Title
Constructing flexible feature representations using nonparametric Bayesian inference

Author
Austerweil, Joseph Larry

Publication Date
2012

Peer reviewed|Thesis/dissertation

Constructing flexible feature representations using nonparametric Bayesian inference

by

Joseph Larry Austerweil

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Thomas Griffiths, Chair
Professor Michael Jordan
Assistant Professor Tania Lombrozo
Professor Stephen Palmer

Fall 2012

Constructing flexible feature representations using nonparametric Bayesian inference

Abstract

Constructing flexible feature representations using
nonparametric Bayesian inference

by

Joseph Larry Austerweil

Doctor of Philosophy in Psychology

University of California, Berkeley

Associate Professor Thomas Griffiths, Chair

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

To my mother, who taught me what is important.

Contents

Acknowledgments

Although it usually takes a village to raise someone, in my case, it took a moderately-sized town to raise me into the person who just completed his doctoral dissertation. I have been extremely fortunate to have the support of a large number of people throughout my life and am glad that I now have the opportunity to acknowledge their support. Firstly, I would like to thank my parents, Theresa and Arthur Austerweil, for their unending support and love. I am especially thankful that considerable time and effort that my father spent teaching me algebra and computer programming when I was a young boy. I am grateful for our time playing with the LOGO turtle and programming Conway's Game of Life in Visual Basic together. I thank my brother, David Austerweil, and his wife, Rovika Rajkishun, for inspiring me to work tirelessly on a career that I feel passionate about and their emotional support. Also, I am very thankful that they brought my nephew Noah into the world.

In addition to the family I was born into, I have been fortunate to be part of a number of families throughout my life. I am extremely grateful to the Jailer-Coleys (Kathy, Jim, Joseph, and Eleanor). In addition to culturing me in good (and bad) science fiction and music, they have been and continue to be incredibly supportive and lots of fun to spend time with (in particular, Joe). Additionally, I thank my friends from growing up: Andrew Pariser, Sam Marcellus, Min Suh, Adam Bloomston, Thomas Wang, Gina Farinaccio, Nikki Ambrosio, and the others I unintentionally left out. I was very lucky to have the guidance of Elaine Labrocca and Scott Lenz on my first few research projects. I am not sure I would have ended up becoming a scientist otherwise. When I arrived at my freshman dormitory in college, I had no idea that the people I lived with in that first dormitory would end up being my family throughout college. I am especially thankful for Erikson Arcaria, Kate Johnston, Rahim Kassam-Adams, David McNamee, Tyler Rorrison, Tucker Peck, Stuart Schüssel, Stephanie Minor, Maggie Mustard, and Elizabeth Dickinson McNamee. Additionally, I am grateful for the incredible graduate students and professors who taught me and inspired me to do research at Brown, including Thomas Griffiths, Stuart Geman, Steven Sloman, Micha Elsner, Frank Wood, Philip Fernbach, Naomi Feldman, and Eugene Charniak.

While at Brown, a new professor, Tom Griffiths, came to give a lecture on machine learning in my introduction to artificial intelligence course. During that lecture, it became clear to me that I wanted to spend my life working on the types of approaches and problems he had been advocating. Despite forgetting my final presentation for his class, Tom invited me to be a research assistant in his laboratory, and he has been my advisor since then. He is the most simultaneously brilliant and humble person that I have ever met, and I hope that I have learned even a tenth of his ingenuity at connecting modern mathematical techniques with psychological issues. Furthermore, he has been an incredibly supportive and compassionate advisor throughout the difficulties that I have had throughout the last eight years. I am incredibly fortunate that he was my advisor.

In addition to being grateful for his academic and emotional support, I am very grateful that Tom brought me to Berkeley because it is where I met the most important and wonderful

person in my life, Karen Schloss, who has been an endless supply of support and love. Furthermore, she is a brilliant scientist, who continues to teach me new things and inspire me to become a better scientist. Her family (Nina Schloss, Lou Schloss, Lori Brickell, Jill DeArmon, and Jaden DeArmon), has been incredibly wonderful to me and I am especially fortunate to have joined their family and share in their joy.

I am thankful for the friendship of a large number of graduate students and postdoctoral scholars over the last six years I have spent in the Psychology Department at Berkeley: Josh Abbott, Michael Pacer, Chris Lucas, Jing Xu, Lei Shi, Florencia Reali, Vincent Berthiaume, Elizabeth Bonawitz, Luke Maurits, Kevin Canini, Taraz Lee, Peter Butcher, Dav Clark, Daphna Buchsbaum, Jessica Hamrick, Michael Pacer, Anna Rafferty, Joseph Jay Williams, Jay Martin, and Saiwing Yeung. I thank all of the research assistants that I have had throughout graduate school: David Belford, Dylan Breen, Matt Cammann, Song Choi, Shubin Li, Ingrid Liu, Benj Shapiro, Hye Young Shin, Brian Tang, Christina Vu, and Julia Ying. I am sure that most of them will have very successful and interesting lives.

I want to thank Stephen Palmer, Bill Prinzmetal, and Karen De Valois have become our Berkeley family. I've enjoyed the countless dinners and nights that we have spent together. I am especially thankful for Steve also being on a committee member and providing substantial feedback on this dissertation and on my research. I am extremely grateful for all of the academic and non-academic advice that Tania Lombrozo has given me. In addition to helping me think through some of the most difficult philosophical issues related to my research, she is incredibly wise and compassionate. I thank Michael Jordan for teaching me statistics and machine learning from both intuitive and rigorous points of view. Additionally, I appreciate his thoughts about my research interests and being a committee member for my qualifying examination, statistics masters, and this dissertation.

Also, I am grateful that all the people I left off, but should have included in the acknowledgments, forgive me for not including them. Last, but not least, I would like to thank my cat, Banana Kaleidoscope Dishes Schlossterweil, who has allowed me to stare at my glowing rectangle for long enough to complete this dissertation.

Chapter 1

Introduction

Modern cognitive psychology explains human behavior as the result of processes acting on sensory inputs from the environment. Figure 1a illustrates the computational formulation of this viewpoint: Behavior is a function from inputs to outputs, and the goal of psychology is to understand and to describe this function [Marr, 1982, Palmer and Kimchi, 1986]. Through the history of experimental psychology, there have been a number of different proposed theoretical frameworks for how to go about explaining this function. According to the Behaviorist tradition, behavior is explained as a series of connections from our sensory inputs to motor outputs and behavior. Shown as a caricature in Figure 1b, the mind was viewed as a comprehensive list of input-output connections. The goal is to find which changes of inputs lead to changes in outputs [Watson, 1913]. From this viewpoint, learning is a process that strengthens input-output pairs that co-occur and weakens those that do not co-occur [Hebb, 1958]. Then, when two inputs (e.g., the auditory sensation from a bell and the olfactory sensation of food) that used to lead to different behaviors (e.g., focusing attention or not much at all, and salivating) are associated (e.g., sounding a bell whenever a dog gets food), they produce the same behavior (e.g., salivating). For Radical Behaviorists, the explanatory emphasis for even the most seemingly complex process (e.g., language learning or decision making) is on the observable factors in the environment, and not on whatever internal events may be occurring in the mind of the agent [Skinner, 1977].

In reaction to the Behaviorists, cognitive psychologists argue that how people react to a stimulus is determined by their representation of the stimulus and not by the stimulus itself [Chomsky, 1957, Pitt, 2008]. From this viewpoint, behavior can be decomposed into (at least) two processes: (1) a mapping the sensory input to some of internal state(s), and (2) a mapping the internal state(s) to behaviors (as shown in Figure 1c). Each mapping can be understood as a machine that takes the inputs as symbols, and processes those symbols using simple rules of computation to produce an output [Gardner, 1987, Palmer and Kimchi, 1986]. One useful property of this explanatory viewpoint is that it can potentially explain why two people who have the same input can react very differently [Chomsky, 1959, Neisser, 1967]. For example, imagine that an art theorist and lay person both view a Jackson Pollock

painting. An art historian may enjoy a Jackson Pollock painting due to her representation of it as a rejection of painting with a brush and exclaim, "That is beautiful!" A lay person viewing the same painting might dislike it because he represents it as a cluttered mess of discordant colors and exclaim, "That is ugly." Representations are the central device for explaining such different reactions to the same stimulus. Thus, understanding how the mind represents stimuli is a fundamental problem for understanding behavior within a cognitive framework.

Before delving further into how the mind represents stimuli, it is important to be clear about what we mean by a representation. Unfortunately, representation is a notoriously difficult concept to define [Palmer, 1978, Cummins, 1989, Markman, 1998]. Perhaps a good starting definition of a representation, which can be traced back at least to Aristotle [Pitt, 2008], "is something that stands in place for something else" [Palmer, 1978, p. 262]. Although this is vague, it gives the gist of what a representation is: something (e.g., a symbol or the activation of artificial neurons in a layer of a neural network) that stands for something else (e.g, an object in the environment or a symbol in a different cognitive process). The internal representation being active indicates the presence of what it represents (i.e., whether or not what it represents is present).

Although using representations to explain human behavior has been (and continues to be) a successful scientific paradigm, it has been difficult to specify a method for computing which representation should be used to encode the sensory input from a given stimulus. Indeed, many cognitive explanations ignore the problem by implicitly assuming that representations are given directly from the inputs of stimuli on our senses [Schyns et al., 1998]. However, it is clear that most information useful for decision making is not directly available from sensory inputs.

