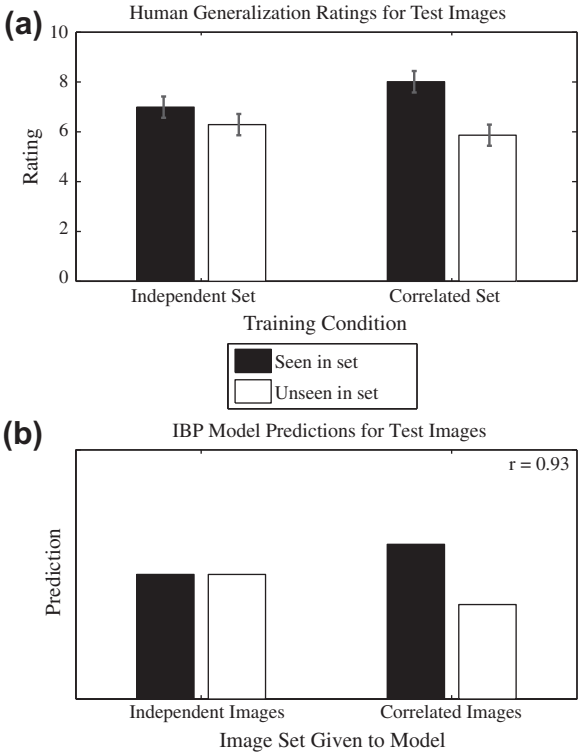

Fig. 12. Results of Experiment 3. (a) The mean ratings participants made for test items as a function of training condition. Error bars show one standard error. (b) Predictions made by the rational model as a function of training set. Notice the close qualitative correspondence to human performance. Replicating Experiments 1 and 2, the participants in the *independent* condition do not differentiate between the *seen* and *unseen* objects. Participants in the *correlated* condition differentiate between the *seen* and *unseen* objects.

predictions made by the best-fitting IBP model ($\gamma = 3.35 \times 10^{-5}$), using the exponentiated Luce choice rule to generate predictions, as in Experiments 1 and 2. The quantitative fit of the IBP with the best-fitting γ value to the human responses is $r = 0.93$ (Pearson's product-moment correlation coefficient).

A mixed-effects ANOVA corroborates the results of our planned t-test. There was an interaction between the training condition of the participants and their judgements on the *seen* vs. *unseen* objects ($F(1,47) = 5.72, p < 0.05$). There was no main effect of training condition ($F(1,47) = 0.33, p = 0.58$), suggesting participants in the two training conditions used the rating scale similarly. This is important as it rules out a potential alternate explanation that participants in the *independent* condition are simply rating everything higher.

8. Experiment 4: Conceptual feature learning

Experiments 1, 2, and 3 show that the features people identify for visually presented objects are sensitive to the distribution of parts across those object. This raises the question of how domain-general this phenomenon might be. Does distributional information only affect how visual features are learned or does it play a role in other domains? Experiment 4 explored whether distributional information affects feature learning when the objects and parts are conceptual. The experiment was analogous to the two previous experiments, other than being conducted in the conceptual domain. Participants learned about different facts about novel animals found from fossils in a Martian meteorite. Then, they were asked to rate how likely other animals with three facts were to be found on the Martian meteorite.

8.1. Methods

8.1.1. Participants

A total of 28 undergraduates from University of California, Berkeley, participated in the experiment in exchange for course credit. There were 14 participants in both the *correlated* and *independent* conditions.

8.1.2. Stimuli

Six facts about animals were used to construct the stimuli: (1) lays eggs, (2) moves fast, (3) has small claws, (4) spends most time on land, (5) has scaly skin, and (6) is a small herbivore. Each animal was described in terms of three of these facts, which play the same role as the three parts of the objects used in Experiments 1 and 2. The design of the experiment was analogous to that of Experiment 3, with the structure of the training and test sets being identical except for the substitution of facts for parts. Each animal was presented as a listing of the facts known about it.

8.1.3. Procedure

Participants were given the sixteen cards appropriate to their condition (width 3.5 in. by 2 in.) that had each fact on its own line (printed in Times New Roman font of size 23) in front of them in a random order. The following cover story was used (slightly different from the previous two experiments).

> Recently a meteorite from Mars landed in Antarctica. Scientists have excavated fossils of organisms from inside the meteorite, which they think correspond to 16 different species of Martian organisms. The scientists have discovered three facts about each species.
> The cards in front of you are examples of facts of all different species that were found.

They were asked to alert the experimenter after "investigating the images" by "laying all the cards out on the table and organizing them in any way you think might help you learn about the images" and told that "no longer than 5–10 min is necessary." After they finished investigating the images, they were given the following test instructions:

> It looks like there are many more fossils in the meteorite that the scientists have not yet had a chance to excavate. If the scientists excavated the meteorite more, what sorts of species do you believe they will find?

Your task is to rate how likely you believe it is that the scientists will excavate a fossil with the given three facts as they excavate more of the meteorite.
In the booklet in front of you are facts about eight species, each on its own page. After you are finished rating each species, turn the page to the next species. Once you have turned to the next species, please DO NOT TURN BACK to any previous species.

To minimize memory effects, the images from the training set were not taken away from the participants. Each image was shown on a single page and participants were asked to generalize to the test set ("Rate from 0-10 how likely you think scientists are to find a species with these facts in the meteorite").

8.2. Results and discussion

Fig. 13a shows the mean responses of participants in the experiment. Participants responses were grouped into the two *test types* (*seen* and *unseen*) and then averaged. The results replicate the main findings of the previous two experiments: Participants who observe parts that occur independently over objects do not differentiate between old and new combinations of the parts ($t(26) = 1.72$, $p = 0.10$), but participants who observe the parts co-vary over objects do ($t(26) = 3.49, p < 0.005$). However, in this case, the participants in the *independent* condition is trending towards significance. This is not surprising as the parts are presented one-by-one each on their own line on a card and so, we

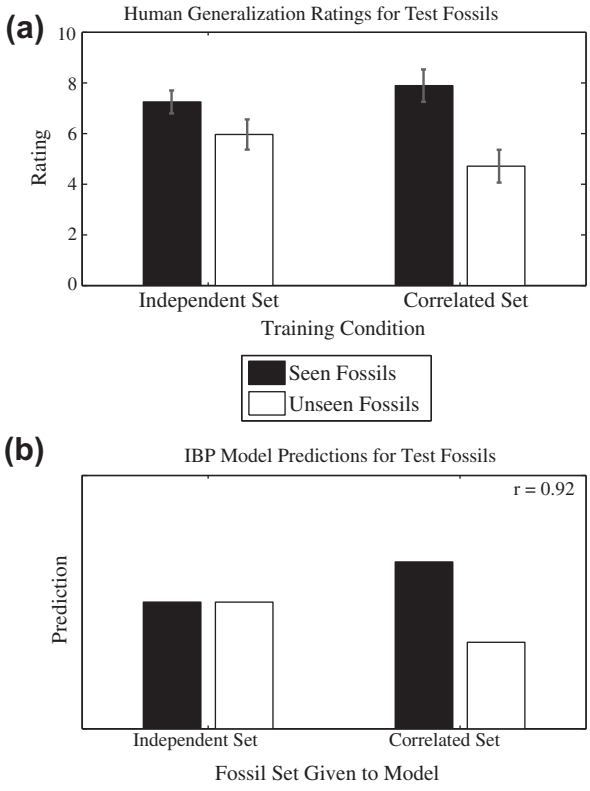

Fig. 13. Results of Experiment 4. (a) The mean ratings participants made for test items as a function of training condition. Error bars show one standard error. (b) Predictions made by the rational model as a function of training set. Notice the close qualitative correspondence to human performance. Although participants in both conditions differentiate between the *seen* and *unseen* objects, participants in the *correlated* condition differentiate more than those in the *independent* condition.

promote a strategy where participants respond by explicitly judging the number of parts in common between test objects and objects in the training set. This strategy predicts participants in the *independent* condition should prefer the *seen* to the *unseen* objects and thus, we should expect some differences in this case. Although a mixed effects ANOVA showed no main effect of *training condition* ($F(1,26) = 0.16$, $p = 0.69$), there was an interaction between the *training* and *test types* conditions ($F(1,26) = 7.73, p = 0.01$) (there was also a main effect of *test type*, $F(1,26) = 42.97, p < 0.001$). The interaction demonstrates that though participants in each condition overall expect the same number of objects, participants in the *independent* condition rate the *seen* and *unseen* objects more evenly than those in the *correlated* condition.