There are many reasons that specifying how representations are computed from sensory input remains a difficult problem. For example, the appropriate representation to encode a stimulus depends on the configuration of values over a large number of receptors (an observation that dates back to at least the Gestaltists; Wertheimer, 1923/1938). This dissertation focuses on one particular reason that the problem of representation is difficult: the representation for any sensory input from a stimulus is an ill-posed or inductive problem in the sense that there are an infinite number of representations logically consistent with any sensory input. Indeed, a large amount of effort in both perception and higher-level cognition has been devoted to describing the rules used by the mind (sometimes called heuristics) to choose among the many possible solutions to the ill-posed problem [Kahneman et al., 1982, Palmer, 1999]. For example, there are (at least) two plausible representations consistent with the object that is labelled a "dax" in Figure 1d. The two representations yield different behaviors (the new object is a dax only when it is represented as two separate vertical bars). We will return to explaining this example in Chapter 5. Thus, it is critical to understand the principles that the mind uses to decide among different representations.

Our investigation focuses on one hypothesis: the mind represents its input using the

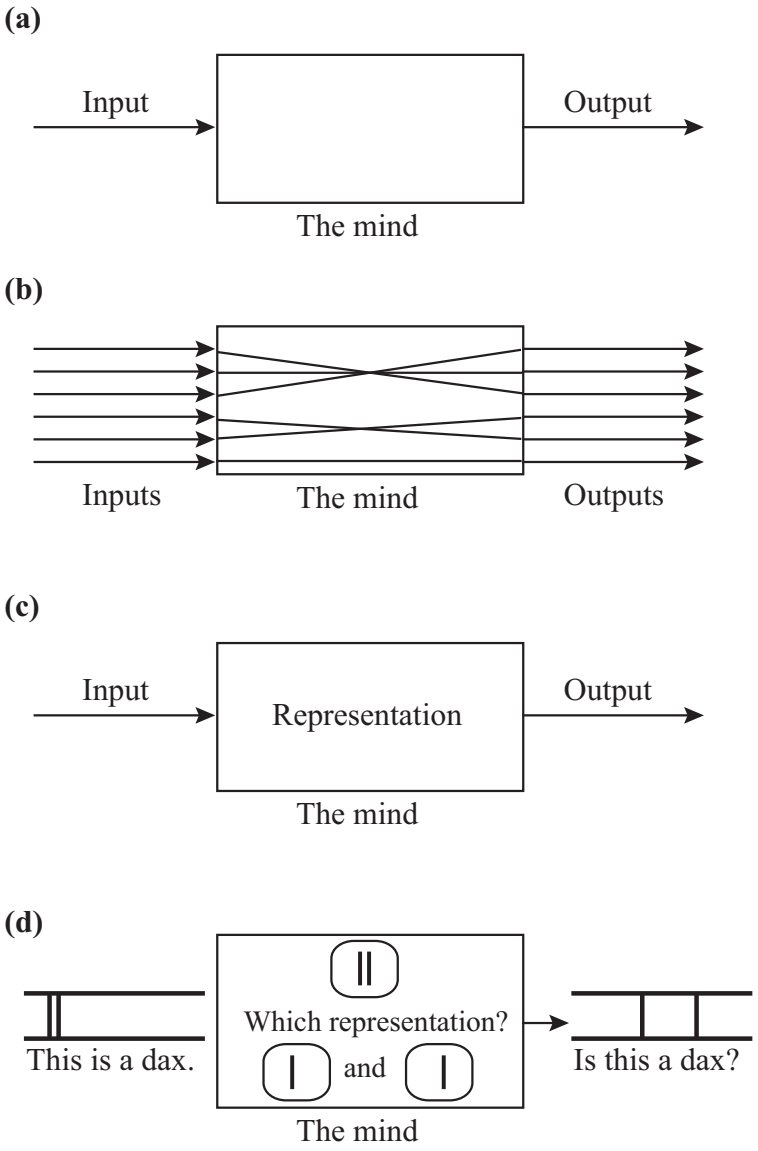


Figure 1.1: Computational explanations of the mind. (a) In a computational explanation of the mind, the mind is viewed as function mapping inputs to outputs. Theories differ in their assumptions about the nature of that mapping and how that mapping changes with experience. (b) According to Radical Behaviorism, simple and complex behaviors are best thought of as a big list of input-output pairs, which are malleable from learning. (c) According to Cognitive Psychologists, sensory inputs are mapped to representations, which in turn, determine behavior. (d) Which representation should be used to encode an object is ambiguous (it is not clear which representation should be used). The mind will have different behaviors depending on how the object is represented, and thus, knowing the principles that the mind uses to infer the appropriate representation of objects in different contexts is an essential part of explaining behavior. Chapter 5 explores how context can determine the appropriate representation for this particular example.

simplest representation (e.g., the one with the fewest features) that is consistent with prior knowledge (e.g., consistent with perceptual expectations or diagnostic for categorization), and which encodes the input and its context (e.g., the set of other sensory inputs that it is presented with). We will formalize and explore what we mean by simplicity, prior knowledge, and context in more detail throughout the dissertation. Variants of this hypothesis have a long history in psychology [Attneave, 1954, Garner, 1974, Helmholtz, 1866, Hochberg and McAlister, 1953, Leeuwenberg, 1971, Selfridge, 1955]. Selfridge (1955) provides a classic example of the effect of context on the representation for some input. When an image (halfway between "A" and "H") is presented with "C" and "T," it is represented as "A" (to form "CAT"), but when it is presented with "T" and "E" it is represented with "H" (to spell "THE"). In this dissertation, we present one of the first comprehensive computational frameworks for explaining and exploring how context and prior knowledge affects the representations inferred for a stimulus.

1.1 Computational explanations in cognitive psychology

The computational approach to the mind, which characterizes how the structure of input stimuli affects human behavior, date back at least to the first studies in experimental psychology.[1] Historically, psychology is one of the last sciences to branch off from philosophy (after physics, chemistry, and biology). Progress in many sciences proceeded by positing variables of interest for a domain, and showing that changes in those variables affect other variables in a domain. In other sciences, this perspective was successful when the variables in the computational description lead to the creation of hypotheses regarding the latent variables affecting the phenomenon of interest, and subsequent rejection or verification of the existence of the latent variables through experimenting [Einstein and Infeld, 1942]. In fact, many theoretical movements within psychology, including Behaviorism and Information Processing [Watson, 1913, Cantril et al., 1949, Marr, 1982], are all interested in describing the functional relationship of how the inputs to our senses are transformed into outputs – the computational perspective.

Arguably, the first experimental results in psychology explored perception and memory using the computational approach to the mind. The first psychological experiments were completed in the mid 1800s by Weber, Fechner, Helmholtz, and Wundt, who investigated perception using psychophysics, and by Ebbinghaus, who investigated memory by statistical analysis of patterns of recollection error [Misiak and Sexton, 1966]. The psychophysicists examined how changes in the structure of inputs lead to different behaviors (usually, agnostic

[1] The computational approach to the mind is not exactly the same as a computational-level explanation. A computational-level explanation is a type of computational approach to the mind, which characterizes the formal solution to a mathematical problem.

to the internal processing). For example, Weber and Fechner both investigated how large of the change in energy of two inputs was necessary for a subject to be able to distinguish between the inputs (i.e., just noticeable differences) in different domains [Boring, 1961]. On the other hand, Ebbinghaus explored how different variations in a list of nonsense words (e.g., list length, word repetition) affected the time it took him to recall the list perfectly [Ebbinghaus, 1913]. In effect, this is possibly the first rigorous exploration of how a mental process within the mind changes over time (memory), as he discovered basic principles for how learning a novel input-output mapping (recalling pairs of words in a list) as a function of properties of the word list. In essence, this is the computational approach in cognitive psychology: describing the structure of input-output pairs for a particular type of behavior and how the input-output mapping may change with experience.

Unfortunately, much of human behavior is too complex to describe as a mere function of the current input [Chomsky, 1957]. As illustrated by the differing responses of the art theorist and lay person looking at the same Jackson Pollock painting, people can respond to the same input differently depending on their representation of that input. This suggests we should study representations to understand human behavior. However, we do not have direct access to representations. If we cannot directly observe a representation, how can we scientifically investigate representations without directly accessing them? In fact, Anderson (1978) proved that without extra constraints (e.g., specifying the processes that use the representations), it is impossible to distinguish between two types of representation from behavioral data. Anderson's argument can be summarized as follows: If two representations preserve the same distinctions to external stimuli, then a function exists mapping one representation to the other representation (and vice versa). After applying the mapping function, the process of the second representation can be applied to the first (as it now is in the format of the second representation) and the first representation will act identically in all behavioral tasks to the second representation. Thus, we cannot distinguish between two representations without specifying processes or other constraints that act on each representation. We return to discussing Anderson's argument in relation to our central thesis in Chapter 8.

One method for studying representations that has been pursued by computational cognitive psychologists is to specify a mathematical rule for "learning" representations based on sensory inputs and how those representations map to outputs (though in some "autoassociator" or unsupervised models, the input and output are the same). This ameliorates the previous mimicry concerns based on Anderson's argument because it assumes a particular process for how to create representations, and how the inputs map to the created representations, which, in turn, map to outputs. One of the most exciting developments in the last century of computational cognitive psychology research has been the development of sophisticated techniques for learning representations in neural networks given a set of input-output pairs [Rumelhart and McClelland, 1986]. Historically, the first computational rules for learning representations were Hebbian learning [Hebb, 1949] and the Perceptron [Rosenblatt, 1958], which learn representations distributed over artificial neurons based on the correlations between sensory inputs with itself (and in the latter case, outputs as well).

These learning rules specify a formulation of the input-output mapping (a linear function) and how it should be updated. They lead to particular predictions, including the most famous theoretical prediction, which is that non-linear functions should be difficult or impossible for people to learn [Minsy and Papert, 1969]. As many important psychological tasks require a non-linear mapping from inputs to outputs (e.g., deciding if a figure is connected, given small patches of it as input; Minsky & Papert, 1969), this ruled out the possibility of the function(s) mapping inputs to representations to outputs being linear. In fact, linear mappings are given by matrix multiplication and so, multiple mappings can always be represented in terms of a single mapping (as the product of multiple matrices can always be reduced to the product of two matrices). This means that the intermediary representations are completely superfluous, and thus, weakens their explanatory status. Further contradictory evidence comes from previous work in categorization, which has shown that linear category rules can be harder to learn than non-linear category rules [Medin and Schwanenflugel, 1981].