Fig. 13b gives the predictions made by the best-fitting IBP model ($\gamma = 1.65$), which was found by minimizing the square distance between the model and human responses and renormalizing. The quantitative fit of the IBP with the best-fitting γ value to the human responses is $r = 0.92$ (Pearson's product-moment correlation coefficient).

9. General discussion

What principles do people use to form the basic units to represent an observed set of objects? In order to explore this question, we performed a rational analysis of how people should solve this problem of feature learning. We used a nonparametric Bayesian model to infer arbitrarily rich feature representations, allowing an infinite number of features but expecting that only a finite subset will be expressed in any collection of objects. This model predicts that people's inductive generalizations should be sensitive to the amount of co-variation present in the parts that are used to create objects. Four behavioral experiments confirmed this prediction. Experiments 1 and 2 demonstrated that when given objects created from parts that occur independently, people generalize to novel combinations of the parts, but do not generalize when given objects whose parts strongly co-vary. Experiment 3 replicated this result with more realistic grayscale three-dimensional rendered objects. Experiment 4 generalized this result to the conceptual domain. Taken together, these results suggest that people use distributional information as a domain-general cue for forming feature representations and making generalizations.

In the remainder of the paper, we consider some of the issues raised by our analysis. First, we compare our approach to previous models of feature learning from the psychological and machine learning literatures. We highlight some of the differences from previous psychological approaches, and show that classic dimensionality reduction techniques from machine learning do not capture the pattern of results we observed in our experiments as well as our model. We then turn to ways in which our account can be extended. We show that imposing a proximity bias on the feature image prior helps the model infer more psychologically plausible visual features, and discuss how our abstract computational-level approach might connect to accounts of the cognitive processes behind feature learning at the algorithmic level. Finally, we consider some of the limitations of our analysis and suggest some directions for future research, and conclude the paper.

9.1. Comparison with other approaches to feature learning

While Experiments 1 and 2 were designed to rule out alternative explanations based on popular psychological models of categorization, previous work in both psychology and machine learning has produced models that can be applied to feature learning. In this section we consider these other models, comparing them with our approach and analyzing their predictions.

9.1.1. Psychological models

Goldstone (2003) introduced a neural network model of feature learning, which uses a soft constraint-satisfaction process to find good feature representations of objects. This model provided an important first step towards automatically inferring the features of objects, but it has one limitation that ours does not possess. The number of features must be specified in advance. This is a serious issue for an analysis of human feature learning because it does not allow us to directly compare different

feature set sizes – a critical factor in capturing unitization and differentiation phenomena.[9] In contrast, the use of a nonparametric Bayesian approach allows us to adaptively learn the appropriate number of features to be used in a representation based on the observed data, while allowing potentially infinitely many features.

Orban, Fiser, Aslin, and Lengyel (2008) investigated how the human perceptual system learns to group objects that seem to arise from a common cause, defining an ideal observer model that can be used to pick out "visual chunks" consisting of objects that tend to co-occur. This work uses a Bayesian model that can vary the number of causes it identifies, but assumes indifference to the spatial position of the objects and that the basic objects themselves are already known, with a binary variable representing the presence of an object in each scene being given to the model as the observed data. By assuming the basic units given to the model are whole parts instead of pixels (as our model does), the input to their model solves a part of the problem we are interested in (how does the human perceptual system identify these primitives?). In addition, it cannot explain results like those in Experiment 3, where the features need to be identified from continuous input data. However, once those primitives are provided, the assumptions of this model are similar to those behind our nonparametric Bayesian model.[10]

Although computational modeling of psychological visual feature learning is a relatively unexplored area, work on language acquisition has provided many computational models of how people learn linguistic representations without explicit instruction. Primarily, researchers have explored how people learn grammars (Solan, Horn, Ruppin, & Edelman, 2005) and segment continuous streams of phonemes into words (Batchelder, 2002; Brent, 1999; Goldwater, Griffiths, & Johnson, 2009; Perruchet & Vinter, 1998; Venkataraman, 2001). These models share with our analysis an interest in the use of distributional information in determining discrete representations. However, they are not directly applicable to inferring the features that appear in two-dimensional images without significant modification because they assume the representational units are negatively correlated, whereas they are independent in our model. The models used by Brent (1999), Goldwater et al. (2009) share with our model a grounding in nonparametric Bayesian statistics, which provides a general-purpose tool for solving the problem of determining the amount of structure expressed in observed data.

9.1.2. Machine learning methods

Another potential approach for deriving human feature learning models is to borrow related computational methods from machine learning. The matrix factorization formulation of the computational problem makes it clear that the goal is to find a smaller encoding for a set of observed objects. In the machine learning literature, this is known as dimensionality reduction. Although a number of different dimensionality reduction approaches have been proposed in machine learning, we focus on two of the most influential methods: principal component analysis and independent component analysis.

Principal component analysis (PCA) encodes a set of given objects in terms of the eigenvectors of its covariance matrix. Usually only some K eigenvectors have non-negligible weight (eigenvalues) and these are taken to be the redescription of the data (for an excellent introduction to linear algebra written for cognitive scientists see Jordan, 1986). This procedure is well motivated mathematically, being equivalent to encoding a set of objects in the K-dimensional orthogonal subspace that captures the greatest proportion of the variance in the observed data (Bishop, 2006) and has been previously proposed as a potential method for inferring psychologically valid features (Abdi, Valentin, & Edelman, 1998). Heuristic methods are usually used to select K, but PCA can also be extended to automatically infer the number of features (Minka, 2001).

Although the features inferred by PCA are uncorrelated, they are not necessarily independent. Independent component analysis (ICA) infers features that are statistically independent form each other

[9] The model introduced by Goldstone (2003) may be able to infer the number of features by specifying a large number of potential features, but encoding a constraint that biases it towards leaving a large number of the features unused. However, it still introduces an artificial theoretical upper bound on the number of features expressed in a set of objects.

[10] Technically, these two models represent different philosophies towards the problem of determining the dimensionality of observed data. The model proposed by Orban et al. (2008) assumes that there is a finite number of visual chunks, and tries to infer this number. Our model assumes that there are infinitely many features, of which only a finite number are observed. The difference between these two approaches has been explored in detail in computer science and statistics (Green & Richardson, 2001; Rasmussen & Ghahramani, 2001).