When Minsky and Papert (1969) pointed out that linear neural networks could not learn a non-linear input-output mapping, it resulted in a general abandonment of neural network models due to the number of non-linear problems that people solve effortlessly [Rumelhart and Zipser, 1985]. This ruled out linear neural networks as a psychologically plausible computational model. However, Minsky and Papert (1969) pointed out that non-linear neural networks could solve these functions, but no rules for learning the representations in non-linear neural networks were known at the time. This led to the use of non-linear neural networks as computational explanations of psychological behavior whose representations were pre-determined by the experimenter [McClelland and Rumelhart, 1981] and the development of mathematical tools for learning representations in non-linear neural networks [Rumelhart et al., 1986]. Although tools for learning representations in non-linear and hierarchical neural networks have been developed [Rumelhart et al., 1986, Hinton and Salakhutdinov, 2006], it is unclear whether they form human-like representations when given the same information as people in a task. However, there is no reason to suppose that they would not infer human-like representations, and so, their application to the problem of representation learning is an interesting open question.

Another important issue in trying to understand behavioral phenomena is how multiple explanations of the same behavioral phenomenon could all be explanatory and non-competing. Marr (1982) proposed a framework for understanding explanations of behavior at three levels of analysis: the *computational level*, the *algorithmic level*, and the *implementational level*.[2] The computational level explores abstract solutions to the problem the cognitive system is trying to solve and compares ideal solutions to the behavior of the cognitive system. The algorithmic level explores how well an actual process solves the problem and compares the behavior of that process, conceived as a structure of more elementary building blocks, to how people behave. Finally, the implementational level explores how the

[2]Interestingly, Egon Brunswik made a similar proposal [Brunswik, 1952], which at the very least is an important precursor to Marr's work.

process described at the algorithmic level is actually instantiated in the world.

One way to distinguish among the possible representations consistent with some observations is to use additional constraints. The one main constraint that we will use in this dissertation to distinguish between representations is to assume that the mind uses representations that are optimal with respect to its environment and some set of prior constraints. Thus, we distinguish between psychological representations by selecting the one with maximal probability given the environmental statistics and prior constraints. This assumption can be controversial [Danks, 2008], but has a long and productive history in psychology [Brunswik, 1956, Gibson, 1966, Hecht et al., 1942] going by a number of names such as "rational analysis" [Anderson, 1990, Chater and Oaksford, 1999], Bayesian modeling [Griffiths et al., 2010, Tenenbaum et al., 2011] or ideal observer modeling [Geisler, 2003]. Bayesian modeling has been successful for exploring how people combine prior knowledge and observations to solve inductive problems (e.g., category, language, and causal learning; Tenenbaum et al., 2011), which suggests it can be a useful to apply to the problem of learning representations. We will demonstrate that this optimality constraint can be used to explore how people form representations based on their prior knowledge and observations. Our claim is not that people always form representations in an optimal manner, but rather that exploring the principles that govern how a Bayesian model forms representations will be a useful tool for understanding how people form representations.

1.2 Dissertation outline

The plan of the dissertation is as follows. Chapter 2 reviews previous empirical, theoretical, and computational research in the literature on perceptual and categorization representations: how people learn them, and how they affect behavior. We discuss a diverse set of empirical results ranging from classic learning effects that predate the cognitive revolution to recent results on how people learn representations with and without feedback. Additionally, we critique previously proposed computational methods for inferring representations.

Chapter 3 begins by defining representation learning as a problem of matrix factorization. It decomposes the problem into the product of two matrices: (1) an ownership matrix, which is a binary matrix that specifies which representations are used to encode each object, and (2) an instantiation matrix, which is a matrix that specifies how the latent representations produce observable properties. Based on the previous research discussed in Chapter 2, we propose criteria for how computational models should learn feature representations. Then, we derive a nonparametric Bayesian model as the limit of a parametric Bayesian model and show how it can be used as a computational framework for inferring feature representations from raw sensory inputs. Finally, we conclude the chapter by demonstrating that the proposed framework satisfies the criteria for a computational framework for learning representations.

Chapter 4 starts by using the computational framework to explain a previous percep-

tual learning result. Based on this explanation, we describe and test a prediction of the computational framework: the features used to represent a set of objects should be affected by the distribution of parts over the objects (i.e., by the degree of co-variation). As the prediction is domain-general, the remainder of the chapter verifies the prediction in three different domains: 2-D binary images, 3-D rendered grayscale objects, and animal concepts. Additionally, we demonstrate that simple prototype or exemplar models accounts cannot explain the behavioral results.

In the previous two chapters, the computational models can only infer features that occur identically on each instantiation. However, the observable properties of features are often transformed across occurrences (e.g., a feature's retinal image after a saccade is approximately a translation over the retina). In Chapter 5, we define the transformed Indian buffet process, a novel nonparametric Bayesian model, which extends the original model to infer transformation-invariant features. After demonstrating its power on a standard machine learning benchmark (the Two Lines problem), we show that it can explain previous results by Fiser and Aslin (2001), who taught participants object chunks through passive observation. We conclude the chapter by returning to how context can be used by people and the model to disambiguate between the two potential representations for the object presented in Figure 1d.

In Chapter 6, we discuss how various expectations can be incorporated into or learned by the model by modifying the generative processes that define a model in our framework. First, we demonstrate how to include a perceptual bias into the features learned by the model, which encourages the model to infer contiguous features. Second, we contrast two possible methods for incorporating categorization information into the model, so that the category membership of the observed objects affects the features inferred by the model. Third, we explain a previous behavioral result and test a novel prediction in our framework by including hidden variables that enable people to learn expectations as to how transformations are dependent. Finally, we present an incremental learning form of the model that captures a previously found effect of the order of object presentation on the features people learn.

Chapter 7 summarizes the key points of the dissertation, discusses the theoretical status and implications of the computational framework, raises some future directions for research and concludes the dissertation.

Chapter 2

How flexible are representations?

In this chapter, we first define features and review how they are used in object recognition, categorization, and to define similarity. Then, we explore the animal, perceptual, and category learning literature that is pertinent for learning representations. Finally, we discuss previous computational proposals in psychology for inferring representations.

2.1 Features

In this section, we first review how feature representations have been used by psychologists working on object recognition, categorization, and assessment of similarity. Features[1] are a form of internal representation that is widely used in psychological theories [Markman, 1998, Palmer, 1999, Tversky, 1977]. They are elementary units that can be simple, such as the presence of a horizontal line at a particular location, or more complex, such as connectedness. They can be discrete, whether binary (present or absent) or one of a countable set of values (e.g., the style of a line might be dashed, dotted or solid), or continuous (e.g., the length of a line).[2] The response to a given input stimulus is generated by first encoding the values of the features for the input, and then making a decision on the basis of those feature values. If two people represent the same stimulus using different features, then their response to that stimulus may be different.[3] How people determine the appropriate features for a stimulus and how those features are inferred from the sensory data is typically

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

[2]Previously, researchers have distinguished between discrete and continuous valued properties, where continuous properties are called dimensions (such as, Krumhansl, 1978). For our purposes, we consider both discrete and continuous properties to be features.

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

left as an open problem: models of human cognition typically assume that people adopt a particular feature representation for a stimulus without specifying how. However, a full explanation of human behavior demands a theoretical account of how representations are determined [Edelman, 1999, Garner, 1974, Goldmeier, 1972, Goldstone, 2003, Goodman, 1972, Kanizsa, 1979, Murphy and Medin, 1985, Navarro and Perfors, 2010, Schyns et al., 1998].

2.1.1 Object recognition

Most cognitive theories posit feature representations that serve as intermediaries between the raw sensory inputs and later psychological processing. Typically, features are used in cognitive theories to mark the perceptual and conceptual commonalities and distinctions between stimuli. For example, one classic theory of object recognition is that objects are recognized depending on the presence and absence of features, which are generated from the retinal image [Selfridge and Neisser, 1960]. For example, an "O" might be represented as having a CONCAVITY FACING DOWN and a CONCAVITY FACING UP, but not a LINE CROSSING. Cognitive processes therefore only have access to information in the retinal image indirectly through the features identified in the image. The typical assumption is that features contain the information necessary to make a decision, but without the noise and ambiguities present in the raw sensory data. In this framework, the activity of the brain can be thought of as the combination of two types of processes: perceptual processes that construct feature representations for observed sensory data and cognitive processes that make decisions based on the feature representations resulting from the output of perceptual processes.

Under this view, features are the primitive elements of our mental representations of objects, concepts, or events that are manipulated by cognitive processes. There are many different ways of using features to form representations in object recognition, ranging from simple unstructured lists of features [Selfridge and Neisser, 1960] to more complicated representations such as *augmented structural descriptions* that arrange the features into relations [Palmer, 1978, Winston, 1975]. The different types of representations have different strengths and weaknesses. For example, a coffee mug can be represented as CYLINDER AND ARC (in either order) or ON SIDE OF(ARC, CYLINDER), where the opposite order would represent an object with a cylinder on the side of an arc (Biederman, 1987). The benefit of feature lists is that they are simple and easy to compute with. The potential cost is that they may not be powerful enough to distinguish a coffee mug from a pail of paint, which could lead to very unfortunate consequences! However, though augmented structural descriptions could distinguish between a coffee mug and a pail of paint, they are notoriously difficult to compute with. Regardless, the debate between unstructured feature lists and augmented structural descriptions in object recognition is not about the use of features, but about the additional structure in object representations beyond features.[4] Other controver-

[4]For a modern perspective on human and machine object recognition, we refer the reader to a review by

sies concern the information content of features [Palmer, 1978]. For example, a long-lasting debate has explored the general structure of representations in mental imagery: Are they analog and image-like [Kosslyn, 1995] or discrete and propositional [Pylyshyn, 2002]? In this case, there has not been a definitive "winner" to the debate, possibly due to the difficulty of directly probing the representations used at various stages of cognitive processing [Anderson, 1978]. In other cases, experimental data has provided a more or less definitive answer as to the type of information in a representation, at least in certain situations. For example, Tarr and Pinker (1989) argued that object representations contain orientation information, whereas Biederman and Gerhardstein (1993) argued that they are orientation-invariant. In a series of cleverly designed behavioral experiments, previous empirical results [Bülthoff and Edelman, 1992, Jolicoeur, 1985, Tarr et al., 1998] have reported differences in reaction time depending on the orientation of the object, demonstrating that object representations contain orientation information in some situations (although this does not rule out object representations being orientation-independent in other contexts). In Chapter 8, we will discuss further how behavioral results can distinguish between representations in relation to the computational framework.

2.1.2 The role of features in categorization and similarity

One essential use of feature representations in cognitive theories is for determining the similarity between stimuli. Similarity has been argued to play a critical role in cognitive domains ranging from memory [Anderson and Bower, 1972, Roediger et al., 2001] to learning (see Mestre, 2005, and Day and Goldstone, 2012, for modern discussions of transfer of learning) to higher-level cognition [Osherson et al., 1990, Shepard, 1987] to perception [Goldmeier, 1972, Palmer, 1999]. Though theorists have proposed many different methods for calculating similarity, most proposals define the similarity of two stimuli as a function of their feature representations. For example, Tversky (1977) defined the similarity of object x to x' to be a weighted sum of three terms: a function of their common features, a function of the unique features of x, and a function of the unique features of x'.[5] Another common definition of similarity is for it to be inversely related to the distance of object x to x' in feature space, with each feature i weighted by its relative salience or importance w_i, giving the similarity $s(x, x')$ as

$$s(x, x') = \exp\left\{ -\left[\sum_i w_i (x_i - x'_i)^r \right]^{1/r} \right\}, \tag{2.1}$$

where (for simplicity) x_i is restricted to be 1 if x has feature i (and 0 if x does not), and r specifies the specific distance metric (usually $r = 1$ or $r = 2$) [Attneave, 1950,

Riesenhuber (2009).

[5] Although Tversky (1977) did not argue that a particular function should be used, but he and others (such as Shepard & Arabie, 1979) have commonly used a function that simply counts the number of features.

Medin and Schaffer, 1978, Nosofsky, 1984, Russell, 1986].[6] Defining similarity to be weighted distance in feature space has been notably successful at capturing human categorization for simple stimuli [Medin and Schaffer, 1978, Nosofsky, 1984], and it forms the foundation for categorization models of more complex stimuli and behavior [Kruschke, 1992, Nosofsky, 1986, Nosofsky, 2011, Shepard, 1987]. In fact, Nosofsky (1984, 1986) proposed that feature weights are chosen to optimize performance of previously learned categorization, which has robust empirical support [Getty et al., 1979, Nosofsky, 1984, Nosofsky, 1986, Reed, 1972, Shepard et al., 1961].

Although similarity is frequently used in psychological theories, there are serious concerns about its explanatory utility. Perhaps the most incisive criticism is that infinitely many features can be used to represent any particular object [Goldmeier, 1972, Goodman, 1972, Medin et al., 1993, Murphy and Medin, 1985]. For example, a book is a collection of papers, but it is also smaller than an airplane, larger than a mosquito, lighter than an elephant, and so forth. Any of these features could be relevant (e.g., being larger than a mosquito could be appropriate if your goal is to stop getting bitten by a mosquito; see Barsalou, 1985). This casts doubt on the explanatory power of feature-based similarity theories because determining the appropriate features (which in turn, define similarity) is a non-trivial problem [Murphy and Medin, 1985]. Many cognitive theories depend on similarity for later cognitive processes (e.g., generalization), which makes this a major weakness for most explanations using feature representations. If there is a way to select attributes independent of similarity, however, this would not be an issue.

Part of what makes determining the features of an object challenging is that these features are not intrinsic to the object, but depend on context. In other words, given only an object, it is not clear what feature representation should be used. For example, "tastes good with tomato sauce" and "needs to be boiled" are features of macaroni pasta in a cooking context, but "looks good with glitter" and "can be glued to construction paper" are features of macaroni pasta in an art context. Although it is not clear what the features of an object should be when it is presented alone and contextless (they may be selected for a "default" context though, see Garner, 1974, and Chapter 8), the appropriate feature representation is less ambiguous for a set of objects in a particular context (i.e., features are not immutable and intrinsic to each object). Note that this is different from many cognitive theories where the same representation is used for an object in any context. Thus, we can determine the appropriate features to represent an object by looking at how to represent it with the set of objects it is in. Arguably, this is one of the main purposes of features: to mark the commonalities and differences between the objects in a set [Garner, 1974]. Because there are arbitrarily many commonalities and differences that can be discovered between any pair of objects, representing objects with features is fundamentally an inductive problem.

[6] We are considering continuous dimensions as features because continuous dimensions can be represented as a series of nested discrete features [Restle, 1959]. See Tenenbaum and Griffiths (2001) for an example of how to do this.

2.2 Flexible representations

In this section, we review some previous results in animal, perceptual, and category learning that are pertinent to how people learn representations. To structure the reviewed results, we use a distinction that has been useful in the learning literature to distinguish between the different types of learning: *supervised*, *semi-supervised*, and *unsupervised* learning. Supervised learning is when the correct response, label, or feedback is given after each observed stimulus. Semi-supervised learning is when the correct response, label, or feedback is given for some (but not all) of the observed stimuli. Unsupervised learning is when the correct response, label, or feedback is not given for any stimulus. All three of these learning paradigms have been successfully used in psychology to study human and animal learning.

2.2.1 Unsupervised learning

Some of the earliest, unequivocal experimental evidence that cognitive agents learn representations even in the absence of feedback dates back to the rat maze learning studies completed by Tolman and his colleagues [Tolman, 1948], which (fittingly) were mostly completed at University of California, Berkeley. Surveying several experiments, Tolman (1948) presented compelling evidence that rats construct spatial representations of their environments, which guide their future behaviors. In one example study, Blodgett (1929), working with Tolman on his dissertation, compared the number of errors (i.e., turns that brought the rat further from the food) made by a control group of rats who were fed food in the maze every day for a week (appearing in the same location each time) to the amount of errors made by an experimental group of rats who ran through the maze for a few days without food, and then were fed food in the maze for the rest of the days of the week. The control group learned the maze from feedback over time. Unexpectedly from a feedback-driven learning point of view, the amount of error that the experimental group made on their second trip through the maze after food was introduced was equal to the amount of error that the control group made after they ran through several times with food. If rats only learn from feedback, then on the second day that food was introduced, the experimental group should make the same amount of errors as the control group on the second day. However, the experimental group on the second day food was introduced made as few errors as the control group after several days of training. Therefore, the experimental group must have learned something while running through the maze without feedback (coined "latent learning"), and this behavior is difficult to explain without positing a sort of map of the maze in the mind of the rat.

There is now a large body of evidence demonstrating that people have a domain-general capacity to form representations through unsupervised learning. Infants, children, and adults are able to learn word boundaries from passively listening to an artificial language, where the the transitional probabilities of phonemes within words are predictable, but unpredictable between words [Saffran et al., 1996, Saffran et al., 1997, Aslin et al., 1998]. In the visual modality, infants and adults are able to learn visual group-

ings over space [Fiser and Aslin, 2001, Fiser and Aslin, 2002b, Fiser and Aslin, 2005] and time [Fiser and Aslin, 2002a, Kirkham et al., 2002] that are statistically coherent from passive viewing. Furthermore, people can form "conceptual units" that are statistically coherent sequences of concepts expressed as natural images (e.g., "kitchen;" Brady & Oliva, 2008). Indeed, people have a strong capability to form representations through statistical information inherent in stimuli while learning in an unsupervised manner.

Other evidence for people being able to infer representations in an unsupervised manner comes from contextual effects [Garner, 1974, Gibson, 1969, Goldmeier, 1972, Goldstone, 2003, Kanizsa, 1979, Palmer, 1999, Schyns et al., 1998, Tversky, 1977]. This can be interpreted as unsupervised "learning," because the context can be thought of as a data set for learning. For example, Tversky (1977) found that people partitioned the same nations, Austria, Sweden, and Hungary, into two groups differently when Norway was included rather then Poland. When Norway was included, the four nations split nicely into neutral (Austria and Sweden) and non-neutral members (Norway and Hungary) based on Cold War political alliances, but when Poland was included, the four nations are segregated based on geographic proximity (Sweden and Norway vs. Austria and Poland). This difference was also seen in people's similarity judgments, and suggests that they represented the nations using different features depending on the nations in the set. Examples such as those provided by Tversky (1977) suggest that people can create or re-weight features "on the fly," forming a representation of an object without any extensive training or repeated exposure.