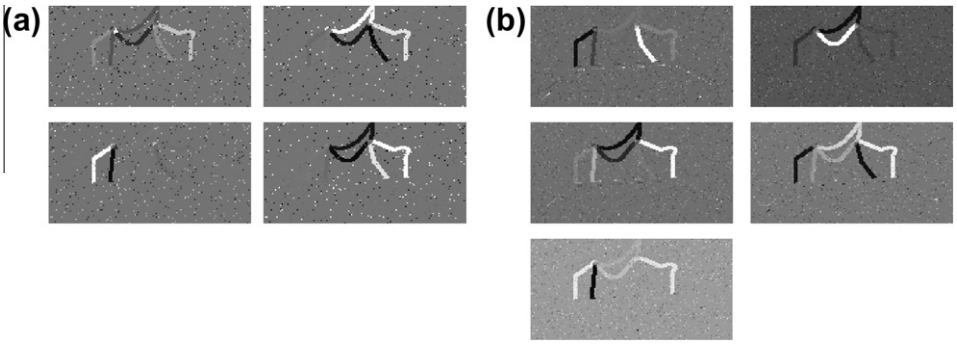

Fig. 14. Features inferred by Principal Component Analysis (PCA) when given objects whose parts either co-vary or vary independently. The features are presented as a color map, where negative weights are colored black and positive weights are colored gray. On left, features inferred by PCA given the *correlated* set in Experiment 1 and on right, features inferred by PCA given the *independent* set in Experiment 1. Only non-noise features are shown.

(Hyvarinen, Karhunen, & Oja, 2001). This idea is similar to approaches used to model visual cortex (Bell & Sejnowski, 1995; Olshausen & Field, 1996), and has recently been explored as a way of explaining human dimensionality reduction in multiple domains (Hansen, Ahrendt, & Larsen, 2005). Olshausen and Field (1996) have shown that ICA with an added sparsity bias given a set of natural scenes infers features that are very similar to proposed neural representations used in the primary visual cortex. However, it is unclear whether or not the close correspondence is due to the sparsity bias or independence between components (or even if the features used in the primary visual cortex are behaviorally relevant in our task). Thus, it is unclear whether or not the way ICA uses independence information will be successful for this simulation.

We can compare the predictions of these two dimensionality reduction techniques with the predictions of our rational model on the *correlated* and *independent* object sets used in our experiments. Additionally, we can investigate how each technique generalizes to new objects by looking at how well the best-fitting predictions given each object set can capture the pattern of results observed in Experiments 1 and 2. The predictions are formed by averaging over each type of test object the amount of reconstruction error formed from projecting each of the test objects onto the subspace learned by the dimensionality reduction technique.[11] The best-fitting predictions are found by transforming the reconstruction distance using the exponentiated Luce choice rule with the parameter value that minimizes the mean squared error between the model predictions and the human data (see Appendix).

Fig. 14 shows the features inferred by PCA in Experiment 1 (similar features are inferred given the objects from Experiment 2), restricting the number of recovered dimensions to include only those that do not capture noise in the data. The features inferred from the *correlated* set of Experiment 1 (shown in Fig. 8) are the shared part and the three sets of anti-correlated parts in the training set (e.g., the feature on the bottom left of Fig. 14a captures that the third and seventh parts never occur together). The features inferred from the *independent* set capture the minor correlations between parts (shown in Fig. 14b). Thus, the features inferred by PCA do not use statistical information in the same way as the IBP model or as our intuition would expect. Fig. 16a and b show the best-fitting predictions from PCA given each object set found in the same manner as for the other models ($\gamma = 0.3082$). It incorrectly predicts a weaker effect of *distribution type* than we found in Experiment 2. Quantitatively, the correlation between its predictions and human responses are $r = 0.84$, $r = 0.98$, $r = 0.86$, and $r = 0.80$ for the *independent* and *correlated* conditions of Experiments 1 and 2, respectively. Though the fit of PCA is slightly higher than the fit of the IBP model for the *independent* condition of Experiment 2 (by 0.01), its fit is much lower (about 0.1) for the *independent* condition of Experiment 1 and *correlated* condition of Experiment 2.

[11] Since the reconstruction error is monotonically related to the predictive probability of the test objects given the training objects, it is equivalent to using the predictive probability directly.

Fig. 15 shows the features inferred by ICA (using the FastICA software package Hyvarinen, 1999) when given (a) the *correlated* and (b) the *independent* sets of objects of Experiment 1 (similar features are inferred given the objects from Experiment 2). For both cases, we asked ICA to learn sixteen components and then only present the non-noise features. The features inferred from the *correlated* set by ICA (shown in Fig. 15a) have the same problem as those inferred by PCA. The features inferred by ICA are somewhat better than those inferred by PCA when given the *independent* set. For example, the top two features inferred by ICA are the parts used to create the *independent* set. However, the feature representation is still intuitively unappealing. The bottom right feature captures the small negative correlation between two of the parts in the *independent* set. This is because sixteen of the total twenty possible objects were shown. So, there are small non-zero correlations between some parts in the object set. Unfortunately, this yields a feature that does not expect two of the parts to occur together in a single object, which means the feature representation inferred by ICA differentiates between the *seen* and *unseen* objects (the two parts occur together in some of the objects in the *unseen* set). Fig. 16c and d show the best-fitting predictions from ICA given each object set found in the same manner as for the other models ($\gamma = 0.0726$). It incorrectly predicts a weaker effect of *distribution type* than we found in Experiment 2. Quantitatively, the correlation between its predictions and human responses are $r = 0.81$, $r = 0.98$, $r = 0.81$, and $r = 0.78$ for the *independent* and *correlated* conditions of Experiments 1 and 2, respectively. Like PCA, its fit to participant responses is much lower (about 0.1) for the *independent* condition of Experiment 1 and *correlated* condition of Experiment 2.

Dimensionality reduction is a growing literature, and while PCA and ICA are the most common methods, it may be that other methods from machine learning or computer vision will be able to capture the effects of distributional information on feature learning. For example, Ullman (2007) describe a method for visual feature learning that infers features by picking those that provide the most amount of information to an object's category relative to the other categories in the data set. This approach cannot be used for the sets of objects we analyzed in this paper, since no category information is provided, but it would be interesting to compare the predictions that result from this kind of approach to human judgments in future work.

9.2. Encoding a proximity bias in the feature image prior

The model that we used in most of our analyses made a very simple assumption about the structure of features, asserting that the prior distribution for each pixel was independent. This simplification makes it easier to work with the model, but ignores a large literature in perception showing that people have strong expectations about the forms that parts of objects will take (Biederman & Cooper, 1991; Braunstein, Hoffman, & Saidpour, 1989; Palmer, 1999). One consequence of this simplifying assumption was seen in the features inferred from the stimuli used by Shiffrin and Lightfoot

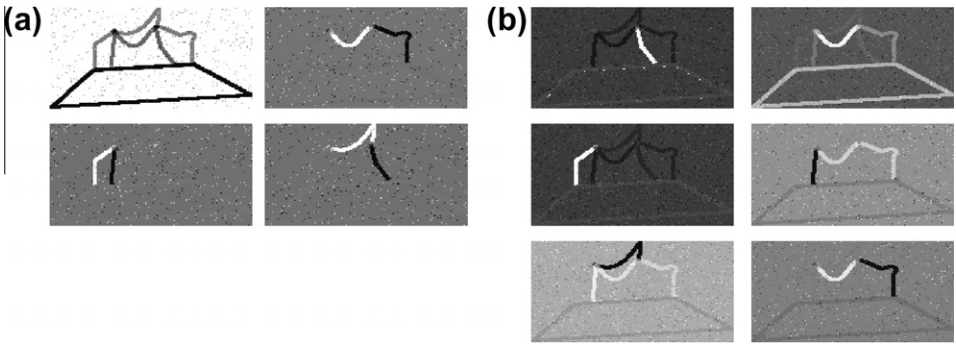

Fig. 15. Features inferred by Independent Component Analysis (ICA) when given objects whose parts either co-vary or vary independently. The features are presented as a color map, where negative weights are colored black and positive weights are colored gray. On left, features inferred by ICA given the *correlated* set in Experiment 2 and on right, features inferred by ICA given the *independent* set in Experiment 2. Only non-noise features are shown.

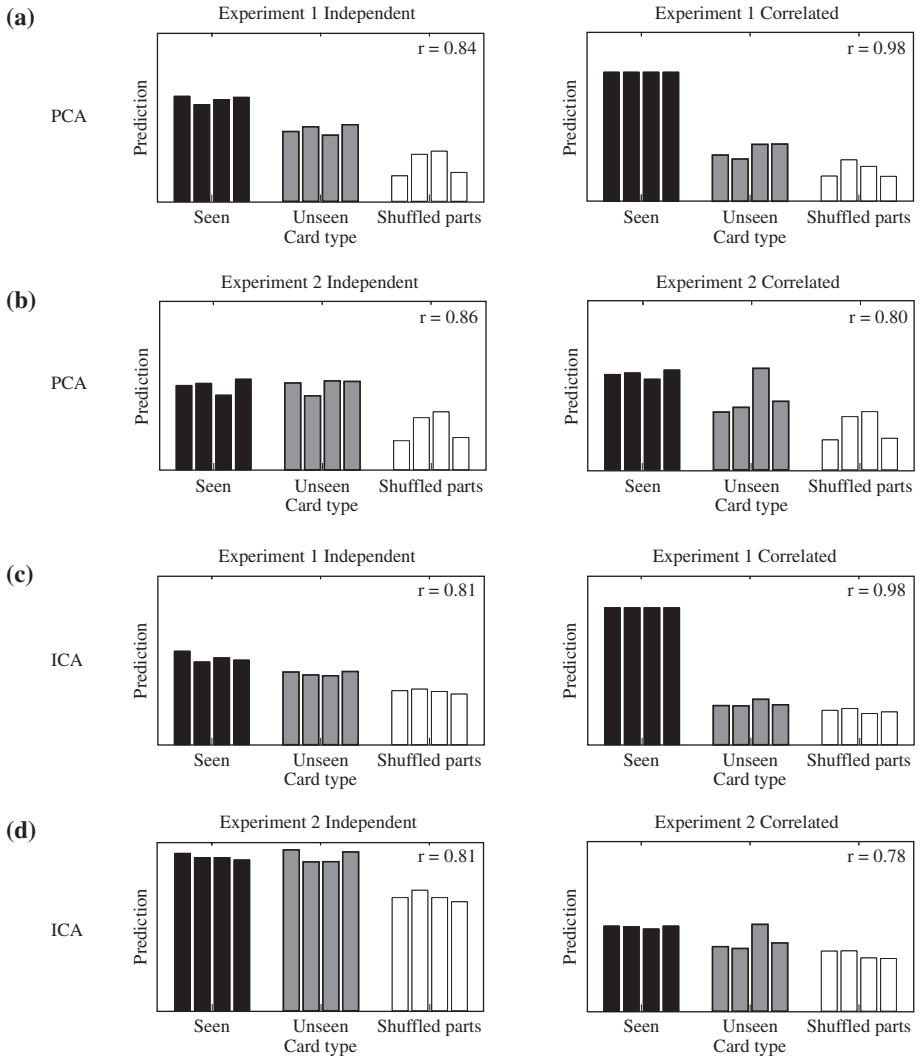

Fig. 16. The best-fitting predictions of Principal Component Analysis (PCA) (a) and (b) to human results of Experiments 1 and 2 respectively and the best-fitting predictions of Independent Component Analysis (ICA) (c) and (d) to human results of Experiments 1 and 2. PCA predicts a weaker effect of *distribution type* than produced by participants in Experiments 1 and 2. ICA predicts a weaker effect of *distribution type* than produced by participants in Experiment 2.

(1997), which were disconnected and contained "speckled holes", both of which seem psychologically implausible. By changing the feature image prior such that it incorporates more realistic expectations, we can infer features that are more psychologically plausible. We will show how this can be done for the noisy-OR likelihood, although a similar extension is possible for the linear-Gaussian likelihood.

For the noisy-OR model, the prior probability distribution on **Y** assumes each element is independently drawn from a Bernoulli distribution. This misses some obvious psychological constraints on the types of features we learn, such as proximity (i.e., all else equal, features should be connected). We can introduce a proximity bias by taking our prior to be a distribution (called the Ising model; e.g., Geman & Geman, 1984) on the elements of **Y** such that neighboring pixels tend to have the same value. Let ρ

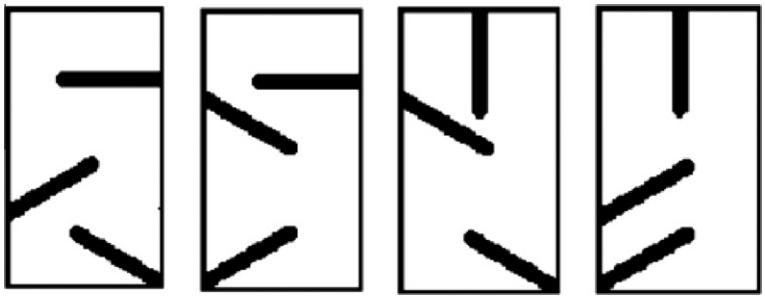

Fig. 17. Features learned by the model given the objects from Experiment 1 of Shiffrin and Lightfoot (1997) using the proximity bias feature image prior. Note that the speckled holes have been filled in using proximity information resulting in more psychologically plausible features.

reflect the propensity for two neighboring pixels to share the same value ($\rho > 0.5$ indicates neighbors are more likely to be the same), and θ be the propensity for each pixel to be on ($\theta > 0.5$ indicates pixels are more likely to be on). Neighboring pixels are connected by edges ($e = \{d_1, d_2\}$, where d_1 is the index of the first pixel), with C the set of edges between neighboring pixels. This gives us the following prior probability distribution on $\mathbf{Y}$ (up to a proportionality constant):

$$P(\mathbf{Y}) \propto \left(\prod_{k,d} \theta^{y_{k,d}}(1-\theta)^{y_{k,d}}\right)\left(\prod_{\{d_1,d_2\}\in C} \rho^{d_1 d_2}(1-\rho)^{1-d_1 d_2}\right). \tag{6}$$

Calculating the normalization constant for this distribution is intractable and thus, approximation techniques are used to find $\mathbf{Y}$ under this prior (see the Appendix for details).

Using the feature image prior with a proximity bias, the model infers features that are more connected, and thus more psychologically plausible than before. For example, consider the aforementioned case of inferring features given the images from one experiment of Shiffrin and Lightfoot (1997). The original model infers features with "speckled holes", as shown in Fig. 3. When the model uses the feature image prior with a proximity bias, the holes get filled in because the prior penalizes the neighboring pixels not all being on (shown in Fig. 17). Thus, we have incorporated a proximity constraint simply by changing the prior probability on feature images. As we develop better models of other perceptual principles (see for example Zhu, 1999), our model can use these in the same way to improve the psychological plausibility of the features it infers.

One future direction is exploring how distributional and proximity information interact using our model with the proximity bias feature image prior.[12] Goldstone (2000) showed that when a category-irrelevant part is flanked by two category-relevant parts (forming one contiguous unit), the conjunction of the three parts are learned as a single feature. This demonstrates the importance of proximity information when features are learned through categorization training. If categorization information is included into our model (by either including it as part of the sensory data of an object as in Austerweil & Griffiths, 2009, or by including hierarchies into our model, Thibaux & Jordan, 2007), it would infer a contiguous feature (containing all three parts) when the feature image prior with a proximity bias is used (but our original model would infer a feature only containing the two category-relevant parts).