2.2.2 Semi-supervised learning

Due to the explosion of unlabeled data on the internet and small amount of labelled data, human and machine learning researchers have become increasingly interested in semi-supervised learning [Zhu and Goldberg, 2009], or learning from observations when some are labelled, but others are not. Although studying human semi-supervised learning is typically viewed as a recent development (e.g., Zhu, Rogers, Qian, & Kalish, 2007, attributed the first study to Stromsten, 2002), it dates back to Gibson and Gibson (1950; 1955) who conducted the first human semi-supervised learning experiment (to the best of our knowledge). In their experiment, participants were first given five seconds to study an item (a scribble with four loops), which is the one labelled example in the study. Then, they went given a deck of 34 cards that had images printed on them, where five of the cards in the deck were the studied item, seventeen were extremely similar to the target, and twelve were extremely dissimilar to the target. Next, as the participant went through the cards sequentially, she classified it as the studied or not the studied example. Importantly, no feedback was given. After going through the deck, if she made more than one error, she studied the labelled example and repeated the procedure until reaching the criterion or failing to do so eight times. Six to eight year olds never reached criterion, but 8 1/2 to 10 year olds and adults both reach criterion. As they learned to recognize the target, they were able to learn in a semi-supervised manner [Gibson and Gibson, 1950, Gibson and Gibson, 1955]. Although Gibson and Gib-

son (1950; 1955) were the first to demonstrate that people can learn in a semi-supervised manner, they did not explicitly investigate the contributions of the labelled and unlabeled information or how an ideal observer would learn from these sources of information. The first exploration of how the distribution of unlabeled information can affect human learning was conducted by Zaki & Nosofsky (2007) who taught participants a "dot" prototype category [Posner and Keele, 1968] in a supervised manner (exemplars generated as Normal deviations from a prototype). The participant did not receive any feedback during testing, and the testing distribution either conformed to a Normal distribution, but had lower variance than the one used in training, or to a Normal distribution with lower variance and a new mean. Participants shift their responses accordingly as predicted by an exemplar or ideal observer model. More recent work has interfaced with machine learning research on semi-supervised learning and used more complex distributions (e.g., mixtures of Normals) to explore the depth of human semi-supervised learning, and to distinguish explicitly between supervised and semi-supervised learning strategies [Zhu et al., 2007, Vandist et al., 2009, Kalish et al., 2010]. As the extent that people learn in a semi-supervised manner is currently unclear, this is an active and undeveloped area of research.

2.2.3 Supervised learning

There is a large body of research on how feedback and learning categories can affect perceptual representations. We begin by reviewing research into expertise, which explores how people who train for years in a particular domain (experts) represent that domain differently than lay people. We then discuss research that has attempted to train people to become experts in laboratory settings, and conclude this part of the review by discussing differentiation and unitization, two perceptual learning phenomena.

Further empirical evidence for flexible features comes from the perceptual learning literature on expertise, where people become capable of performing a task that they previously were unable to perform, which indicates that they created a new feature or re-weighted features that previously had small weights. Experts in a domain list different features from novices [Tanaka and Taylor, 1991] and when quick and accurate perceptual judgments are necessary, experts infer useful domain-specific features [Biederman and Shiffrar, 1987, Sowden et al., 2000]. For example, chick-sexing experts can tell the gender of about 1,400 newly hatched chicks in an hour with 98% accuracy [Lunn, 1948]. In a controlled study using images of newly hatched chicks, naive participants could barely tell the gender of chicks above chance, but when told the diagnostic feature for distinguishing males from females, could perform on par if not better than the experts [Biederman and Shiffrar, 1987].

For the most part, expertise studies are correlational as it is unclear (though unlikely) whether people became experts because they happen to have features that enabled them to perform tasks that novices cannot do (or because they worked hard to become an expert). Additionally, it is not clear what aspect of becoming an expert is the cause of feature learning. Using categoriza-

tion training [Goldstone, 2000, Goldstone and Steyvers, 2001, Lin and Murphy, 1997, Pevtzow and Goldstone, 1994, Schyns and Murphy, 1994, Schyns and Rodet, 1997], repeated visual search [Shiffrin and Lightfoot, 1997], and prolonged exposure [Gauthier and Tarr, 1997, Gauthier et al., 1998], researchers have successfully emulated the process of acquiring expertise and taught participants new features in the laboratory. In categorization training studies, a set of parts are diagnostic of a category the experimenter has in mind, but initially they are not inferred by participants as features. After repeated feedback, participants discover that the part is diagnostic and it is inferred as a feature. For example, Pevtzow and Goldstone (1994) created a stimulus set in which each object shares one part with two other objects. Participants learned categories in which each object shared one diagnostic part with the other object in its category. After categorization training, they inferred the part diagnostic for categorization as a feature, but not the part non-diagnostic for their categorization training. Furthermore, perceptual expectations guide how representations are learned through feedback (such as expecting that features be contiguous in vision; Goldstone, 2000). Finally, prior conceptual knowledge can also affect the features used to represent an observed set of objects while learning a category [Lin and Murphy, 1997, Wisniewski and Medin, 1994]. For example, Wisniewski and Medin (1994) found that people represented the same children's drawings using different features depending on the types of children they were told produced the drawings (e.g., creative vs. non-creative or low vs. high IQ children).

Another important area in perceptual learning concerns how people learn to unitize and differentiate between perceptual dimensions. Research on perceptual learning has discussed two ways in which features can change, termed *unitization* and *differentiation*. Unitization occurs when features that were previously perceived as distinct are perceived as a single feature. Evidence that people are capable of unitizing features comes from Shiffrin and Lightfoot (1997)'s visual search experiment, in which elements that were previously perceived as having separate features (as indicated by serial visual search) come to be represented by a single feature (as indicated by "pop-out"). In contrast, differentiation occurs when a single feature splits into multiple new features. Evidence for this occurring comes from Goldstone (1994), who demonstrated that color novices who previously could not distinguish between a color's saturation and brightness can learn to distinguish between these two dimensions.

2.2.4 Feature creation or re-weighting?

In order for features to be effective, there should be some feature representation that corresponds to the raw sensory input of any possible stimulus; otherwise, how could you represent the input when it occurred?. Satisfying this constraint is a formidable problem: How do you define a set of features that is able to represent any possible object? Some approaches, such as the Geon theory of object recognition [Biederman, 1987], argue for a *fixed* feature set, where stimuli are represented using a space of a fixed set of features composed with a fixed set of relations (e.g., CYLINDER and ON TOP OF). The power of composing features

with relations is that a surprisingly large number of objects can be represented even with a relatively small number of features and relations. For example, over 10^{14} distinct objects can be created using five out of ten features that can be related in eight different ways [Goldstone, 2003]. Thus, a proposal like the Geon theory, which posits 36 features and five relations can represent an astronomically large number of objects.

Even though it may be possible to define a set of features that can represent all the objects that a person could plausibly encounter in her lifetime,[7] some researchers argue that these approaches do not adequately capture the representational power and flexibility of human cognition. Instead, they argue for *flexible* feature theories [Goldstone, 2003, Goldstone et al., 2008, Hoffman and Richards, 1985, Schyns et al., 1998], where the set of possible features can adapt with experience. They argue that it is implausible for our perceptual systems to be hard-wired in such a way that it would be possible to recognize quickly objects that are now important to people (such as cell phones or cars) that are very different from the objects that were present during the evolution of our species [Goldstone, 2003]. Furthermore, Hoffman and Richards (1985) argued that people infer features of objects at changes in concavity and Braunstein et al. (1989) demonstrated that people segment (some) novel objects into features at these points.

Although the need of our perceptual systems to recognize objects not present during the evolution of our species is hardly controversial, what is contentious is whether the theoretical arguments and empirical studies discussed above should be treated as evidence for feature creation or as evidence for re-weighting (or recombination) of pre-existing features (see the discussion portion of Schyns et al., 1998). To some degree, this may merely be a matter of terminology. How different is a cognitive model with flexible features from one with a fixed feature set of infinite size (encompassing all possible features) with varying feature weights? "Learning new features" or "forgetting old features" can be thought of as feature weights increasing from or decreasing to zero, respectively. When a feature differentiates into two or more features, the original feature's weight decreases to zero and the weights of the previously unused features increase from zero (and conversely for unitization). Thus, at the computational level [Marr, 1982], which analyzes human behavior in terms of the abstract problem it is trying to solve, we cannot distinguish between these two approaches. However, the need for an explanation of how people infer features flexibly (whether through feature weight change or through feature construction) depending on context and experience is uncontroversial, and we will focus on this problem for a majority of this dissertation.

2.2.5 Summary

Feature representations play an essential role in psychological theories. They determine many aspects of human behavior in many domains. Because the same object may be represented

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

differently in different contexts, there are many possible feature representations for an object whose context has not been specified. Thus, inferring a feature representation for an object is an *inductive* problem. Because features determine how people generalize from one object to another (assuming they only generalize on the basis of common features), we can test how people infer the feature representation of an object in different contexts based on how they generalize to other objects.

2.3 Computational approaches to inferring representations

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

2.3.1 Weight change approaches

One potential solution for capturing the effects of context on features is to assume a large set of features is known *a priori* (e.g., raw sensory data) and provide a systematic method for determining when and how the weights of these features change. This approach has been explored using both connectionist [Goldstone, 2003, Grossberg, 1987, Rumelhart and Zipser, 1985] and Bayesian [Zeigenfuse and Lee, 2010] models.

One early proposal by Nosofsky (1984; 1986) was that feature weights are chosen to optimize performance for previously learned categorization tasks. Indeed, this proposal has been verified extensively [Nosofsky, 1984, Nosofsky, 1986, Nosofsky, 1991]. One of the most successful models for changing the weights of features is ALCOVE [Kruschke, 1992, Lee and Navarro, 2002]. According to the ALCOVE model, the similarity defined by the generalized context model (GCM) feeds into a neural network, which categorizes an object based on its similarity to stored exemplars of each category. The neural network learns feature weights w_i according to error-driven learning. Though this is very successful for the small feature sets typically used in categorization studies (e.g., Gluck and Bower, 1988), the number of potential features people may use is potentially infinite or at the very least, exponential in the number of raw sensory units (e.g., every combination of pixels

in an image is potentially a feature).[8] Due to the bias-variance tradeoff [Geman et al., 1992], very strong constraints on the types of features (i.e., strong inductive biases, Griffiths, 2010) will be needed to learn features effectively from somewhat realistic sensory data as the space of possible features (e.g., all combinations of pixels) is astronomically large if not infinite [Schyns and Rodet, 1997, Schyns et al., 1998]. Although categorization is an important catalyst for feature representation change, different features can be inferred to represent the same object in different contexts without any category information, as shown by our earlier discussion and the results we will present in Chapters 4, 5, and 6. In other words, categorization is sufficient, but not necessary for changing the features used to represent an object. Another sufficient condition for feature representation change is the distribution of parts over the context of an object. Thus, methods based on ALCOVE are insufficient as they do not infer feature weight values in the absence of category information (see Pothos and Bailey, 2009, for extensions of the GCM that learn attention weights in the absence of category information).