Our model predicts that parts that are less good from a proximity standpoint need to vary more independently than those that are good from a proximity bias. Another interesting line for future work would be to explore this prediction and more generally how proximity information affects perceptual feature learning. Although Goldstone (2000) showed that proximity information is used to infer features through categorization training, it is unclear how proximity information is used when features are inferred without categorization training or how proximity interacts with distributional and category information.

[12] We thank Rob Goldstone for this suggestion.

9.3. Connecting to the algorithmic level

Our goal in this paper was to provide a rational analysis of the effects of distributional information on feature learning, which is framed entirely at Marr's (1982) "computational" level, identifying the nature of the underlying computational problem and its optimal solution. The results of our experiments suggest that people are sensitive to distributional information in the way that this optimal solution predicts. However, this raises the question of what cognitive processes people might be using to solve this problem – what Marr termed the "algorithmic" level. Linking the computational and algorithmic levels represents an important challenge for future work on feature learning.

Previous psychological models of feature learning provide clues as to possible algorithmic-level models that might approximate the behavior of our rational model. The model introduced by Goldstone (2003) is one promising candidate, particularly if a similar penalty for the number of features is introduced. Our experiments were designed to provide people with resources similar to those of an ideal agent, not imposing time or memory constraints by allowing people free inspection of the stimuli. Conducting experiments in which these factors are manipulated may provide a way to tease apart the predictions of different algorithmic-level models.

Order effects provide another important source of evidence about underlying cognitive processes. Our analysis assumes that all stimuli are available when the learner is inferring a representation, which was the case in our experiments. However, presenting the stimuli one at a time might provide a way to identify effects of the order of presentation. Our rational model cannot predict order effects, since the posterior distribution $P(\mathbf{Z}, \mathbf{Y}|\mathbf{X})$ is invariant to the order of the stimuli. However, recent work exploring approximations to rational models based on Monte Carlo methods provide tools for defining models that incrementally make inferences as data are presented (Levy, Reali, & Griffiths, 2009; Sanborn, Griffiths, & Navarro, 2006). One popular algorithm for incremental inference, a particle filter, has already been derived for models based on the IBP (Wood & Griffiths, 2007), and in future work we hope to examine whether this "rational process model" can account for order effects in feature learning (e.g., Schyns & Rodet, 1997).

9.4. Limitations and future directions

In this article, we presented nonparametric Bayesian modeling as a general tool for understanding how people learn representations. We showed how nonparametric Bayesian models naturally incorporate arbitrary amounts of structure into their representations just like cognitive representations, and demonstrated that people make similar inferences from distributional information. However, there are a number of directions in which this work could be extended, including providing a more direct connection to perceptual learning, and evaluating models that make different assumptions about the relationship between the features of objects.

Our experiments were not designed to evaluate whether people had genuinely altered their perception of the stimuli. Like Schyns and Rodet (1997), we infer what features our participants use based on the distinctions they make in their generalization behavior. This method leaves open the possibility that people did not form new percepts of these stimuli, but rather that they changed the amount of attention or weight that they assigned to existing features. Other experimental methods that have been used in the literature on perceptual learning, such as participant delineation of features (Schyns & Murphy, 1994) and speeded part-whole judgments (Pevtzow & Goldstone, 1994), provide a more direct test of whether there has been a change in the percepts that people form. In future work, it would be interesting to examine whether the manipulation of distributional information we used in our experiments actually alters people's percepts in this way.

Another interesting avenue for research is to extend our rational model to capture other aspects of human feature learning. For example, translation invariance is an important part of recognizing shapes, and may play a similarly significant role in people's recognition of features across objects. We recently introduced a variant of the IBP that incorporates transformations, allowing features to be modified when they occur in an object (Austerweil & Griffiths, 2010). This model associates a hidden transformation with each feature, which is learned from the observed objects (though a basic set of transformations is assumed). Allowing spatial transformations provides the opportunity to learn

that features are translation invariant. This presents the opportunity to compare this model against other methods from machine learning for inferring translation-invariant features (e.g., Spratling, 2006) as part of a broader account of human feature learning.

The IBP assumes that the features of objects are independent of one another *a priori* (the features are not independent in the posterior distribution), and that there is no relationship between the features of particular objects beyond the general popularity of a feature. Both of these assumptions can be relaxed, resulting in models that may capture people's expectations more accurately in different domains. Infant studies by Spelke (1990) show that the earliest cue to object continuity is time. This suggest that parts of objects moving together over time could be another potentially strong cue to learning not only what objects are, but also the features of objects. Thus, another future direction would be to compare feature learning by IBP models that are sensitive to time (e.g., Van Gael, Teh, & Ghahramani, 2009) to how people learn features with temporal information. Other variants on the IBP make it possible to specify relationships among objects (Miller, Griffiths, & Jordan, 2008), or to assume that objects belong to different categories (Thibaux & Jordan, 2007), resulting in priors that may provide a more accurate account of human feature learning in cases where these assumptions are appropriate.

Finally, our exploration of a proximity-based prior on feature images illustrates how particular constraints on the form of features can be introduced into our model. However, this only scratches the surface of such constraints. In perception, there is a long literature demonstrating that people prefer to interpret objects using features that have Prägnanz or "goodness," where the "goodness" of a feature refers to a subjective feeling of being better than other possible feature decompositions (generally based on classic Gestaltist principles of organization, Palmer, 1977, 1999). There are many computational proposals to formalize the introspective feeling of goodness; usually based on a measure of simplicity in a coding language (Hochberg & McAlister, 1953; Leeuwenberg, 1978; van der Helm & Leeuwenberg, 1996) or the amount of invariance across different transformations (Garner, 1974; Palmer, 1991). In our framework, any of the proposals can be formulated as a more elaborate prior distribution on the types of feature images allowed. Our approach may thus provide a means of defining novel computational models of different Gestalt constraints.

9.5. Conclusion

Through a series of four behavioral experiments, we demonstrated that people use distributional information to infer feature representations for novel objects in a way that is predicted by a rational analysis of feature learning. In particular, we have shown that people generalize category membership from the same objects differently depending on the statistical distribution of parts in the set of objects. The differences in their generalization behavior suggest that they represent the same objects using different features. This is a result that was predicted by our rational model, which is based on nonparametric Bayesian statistics. The results of our experiments cannot be explained using exemplar or prototype models, previous models of feature learning, or dimensionality reduction techniques from machine learning. We view these results as a first step towards answering the question raised by Garner (1974) and the Gestaltists of how a cognitive system forms representations for sets of stimuli, allowing an object to be represented in different ways based on its context.

Acknowledgments

We thank Rob Goldstone, Stephen Palmer, Karen Schloss, Tania Lombrozo, Greg Murphy, Charles Kemp, Amy Perfors and Eleanor Rosch for insightful discussions, Frank Wood for providing code for the Noisy-OR IBP model, and Brian Tang, David Belford, Shubin Liu, and Julia Ying for help with experiment construction, running participants, and data analysis. A preliminary version of some of the computational results was presented at the 21st Neural Information Processing Society and Experiments 1 and 2 were presented at the 31st Annual Meeting of the Cognitive Science society. This work was supported by Grant No. FA9550-07-1-0351 from the Air Force Office of Scientific Research and Grant No. IIS-0845410 from the National Science Foundation.

Appendix A. Simulation details

In this section, we describe how we calculated the predictions of each model for the experiments.

A.1. Modeling human responses using the Indian buffet process

The Indian Buffet Process (IBP) provides a prior probability distribution on the feature ownership matrix $\mathbf{Z}$, but this does not fully specify how to infer the probability of new objects $\mathbf{x}_{\text{test}}$, given a set of previously observed objects $\mathbf{X}$ (the form of the human data from our three experiments). First, we need to infer $P(\mathbf{Z}|\mathbf{X})$, which is the inferred feature representation for the set of observed objects. To infer $P(\mathbf{Z}|\mathbf{X})$, we invert it using Bayes Theorem, which gives us two components $P(\mathbf{X}|\mathbf{Z})$ (the likelihood or the probability of the objects given the feature representation) and $P(\mathbf{Z})$ (given to us by the IBP). We use the noisy-OR likelihood function (given by Eq. (5)) for binary input data (Experiments 1, 2 and 4) and the linear-Gaussian likelihood function (given by Eq. (4)) for grayscale input data (Experiment 3). The matrix of feature weights (the image of each feature when instantiated) is $\mathbf{Y}$. To infer the feature weights for the noisy-OR model, we put a Bernoulli prior with parameter ϕ (controlling the number of pixels we expect to be on in each feature) on each pixel of the noisy-OR likelihood ($y_{kd} \overset{iid}{\sim} \text{Bernoulli}(\phi)$). To infer the feature weights for the linear-Gaussian model, we put a multivariate Normal prior on the feature weight matrix ($\mathbf{Y} \sim \text{Normal}(0, \sigma_Y^2\mathbf{I})$) where σ_Y^2 is a parameter controlling how much variation we expect within a feature. So, we infer the feature weight matrix and feature ownership matrix jointly given the observed objects ($P(\mathbf{Z},\mathbf{Y}|\mathbf{X})$), which we will then use to predict new objects.

Unfortunately computing the posterior distribution on feature weight matrices is intractable. Thus, we approximate it using Gibbs sampling, a Markov chain Monte Carlo method in which a sequence of samples are generated that ultimately converge to the posterior distribution. Gibbs sampling algorithms have been previously derived by Wood et al. (2006) for the noisy-OR model and Griffiths and Ghahramani (2006) for the linear-Gaussian model. An alternative to Gibbs sampling is variational inference (Doshi-Velez, Miller, Van Gael, & Teh, 2009), which can be more efficient for larger sets of images than those we examine here. For the noisy-OR model, the joint distribution used in Gibbs sampling is given by

$$P(\mathbf{Z}, \mathbf{Y}|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{Y}, \mathbf{Z})P(\mathbf{Y})P(\mathbf{Z}). \tag{7}$$

For the linear-Gaussian model, remember that $P(\mathbf{X}|\mathbf{Y},\mathbf{Z})$ is a Gaussian distribution with mean $\mathbf{ZY}$ and variance $\sigma_Y^2\mathbf{I}_D$, where $\mathbf{I}_D$ is the $D \times D$ identity matrix. Though we could perform Gibbs sampling in the same fashion as the noisy-OR case, $\mathbf{Y}$ can be integrated out and so, we sample $\mathbf{Z}$ directly from $P(\mathbf{Z}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{Z})P(\mathbf{Z})$ where $p(\mathbf{X}|\mathbf{Z})$ can be shown to be (Griffiths & Ghahramani, 2006)

$$p(\mathbf{X}|\mathbf{Z}) = \exp\left\{-\frac{1}{2\sigma_X^2}\text{tr}\left(\mathbf{X}^T(\mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_Y^2}\mathbf{I})^{-1}\mathbf{Z}^T)\mathbf{X}\right)\right\} \tag{8}$$

where σ_X^2 is the expected variance of the observed objects, $\mathbf{B}^T$ denotes the transpose of matrix $\mathbf{B}$, and $\text{tr}(\mathbf{B})$ denotes the trace of the matrix $\mathbf{B}$ (the sum of its elements on the primary diagonal).[13] If $\mathbf{Y}$ is needed explicitly (which it will be to predict the new objects), we calculate its posterior mean given the observed objects and $\mathbf{Z}$ (Griffiths & Ghahramani, 2006):

$$E[\mathbf{Y}|\mathbf{X}, \mathbf{Z}] = \left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_Y^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{X}. \tag{9}$$

For each experiment, we ran the sampler for at least 500 samples and took the inferred $\widehat{\mathbf{Z}}$ (and $\widehat{\mathbf{Y}}$) to be the mode of the distribution given by the Gibbs sampler after a burn-in of 50 iterations (removing the

[13] Although the IBP model with the noisy-OR likelihood has four parameters and the IBP model with the linear-Gaussian likelihood has three parameters, these are not free parameters because their values are set without looking at the test objects or human responses. Usually they are set using the gross statistics of the training set, though we obtain comparable results if we put a prior distribution over these parameters and sample them as well.

first 50 iterations from analysis, though the results do not depend on the length of the burn-in). Next, we used the inferred values to approximate $P(\mathbf{x}_{\text{test}}|\mathbf{X})$, the probability of observing new objects $\mathbf{x}_{\text{test}}$ after observing the object set $\mathbf{X}$:

$$P(\mathbf{x}_{\text{test}}|\mathbf{X}) = \sum_{\mathbf{Z}_{\text{test}},\mathbf{Z},\mathbf{Y}} P(\mathbf{x}_{\text{test}}|\mathbf{Z}_{\text{test}},\mathbf{Y})P(\mathbf{Z},\mathbf{Y}|\mathbf{X}) \tag{10}$$

$$\approx P(\mathbf{x}_{\text{test}}|\widehat{\mathbf{Z}}_{\text{test}},\widehat{\mathbf{Y}}) \tag{11}$$

where we pick the representation of the new object(s) $\widehat{\mathbf{Z}}_{\text{test}}$ to be the one that maximizes the probability of the test objects ($\widehat{\mathbf{Z}}_{\text{test}} = \arg\max_{\mathbf{Z}} P(\mathbf{x}_{\text{test}} \mid \mathbf{Z},\widehat{\mathbf{Y}})$). For our experiments, the number of inferred features is small and so we find $\widehat{\mathbf{Z}}_{\text{test}}$ by explicitly searching through all possibilities.

Using the steps described so far, we now have an unnormalized approximate probability of each test object under the rational model (or each type of test object by summing over the objects of the same type). To compare this to the data from our experiments (human generalization likelihoods), we need to rescale the unnormalized probabilities. To do this, we transform the unnormalized probabilities through the exponentiated Luce choice rule $P(\mathbf{x}_{\text{test}}|\widehat{\mathbf{Z}}_{\text{test}},\widehat{\mathbf{Y}})^{\gamma}$, where γ is a parameter fit by minimizing the squared distance between the model's responses and the human data, and then renormalize over each type of object.[14] This is the only parameter that is used to fit the model to the human results.

For the simulations with binary data (the Shiffrin and Lightfoot modeling, and Experiments 1, 2, and 4), the parameters, α, ϵ, λ, and ϕ of the Gibbs sampler were set to 2, 0.01, 0.99, and 0.1 respectively and chosen without looking at the test objects. Changing α had little effect on the features that were inferred by the model (though there are ways of inferring α from the data using the Metropolis-Hastings algorithm as in Wood et al., 2006). Lowering λ resulted in poor performance due to the model not believing in the efficacy of the features it was trying to infer (roughly, λ is the probability that a feature fails to turn on a pixel it says to turn on). Varying ϵ affects how much noise the model suspects there is in the data set. As ϵ increases, the model thinks everything is noise and as ϵ decreases, the model does not think there is any noise and it infers a large number of features to capture every aspect of each object. Changing ϕ effects the number of pixels that are on in each feature. Small values of ϕ encourage features that are small and large values of ϕ encourage features that are large (though the model can infer small and large features when it is appropriate for the observed set of data). For the simulations with grayscale data (Experiment 3), the parameters, α, σ_Y^2, and σ_X^2, of the Gibbs sampler were set to 2, 15, and 20. σ_X^2 is similar to ϵ in that as it increases, the model thinks everything is noise and as it decreases the model does not think there is any noise. σ_Y^2 has a similar effect to ϕ except now low values of σ_Y^2 penalize extreme values in the feature images.