Using competitive learning rules, artificial neural network methods have been developed to infer feature representations from the distribution of parts over a set of objects without any category information [Grossberg, 1987, Rumelhart and Zipser, 1985]. The networks start with a fixed number of features whose weights are initialized to random values, which means that they are not "specialized" to detect any particular visual properties. When an object is presented to the network, the network attempts to recreate it using its features. According to competitive learning rules, the feature that best recreates the observed object is considered the "winner" and its weights are updated using error-driven learning based on the difference between the observed and recreated objects (at a faster rate than the weights of "losing" features). As this procedure is repeated over an object set, the features differentiate and become specialized to detect different visual properties. Once the procedure converges, each object is represented using some number of its vocabulary of inferred features.

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

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

information to infer features.

Zeigenfuse and Lee (2010) describe a Bayesian feature weight change approach, in which they assume that human similarity judgments are determined by a slightly more complex form of Equation 2.1 and then use Bayesian inference to determine the weight of each feature's importance in human similarity judgments. When their method is given a set of objects in a particular domain (e.g., farm animals) represented by a given large set of features and human judgments as to the similarity of various pairs of objects in the domain (e.g., horse and cow), it infers a smaller number of features with non-negligible weights that adequately reproduces human similarity judgments. Although this is a useful tool for reducing the number of features needed to adequately capture human similarity judgments and for investigating which statistical properties yield large feature weights, it does not alter its feature representation depending on context or experience. Thus, it is not appropriate for our purpose of understanding how people infer feature representations.

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

2.3.2 Structure change approaches

Though all current approaches infer features for which the model's structure is known *a priori* (e.g., knowing the number of features ahead of time), there is no reason to believe that it is impossible to develop a model free of such requirements. In fact, one method that would most likely be successful would be to specify a large number of potential features with a penalty for non-zero weights, which would bias them toward leaving a large number of the features unused. In fact, this is similar to the framework we introduce in the next section.

A similar approach exists in the neural network literature, which uses a mechanism for altering the architecture of the neural network, such as cascade-correlation [Fahlman and Lebiere, 1990]. Cascade-correlation recruits new nodes in the neural network when the current architecture does not adequately capture the pattern of input and outputs. This has been used successfully to capture developmental stages [Shultz and Sirois, 2008] and intermediate representations used for categorization [Love et al., 2004]. A flexible connectionist model for learning features using either mechanism would be very interesting and most likely could capture many of the feature learning results in Chapter 4; however, it would still have difficulty learning features whose images are transformed when instantiated in objects.

One exception is the Bayesian chunk learner [Orban et al., 2008], which was used to investigate 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 do the models we have discussed so far), the input to their model solves one aspect of the problem we are interested in: namely, how does the human perceptual system identify primitives?. However, once those primitives are provided, the assumptions of this model are similar to those behind the initial model constructed in the framework we will propose in the next chapter.[9]

2.3.3 Machine learning approaches

Another potential approach for deriving human feature learning models is to borrow related computational methods from machine learning. One of the main techniques for feature learning in machine learning is dimensionality reduction, which finds a smaller encoding for a set of observed objects. 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 (PCA) and independent component analysis (ICA). PCA encodes a set of given objects in terms of the eigenvectors of its covariance matrix (the eigenvectors can be thought of as features). Usually only a subset of the 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 low-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 et al., 1998, Edelman and Intrator, 1997]. [10] Heuristic methods are usually used to select the number of eigenvectors to use, 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 inde-

[9]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 models will assume 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 and Richardson, 2001, Rasmussen and Ghahramani, 2001].

[10]Another popular dimensionality reduction technique is projection pursuit [Friedman and Tukey, 1974], where an object set is reduced to a smaller set of dimensions that optimize some measure of the "interestingness" of each dimension. When the variance of stimuli on the dimension is the definition of the interestingness of a dimension, projection pursuit is equivalent to PCA [Carreira-Perpinán, 1997].

pendent from each other [Hyvarinen et al., 2001]. This idea is similar to approaches used to model visual cortex [Bell and Sejnowski, 1995, Olshausen and Field, 1996], and has recently been explored as a way of explaining human dimensionality reduction in multiple domains [Hansen et al., 2005]. Olshausen and Field (1996) have shown that ICA with an added sparsity bias infers features that are very similar to proposed neural representations used in the primary visual cortex to encode a set of natural scenes. While PCA and ICA might seem like good candidates for how people form low-dimensional representations of complex stimuli, in Chapter 4 we will present evidence that features identified by these methods are not behaviorally relevant.

Work on language acquisition has provided many computational models of how people might learn linguistic representations without explicit instruction. Primarily, researchers have explored how people learn grammars [Goldstein et al., 2010, Solan et al., 2005] and segment continuous streams of phonemes into words [Batchelder, 2002, Brent, 1999, Goldstein et al., 2010, Goldwater et al., 2009, Perruchet and Vinter, 1998, Venkataraman, 2001]. These models share with our analysis an interest in the use of distributional information to determine 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 will be independent in our model). The linguistic models infer one hidden unit per item, which is appropriate for segmenting phonemes into words (each phoneme should be assigned to only one word), but not for inferring features to represent images (where multiple features should be assigned to each image). The models used by Brent (1999) and Goldwater et al. (2009) are grounded in nonparametric Bayesian statistics, which provides a general-purpose tool for solving the problem of determining the amount of structure expressed in observed data. In the next chapter, we explore applying this tool to the problem of feature learning.

Chapter 3

Computational framework

In this chapter, we describe a rational analysis of how features should be inferred to represent a set of objects [Anderson, 1990, Oaksford and Chater, 1998]. A rational analysis compares the behavior produced by a cognitive process to the ideal solution of the underlying computational problem. It is at the abstract computational level [Marr, 1982], which focuses on the optimal solution to the mathematical problem of inferring context-sensitive feature representations. This constrains the cognitive processes people might be using to infer feature representations, which we explore in more depth in Chapter 6.

To perform the rational analysis, we outline criteria for a solution to the problem of context-sensitive inference of feature representations. We then present a computational framework for solving this problem using nonparametric Bayesian models. We first develop an initial model that can satisfy some of the criteria, the implications of which we will explore in Chapter 4. Then, we go beyond this previous work by defining a set of new models that can address the remaining criteria in Chapters 5 and 6.

3.1 Criteria for a Solution to the Computational Problem

Based on the literature on human feature learning reviewed above, we suggest that any computational framework for solving the problem of forming context-sensitive feature representations should be able to satisfy the following criteria:

Criterion 1: Sensory primitives

As it is difficult to define a set of primitives that is capable of representing all ways an object can be represented in different contexts, the framework should construct features from raw sensory data [Goldstone, 2003, Hoffman and Richards, 1985, Schyns et al., 1998].

Criterion 2: Unlimited features

Consistent with the human capability of being able to identify a potentially infinite number of features to represent any object [Goldmeier, 1972, Goodman, 1972, Medin et al., 1993, Murphy and Medin, 1985], there should be no arbitrary limit on the number of features that can be used. The number of features should be inferred based on the observed set of objects.

Criterion 3: Context sensitivity

Because one main purpose of features is to denote the commonalities and differences between the objects in a set, feature representations are inferred with respect to a set of objects [Garner, 1974, Selfridge, 1955, Tversky, 1977]. Note that this criterion predicts contextual effects: when an object is presented with two different sets of objects, it may be represented using different feature representations.

Criterion 4: Prior expectations

The framework should easily include perceptual and conceptual constraints, if such constraints are relevant. For example, people are biased towards inferring contiguous features [Goldstone, 2000], and simpler feature representations [Chater and Vitanyi, 2003, Hochberg and McAlister, 1953]. However, it should still be able to infer features when any of the above types of information are absent.

Criterion 5: Transformational invariance

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

Criterion 6: Category diagnosticity

When objects are categorized, the features inferred by a model in the framework should be affected appropriately. For example, the model should infer features diagnostic for categorization [Pevtzow and Goldstone, 1994, Schyns and Murphy, 1994].

Criterion 7: Incremental learning

People learn features incrementally as they are needed to represent novel images or explain category information. The order in which objects are presented should therefore be able to affect the features inferred by the model [Schyns and Rodet, 1997].

Although many previous models satisfy some of the criteria outlined above, no existing computational framework can address all of them. In the remainder of this chapter, we formalize the mathematical problem of learning feature representations and outline an approach that we will argue can satisfy all of the criteria.

3.2 Problem Formalization

Our computational problem is as follows: find a feature representation for a given set of objects satisfying the criteria outlined above. The set of objects are grouped into a matrix $\mathbf{X}$, whose N rows contain each object and D columns contain their observable properties. Although our framework is agnostic about the specific modality (it has been applied to auditory inputs and extended to multisensory data, Yildirim & Jacobs, 2012), the observable properties of each object $\mathbf{x}$ (a row of the matrix) are an array of pixels representing the intensity of light reflected onto the retina. The two-dimensional array is converted into a one-dimensional vector of size D.[1] In this chapter, we further simplify the input to assume that each pixel is binary, though it is straightforward to use other types of observable properties (e.g., we will use grayscale images in Chapter 4).