Due to the high dimensionality of the images used in the Shiffrin and Lightfoot and Experiments 1 and 2 simulations (each image in Experiments 1 and 2 is $86 \times 146 = 12{,}556$-dimensional), the Gibbs sampler gets stuck in local minima and thus, simulated annealing (Geman & Geman, 1984) and "split-merge" steps as discussed in detail later were used to help the Gibbs sampler find the feature representation with highest posterior probability (a standard practice for solving hard global optimization problems). Given a temperature parameter t, the Gibbs sampler for $P(\mathbf{Z}|\mathbf{X},\mathbf{Y})$ using simulated annealing[15] samples from $P(\mathbf{Z}|\mathbf{X},\mathbf{Y})^{1/t}$. As $t \to \infty$, the Gibbs sampler with simulated annealing chooses values uniformly at random (regardless of $\mathbf{X}$ and $\mathbf{Y}$). As $t \to 1$, the sampler draws values from $P(\mathbf{Z}|\mathbf{X},\mathbf{Y})$ and as $t \to 0$, the sampler converges to the global maximum of $P(\mathbf{Z}|\mathbf{X},\mathbf{Y})$. Thus, to help the Gibbs sampler search get out of local minima, we start with small t (so the sampler chooses new values almost uniformly at random) and slowly increase t we iterate using the Gibbs sampler. The temperature t used by the Gibbs sampler on iteration i was $\frac{250}{\log(i+1)}$. In the simulations for Experiment 1, we found that the Gibbs sampler converged to the reported feature representation by approximately iteration 400.

[14] Due to the high dimensionality of the input data for Experiments 1, 2, and 3, the model predictions were expressed computationally as log probabilities. We transformed the log probabilities by $\exp\{\gamma|\log P(\mathbf{x}_{\text{test}}|\widehat{\mathbf{Z}}_{\text{test}},\widehat{\mathbf{Y}})|\}$. This results in the same transformation on the probabilities defined previously.

[15] In our simulations, Gibbs sampling using simulated annealing to sample from $P(\mathbf{Y}|\mathbf{Z},\mathbf{X})$ resulted in worse performance; however, it was necessary to use Gibbs sampling with simulated annealing to sample effectively from $P(\mathbf{Z}|\mathbf{X},\mathbf{Y})$.

For Experiments 1 and 2, the Gibbs sampler was run for 1000 iterations and the best-fitting γ was 1.5677×10^3. For Experiment 3, the Gibbs sampler was run for 500 iterations and the best-fitting γ was 3.35×10^{-5}. To aid inference, the images were cropped and brightened to increase the contrast of each image. For Experiment 4, the Gibbs sampler was run for 2000 iterations and the best-fitting γ was *1.65*. Importantly, the parameters were the same given either set of observed objects (and for different experiments using the same likelihood model) and aside from γ, chosen without looking at the test objects. Thus, the model's difference in generalization when given the *correlated* set versus the *independent* set is due to the statistical properties of the given object set.

In the General Discussion, we explore imposing a proximity bias on the feature image prior for modeling Shiffrin and Lightfoot (1997) and this filled in the speckled holes that were in inferred when we used the simple feature image prior. To use the proximity bias, every time $P(\mathbf{Y})$ occurs in the Gibbs sampler, we used the feature image prior given by Eq. (6) instead of independent Bernoulli coinflips. Though this does help with inferring more psychologically plausible features, it adds extra complexity to the probability distribution, which makes inference even more difficult. To deal with this issue, we performed a stochastic search, where we repeatedly perform four Gibbs sampling steps and then a "split-merge" step. A split-merge step creates a proposal new state that either splits a feature into two new features or merges two features into one feature with equal probability. The probability the proposal is chosen as the next state is proportional to the ratio of the probability of the objects under the current feature set versus the proposal feature set multiplied by the temperature t (so this step underwent the same simulated annealing as discussed above). The split-merge steps are necessary because the Gibbs sampler for the model with the proximity bias feature image prior rarely changes dimensionality. The results reported in the article use $\theta = 0.15$ and $\rho = 0.999$ (the other parameters were the same).

A.2. Modeling human responses using a prototype and exemplar model

To model the results of Experiments 1 and 2 with a prototype and exemplar model, we formulated a simple prototype and exemplar model based on Nosofsky (1986). To define each model, we defined a distance metric between two objects ($d(\mathbf{x}_{\text{test}},\mathbf{x})$) and a similarity function of a test object to the set of observed objects using that distance metric ($\eta_{\mathbf{x}_{\text{test}},\mathbf{X}}$), and then computed the generalization likelihood response function using the exponentiated Luce choice rule. The distance metric we used was

$$d(\mathbf{x}_{\text{test}},\mathbf{x}) = \sqrt{(\mathbf{x}_{\text{test}}-\mathbf{x})^T(\mathbf{x}_{\text{test}}-\mathbf{x})}, \tag{12}$$

where $\mathbf{x}$ is an observed object to have the property of interest, $\mathbf{x}_{\text{test}}$ is another object, and $(\mathbf{x}_{\text{test}}-\mathbf{x})^T(\mathbf{x}_{\text{test}}-\mathbf{x})$ is the dot product of the vector $\mathbf{x}_{\text{test}}-\mathbf{x}$ with itself.

The similarity function for both models was based on exponential decay (Nosofsky, 1986; Shepard, 1987), where the prototype model calculates the distance of the test object to the mean of the observed objects ($\mathbf{x}_{\text{mean}}$) and the exemplar model sums over the distances of the test object to each of the given objects. Thus the similarity function $\left(\eta^p_{\mathbf{x}_{\text{test}},\mathbf{X}}\right)$ for the prototype model was

$$\eta^p_{\mathbf{x}_{\text{test}},\mathbf{X}} = \exp\{-\kappa d(\mathbf{x}_{\text{test}},\mathbf{x}_{\text{mean}})\} \tag{13}$$

and the similarity function for the exemplar model $\left(\eta^e_{\mathbf{x}_{\text{test}},\mathbf{X}}\right)$ was

$$\eta^e_{\mathbf{x}_{\text{test}},\mathbf{X}} = \sum_{\mathbf{x}\in\mathbf{X}} \exp\{-\kappa d(\mathbf{x}_{\text{test}},\mathbf{x})\} \tag{14}$$

where κ is a "specificity" parameter, scaling the distances in both models. Then, we connected the unnormalized generalization probabilities to the human responses using the same exponentiated Luce choice rule as was used in the previous section for the IBP. For Experiments 1 and 2, we found the best-fitting γ and κ on averaged responses for each model separately ($\gamma = 1.5 \times 10^{-3}$, $\kappa = 6.2573 \times 10^{-4}$ for the exemplar model and $\gamma = 0.38$, $\kappa = 0.3705$ for the prototype model).

A.3. Modeling human responses using principal component analysis and independent component analysis

To model the results of Experiments 1 and 2 using principal component analysis (PCA) and independent component analysis (ICA), we used the first K non-noise dimensions learned from the training set as our representation space for the objects. For PCA, there were three non-noise dimensions for the *correlated* set and five non-noise dimensions for the *independent* set. For ICA, there were four non-noise dimensions for the *correlated* set and six for the *independent* set. To calculate the likelihood of the test objects after observing each image set, we projected the test objects into the subspace learned to represent the observed image set. The average reconstruction error of the test objects projected into the learned subspace to the original test object (mean squared error) was taken as a measure of how likely the test object is after observing the previous objects. For PCA, this is monotonically related to the predictive probability of the test objects given the observed objects.[16] Then, we convert the average reconstruction error to the results from Experiments 1 and 2 using the same transformation as was used in the previous section for the IBP. For Experiments 1 and 2, we found that the best-fitting parameter value was $\gamma = 0.3082$ for PCA and $\gamma = 0.0726$ for ICA.