One way to view this problem mathematically is as a problem of matrix factorization. Tentatively, let us assume that we know the number of features is K (we will relax this assumption later). As illustrated in Figure 3.1, the goal of a feature learning model is to reconstruct $\mathbf{X}$ as the product of two matrices: (1) a $N \times K$ binary feature ownership matrix $\mathbf{Z}$, where $z_{nk} = 1$ if object n has feature k and (2) $K \times D$ feature "image" matrix $\mathbf{Y}$ that encodes how each feature gets instantiated in an object (e.g., if a feature of a mug was its handle, the corresponding $\mathbf{y}$ would be the handle's image.). Informally, this amounts to recreating each observed object $\mathbf{x}$ by superimposing the images $\mathbf{Y}$ of the features it has (given by its feature ownership vector $\mathbf{z}$). In this view, the model is solving the matrix equation $\mathbf{X} = \mathbf{ZY}$ for two unknown variables, the matrices $\mathbf{Z}$ and $\mathbf{Y}$ given only one variable, the matrix $\mathbf{X}$. Thus, the solution is underconstrained (it is like solving a linear regression equation for the predictors and their weights simultaneously), meaning that additional information is necessary to solve for $\mathbf{Z}$ and $\mathbf{Y}$.

The solution to this problem is given by Bayesian inference [Geisler, 2003, Griffiths et al., 2008]. Here, the observed set of objects is our data and the matrices $\mathbf{Z}$ and $\mathbf{Y}$ form our hypothesis. Thus, applying Bayes' rule gives us the following solution:

$$P(\mathbf{Z}, \mathbf{Y}|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}) P(\mathbf{Z}) P(\mathbf{Y}). \tag{3.1}$$

Thus, we have decomposed our original problem into three subproblems: find feature ownership and image matrices $\mathbf{Z}$ and $\mathbf{Y}$ such that they reconstruct the observed set of objects

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

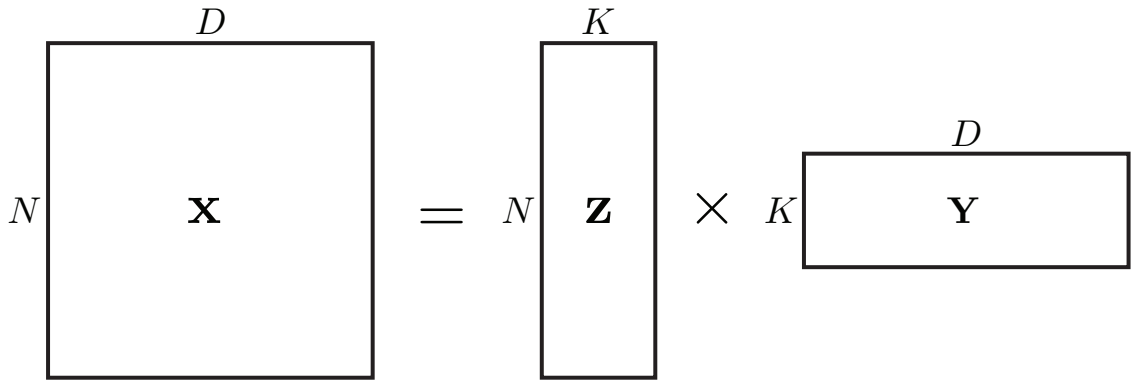


Figure 3.1: Formally, the problem of feature learning reduces to a matrix factorization problem. Represent a matrix $\mathbf{X}$, whose rows are the objects $\mathbf{x}$, using the product of a binary matrix $\mathbf{Z}$ (the feature ownership matrix), whose rows correspond to the features each object has, and a matrix $\mathbf{Y}$ that encodes the observable consequences of having each feature (the feature image matrix). For example, if the objects are binary images, then each element of $(\mathbf{YZ})_{nd}$ encodes the number of object n's features whose image has pixel d on. This is not quite correct as $\mathbf{X}$ is a binary matrix. More precisely, as $(\mathbf{YZ})_{nd}$ increases, so should the probability that $x_{nd} = 1$. In this chapter, this is done by assuming that x_{nd} is noisy-OR distributed (Pearl, 1988) with parameter given by $(\mathbf{YZ})_{nd}$

$\mathbf{X}$ well (designated by the "likelihood" $P(\mathbf{X}|\mathbf{Z}, \mathbf{Y})$), and capture our prior expectations as to what makes a good feature ownership matrix ($P(\mathbf{Z})$) and a good feature image matrix ($P(\mathbf{Y})$). This is our proposed computational framework. A model in the computational framework is specified by how these three components are defined as well as the method of inference.

To relax the assumption that we know the number of features *a priori*, we use a nonparametric Bayesian prior probability distribution on $\mathbf{Z}$. Nonparametric Bayesian models infer probability distributions that are potentially infinitely complex (in that the number of parameters required by a nonparametric Bayesian model to infer a distribution is infinite), but prefer simpler probability distributions, in the sense that they are biased to infer probability distributions that can be represented using fewer parameters [Jordan, 2010]. Intuitively, the result is that the model infers the minimal amount of structure necessary to represent the observed data because the model trades off encoding the observed data with a bias towards simplicity.

In our particular case, the most common nonparametric Bayesian model for $\mathbf{Z}$, a matrix with a finite number of rows (the objects) and an infinite number of columns (the K features), is the Indian buffet process (IBP; Griffiths & Ghahramani, 2005; 2011). It generates matrices with an infinite number of columns, but only a finite, random, number of them have at least

one non-zero element.[2] As features are only part of a feature representation for a set of objects when they are actually used by at least one object, the infinite columns containing all zeroes have no consequence. Thus, using this process, we can relax the assumption of knowing the number of features *a priori* and infer the number of features needed to represent the observed set of objects.

3.3 An Initial Model

Based on our analysis above, we define a model in the computational framework by specifying three probability distributions: the likelihood $P(\mathbf{X}|\mathbf{Z}, \mathbf{Y})$, the prior on feature ownership matrices $P(\mathbf{Z})$, and the prior on feature images $P(\mathbf{Y})$. Once these three probability distributions are defined, we can adapt approximation techniques from machine learning to infer feature representations for a given set of objects (which we discuss further in Appendices A and F, and Chapter 6). As we will see later in Chapters 5 and 6, how these three components are defined, what sorts of other data are incorporated with them (e.g., categories), and the machine learning method used to infer the feature representations, determines the behavior of the feature learning model. We begin by first explaining in depth the prior on feature ownership matrices, and then provide an example of the simplest possible model in this framework, and evaluate it against our criteria.

Feature ownership matrices are binary matrices and so (supposing again that the number of features is known *a priori*), one simple method for defining a probability distribution on them is to flip a coin for each entry, putting a 1 in that entry for "heads" and 0 for "tails." One issue with this simple model is that all features have the same probability of being in an object. This is easily remedied by using a different "coin" for each column k, where the probability of heads for the coin used in column k is a parameter π_k. With this probability model for binary matrices, we can relax the assumption of knowing the number of features *a priori*. If we simultaneously increase the number of columns K to infinity and decrease the probability of a feature being in an object π_k to zero at corresponding rates, the result is a probability model that only generates matrices with a finite number of columns having at least one non-zero element and an infinite number of columns containing all zeroes. This is because for there to be "nice" limiting behavior as K increases, the probability of heads has to approach so close to zero that the probability of there being a one in all future columns approaches zero (i.e., at some point, the infinite "remaining features" will not be in any object). For more technical details, see Griffiths and Ghahramani (2011).

The probability model resulting from this limiting construction is equivalent to an implicit probability distribution on binary matrices given by a sequential culinary metaphor, where objects (or rows) are "customers" and features (or columns) are "dishes" in an Indian buffet.

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

Imagine an Indian buffet where the dishes are arranged sequentially (in order of the first customer that took a dish). Customers enter the Indian buffet one at a time, and we record in a matrix the dishes that they take according to the following set of rules: The first customer samples a number of dishes that is drawn from a Poisson distribution with parameter α. Each successive customer i takes a dish with probability m_k/i, where m_k is the number of people who previously sampled dish k, and then a random number of sample new dishes from a Poisson distribution with parameter α/i (where dividing by i captures the intuition that as more customers enter the restaurant, it should be less likely that a new dish is sampled). The number of previous customers who have sampled the dish, m_k, is normalized by the "arbitrary" customer index i because $i-1$ is the number of previous customers who could have taken the dish (and the denominator is one larger due to our uniform prior beliefs over the probability that a customer takes the dish). This encodes the intuition that features which have been sampled by a large number of the previous customers are more likely to be sampled by new customers. The recorded matrix has the same probability under the culinary metaphor and the original limiting process.

Figure 3.2 illustrates how the culinary metaphor works for an example with three customers. The circles in Figure 3.2a denote dishes, the numbers around each dish denote the customers taking that dish, and the image above each dish is the observable consequence associated with taking that dish. (We will explain how these are generated shortly). The culinary metaphor generates a corresponding feature ownership matrix $\mathbf{Z}$, which is shown in Figure 3.2b. The observable consequences of each feature $\mathbf{Y}$ are displayed in Figure 3.2c. Finally, Figure 3.2d demonstrates how each object is created by superimposing the observable consequences of the features it has.

The explicit form for the probability of the feature ownership matrix[3] after N customers who have taken K_+ dishes at least once is

$$\frac{\exp\{-\alpha H_N\}}{\prod_{h=1}^{2^N-1} K_h!}\alpha^{K_+}\prod_{k=1}^{K_+}\frac{(N-m_k)!(m_k-1)!}{N!} \tag{3.2}$$

where K_h is the number of features with "history" h (the history is the natural number equivalent to the column of the feature interpreted as a binary number), K_+ is the number of columns with at least one non-zero entry (number of "used" features), and H_N is the Nth harmonic number ($H_i = \sum_{j=1}^{i} j^{-1}$). For example, the history for the second feature in Figure 3.2b is 3 ($(1,1,0) = 2^0 + 2^1 = 3$). Note $K_h!$ is 1 unless more than one feature has history h. We refer the reader to Griffiths and Ghahramani (2011), for the full derivation details, and instead provide intuitions for the relationship between the culinary metaphor and Equation 3.2. The term $\exp\{-\alpha H_N\}(\alpha)^{K_+}$ is due to each time a new dish is sampled (compare it to the form of a sum of Poisson distributed random variables), the term $\prod_{k=1}^{K_+}\frac{(N-m_k)!(m_k-1)!}{N!}$ encodes

[3]Technically, this is the probability of all matrices equivalent to $\mathbf{Z}$ in the sense that the columns are the same (ignoring any differences due to the order of the columns).