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.

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

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321–324.

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

Austerweil, J. L., & Griffiths, T. L. (2010). Learning invariant features using the transformed indian buffet process. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., & Culotta A. (Eds.), *Advances in neural information processing systems* (Vol. 23, pp. 82–90).

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition, 83*, 167–206.

Bell, A., & Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*(6), 1129–1159.

Biederman, I., & Cooper, E. E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology, 23*, 393–419.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

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

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning, 34*, 71–105.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *Quarterly Journal of Experimental Psychology, 52A*, 273–302.

Chater, N., & Vitanyi, P. (2003). Simplicity: A unifying principle in cognitive science. *Trends in Cognitive Science, 7*, 19–22.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367–405.

Doshi-Velez, F., Miller, K. T., Van Gael, J., & Teh, Y. W. (2009). Variational inference for the Indian buffet process. In *Proceedings of the 12th international conference on artificial intelligence and statistics, Journal of Machine Learning Research* (pp. 137–144). Clearwater Beach, FL.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, 429–433.

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

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

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721–741.

Goldmeier, E. (1936/1972). Similarity in visually perceived forms. *Psychological Issues, 8*, 1–136.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition, 52*(2), 125–157.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49*, 585–612.

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

Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In *Perceptual organization in vision: Behavioral and neural perspectives* (pp. 233–278). Mahwah, NJ: Lawerence Erlbaum Associates.

[16] For ICA, it is less clear whether or not the average reconstruction error is monotonically related to the predictive probability of the test objects. See Hyvarinen et al. (2001) for details.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112*, 21–54.
Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects* (pp. 35–41). New York: The Bobbs-Merrill Co.
Green, P., & Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics, 28*, 355–377.
Griffiths, T. L., & Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.). *Advances in neural information processing systems* (Vol. 18). Cambridge, MA: MIT Press.
Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling* (pp. 59–100). Cambridge: Cambridge University Press.
Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 354–384.
Hall, G. (1991). *Perceptual and associative learning*. Oxford: Clarendon.
Hansen, L. K., Ahrendt, P., & Larsen, J. (2005). Towards cognitive component analysis. In *International and interdisciplinary conference on adaptive knowledge representation and reasoning* (pp. 148–153). Pattern Recognition Society of Finland, Finnish Artificial Intelligence Society, Finnish Cognitive Linguistics Society.
Hochberg, J., & McAlister, E. (1953). A quantitative approach to figural "goodness". *Journal of Experimental Psychology, 46*(5), 361–364.
Hoffman, D. D., & Richards, W. A. (1985). Parts in recognition. *Cognition, 18*, 65–96.
Hyvarinen, A. (1999). Fast and robust fixed-point algorithm for independent component analysis. *IEEE Transactions on Neural Networks, 10*(3), 626–634.
Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: Wiley.
Jordan, M. (1986). An introduction to linear algebra in parallel distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Foundations* (Vol. 1, pp. 365–421). Cambridge, MA: MIT Press.
Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence of a domain general learning mechanism. *Cognition, 83*, B35–B42.
Koerding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427*, 244–248.
Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.
Leeuwenberg, E. (1978). Quantification of certain visual pattern properties: Salience, transparency, similarity. In *Formal theories of visual perception* (pp. 277–297). New York: Wiley.
Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In Koller, D., Schuurmans, D., Bengio, Y., & Bottou, L. (Eds.), *Advances in neural information processing systems* (Vol. 21, pp. 937–944).
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*(4), 1153–1169.
Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55*, 232–257.
Luce, R. D. (1959). *Individual choice behavior*. New York: John Wiley.
Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.
Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*(1), 37–50.
Medin, D. L., Goldstone, R., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*, 254–278.
Miller, K. T., Griffiths, T. L., & Jordan, M. I. (2008). The phylogenetic Indian buffet process: A non-exchangeable nonparameteric prior for latent features. In D. McAllester & P. Myllymaki (Eds.), *Proceedings of the twenty-fourth conference on uncertainity in artificial intelligence* (pp. 403–410). Corvallis, Oregon: AUAI Press.
Minka, T. (2001). *Automatic choice of dimensionality for PCA. Advances in neural information processing systems* (Vol. 13). Cambridge, MA: MIT Press, pp. 598–604.
Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.
Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition*. Oxford: Oxford University Press.
Olshausen, B., & Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*, 607–609.
Orban, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences, 105*(7), 2745–2750.
Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology, 9*, 441–474.
Palmer, S. E. (1991). Goodness, Gestalt, groups, and Garner: Local symmetry subgroups as a theory of figural goodness. In G. R. Lockhead & J. R. Pomerantz (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 23–39). Washington, DC: American Psychological Association.
Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, Massachusetts: MIT Press.
Palmer, S. E. (2003). Perceptual organization and grouping. In *Perceptual organization in vision: Behavioral and neural perspectives* (pp. 3–43). Mahwah, NJ: Lawerence Erlbaum Associates.
Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39*, 246–263.
Pevtzow, R., & Goldstone, R. L. (1994). Categorization and the parsing of objects. In *Proceedings of the sixteenth annual conference of the cognitive science society* (pp. 712–722). Hillsdale, NJ: Lawrence Erlbaum Associates.
Rasmussen, C. E., & Ghahramani, Z. (2001). Occam's razor. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.). *Advances in neural information processing systems* (Vol. 13, pp. 294–300). Cambridge, MA: MIT Press.
Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 393–407.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science, 274*, 1926–1928.
Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 726–731). Mahwah, NJ: Erlbaum.

Schyns, P. G., & Murphy, G. L. (1994). The ontogeny of part representation in object concepts. *The psychology of learning and motivation* (Vol. 31, pp. 05–354). San Diego: Academic Press.
Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 681–696.
Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science, 237*, 1317–1323.
Shiffrin, R. M., & Lightfoot, N. (1997). Perceptual learning of alphanumeric-like characters. *The psychology of learning and motivation* (Vol. 36, pp. 45–82). San Diego: Academic Press.
Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences, 102*(33), 11629–11634.
Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14*, 29–56.
Spratling, M. W. (2006). Learning image components for object recognition. *Journal of Machine Learning Research, 7*, 793–815.
Thibaux, R., & Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In Meila M. & Shen X. (Eds.), *Eleventh international conference on artificial intelligence and statistics (AISTATS 2007)* (pp. 564–571). San Juan, Puerto Rico: Omnipress.
Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.
Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences, 11*(2), 58–64.
van der Helm, P. A., & Leeuwenberg, E. L. J. (1996). Goodness of visual regularities: A nontransformational approach. *Psychological Review, 103*(3), 429–456.
Van Gael, J., Teh, Y. W., & Ghahramani, Z. (2009). The infinite factorial hidden Markov model. *Advances in neural information systems* (Vol. 21, pp. 1697–1704). Cambridge, MA: MIT Press.
Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics, 27*, 352–372.
Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience, 5*, 598–604.
Wertheimer, M. (1938). Laws of organization in perceptual forms. In *A source book of Gestalt psychology* (pp. 71–88). London: Routledge and Kegan Paul.
Wood, F., & Griffiths, T. L. (2007). Particle filtering for nonparametric Bayesian matrix factorization. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.). *Advances in neural information processing systems* (Vol. 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 *Proceeding of the 22nd conference on uncertainty in artificial intelligence* (pp. 536–543). Cambridge, MA: AUAI Press.
Zhu, S.-C. (1999). Embedding Gestalt laws in Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(11), 1170–1187.