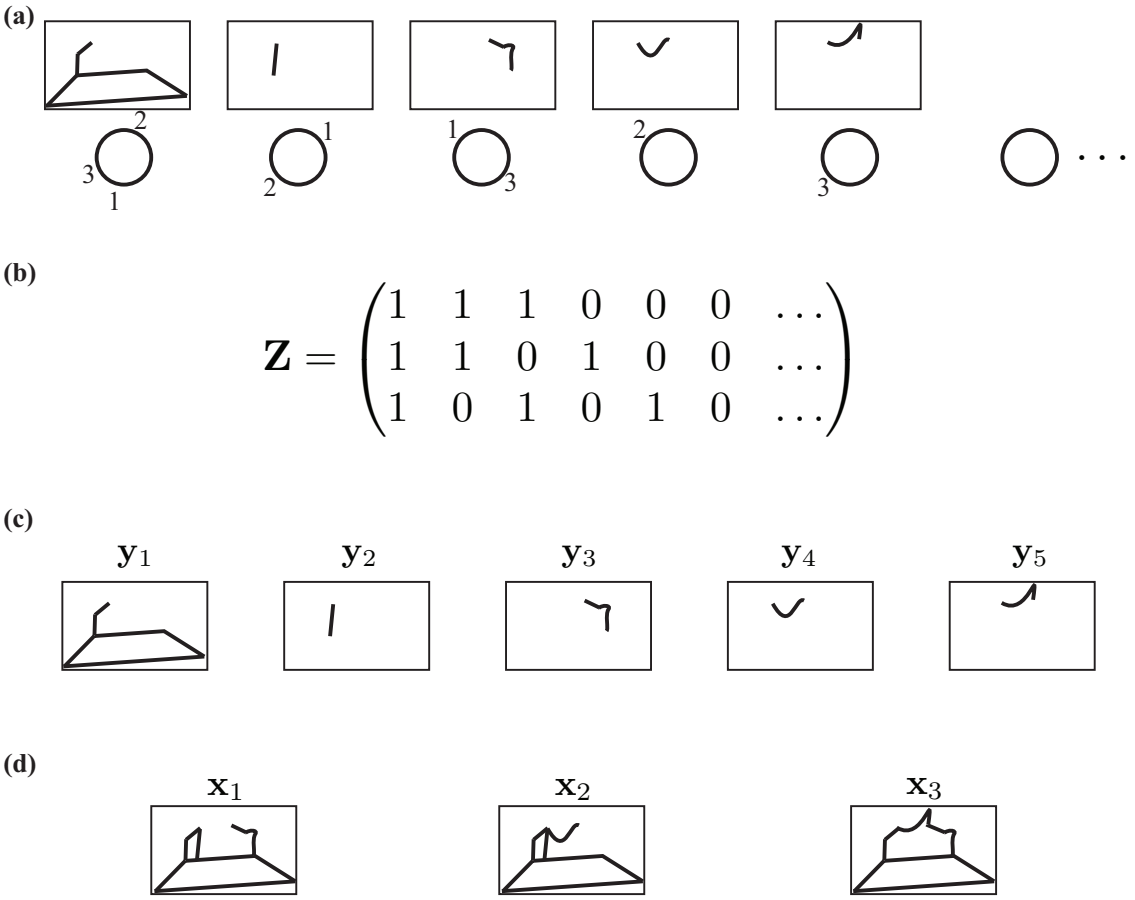


Figure 3.2: An illustration of the relation between the culinary metaphor, the Indian buffet process, and the model. (a) The culinary metaphor for the Indian buffet process. The numbers are customers (objects) and the circles are dishes (features). A number adjacent to a table represents its corresponding object taking that feature. A feature image is generated for each dish, which appears above the circle for each dish. (b) The equivalent feature ownership matrix represented by the culinary metaphor above. (c) The feature images generated from the feature image prior for each feature. (d) The reconstructed objects using $\mathbf{Z}$ and $\mathbf{Y}$ defined in (b) and (c) respectively.

the probability customers chose previously sampled dishes, and $\frac{1}{\prod_{h=1}^{2^N-1} K_h!}$ is a normalization constant that accounts for dishes that are "equivalent" to the IBP (i.e., customers who made the same decisions for two or more dishes).

To define the likelihood, $P(\mathbf{X}|\mathbf{Z}, \mathbf{Y})$, we form our observation matrix $\mathbf{X}$ that consists of the values of D observable binary properties (e.g., pixels in an image) for N objects ($\mathbf{X} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T]$, where $\mathbf{x}_i^T \in \{0, 1\}^D$). The likelihood compares the reconstructed object using the inferred feature representation ($\mathbf{YZ}$) to the observed set of objects. The format of the observable properties in $\mathbf{X}$, which for purposes of exposition are always the binary pixels of an image (though we will use "concept primitives" and grayscale pixels in Chapter 4). Each element of $(\mathbf{YZ})_{nd}$ encodes the number of object n's features whose image has pixel d on. As $(\mathbf{YZ})_{nd}$ increases, so should the probability that $x_{nd} = 1$. Thus, the appropriate likelihood for our purposes is the noisy-OR distribution [Pearl, 1988, Wood et al., 2006], which seems to capture the assumptions people have for how an observed effect is produced by multiple hidden causes [Cheng, 1997, Griffiths and Tenenbaum, 2005]. The total number of features that can turn on pixel d in object n is given by the inner product of the feature assignment vector for the object $\mathbf{z}_n$ with the column vector $\mathbf{y}_d$, which indicates whether or not pixel d is in the different feature images. Assuming each pixel is generated independently, the probability that a pixel is on (takes the value 1) is

$$P(x_{nd} = 1|\mathbf{Z}, \mathbf{Y}, \lambda, \epsilon) = 1 - (1 - \lambda)^{\mathbf{z}_n \mathbf{y}_d}(1 - \epsilon) \tag{3.3}$$

where ϵ is the probability a pixel is turned on by chance and λ is the efficacy of a feature (the probability a feature with the pixel on in its image turns on the pixel in its object). One interpretation of Equation 3.3 is that it assumes each pixel is off *a priori*, but each feature that object n has with pixel d on turns the pixel on with probability λ and it is on by chance with probability ϵ.

The last portion of the model to specify is the prior on feature images $\mathbf{Y}$. This is where domain-specific expectations (e.g., perceptual biases such as good continuation) can be included into our framework. However, for most of our simulations, a simple "knowledge-less" prior on feature images suffices, where each pixel is "on" with probability ϕ independent of the other pixels in the image. Formally, this is a Bernoulli prior on each pixel with parameter ϕ specifying the probability the pixel is on or $P(\mathbf{Y}) = \prod_{k,d} \phi^{y_{kd}}(1 - \phi)^{1-y_{kd}}$. This prior distribution on $\mathbf{Y}$ in the model ignores spatial factors. 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 (and it simplifies inference), but it is not necessary and it is inappropriate for visual features. We can introduce a proximity bias (roughly that contiguous features are preferred) by taking our prior to be a distribution (called the Ising model; Geman & Geman, 1984) on the elements of $\mathbf{Y}$ such that neighboring pixels tend to have the same value. Let ρ 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 ϕ be the propensity for each pixel to be on ($\phi > 0.5$ indicates

pixels are more likely to be on). Neighboring pixels are connected by edges ($\{d_1, d_2\}$, 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):

$$P(\mathbf{Y}) \propto \left(\prod_{k,d} \phi^{y_{k,d}} (1-\phi)^{y_{k,d}} \right) \left(\prod_{\{d_1,d_2\} \in C} \rho^{d_1 d_2} (1-\rho)^{1-d_1 d_2} \right). \tag{3.4}$$

Calculating the normalization constant for this distribution is intractable and thus, approximation techniques are used to find $\mathbf{Y}$ under this prior (see Appendix A for details). We contrast the effects of using the two feature image priors in Chapter 6.

In Chapter 4, we will also explore modeling grayscale images. For grayscale images, we use the linear-Gaussian likelihood and an implicit Gaussian prior over feature images [Griffiths and Ghahramani, 2011]. 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

$$p(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \sigma_X) = \frac{1}{(2\pi\sigma_X^2)^{ND/2}} \exp\left\{ -\frac{1}{2\sigma_X^2} \mathrm{tr}((\mathbf{X} - \mathbf{ZY})^T(\mathbf{X} - \mathbf{ZY})) \right\}, \tag{3.5}$$

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 Equation 3.5 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 $\mathbf{Y}$.

3.4 Evaluating Nonparametric Bayesian Models Against the Criteria

We will now argue that the first two criteria are satisfied by the framework itself, as a result of the problem formulation and solving it using methods from nonparametric Bayesian statistics. In Chapters 4, 5 and 6, we will develop models that satisfy the challenges posed by the remaining criteria. For now, we merely present a sketch of these results to show how the criteria are satisfied, and defer the full presentation for later.

We now evaluate how well the proposed computational framework satisfy the criteria. Criterion 1 (sensory primitives) is satisfied by using the IBP as our prior on the feature

ownership matrix. This is because it identifies a feature image with each dish, which is inferred from the observable properties of the corresponding objects of the customers that took the dish. Through the limiting construction of the IBP, Criterion 2 (unlimited features) is satisfied as the number of features approaches infinity.[4] Although this might at first seem alarming and mystifying, the IBP penalizes each feature that is actually used (is assigned to at least one object). There are two ways that additional features are penalized by the IBP. The first way is from the third term in Equation 3.2 as it is the product of probabilities (for each feature, there is an additional $\frac{(N-m_k)!(m_k-1)!}{N!}$ term, which is less than one) and so each additional feature decreases the probability of the feature ownership matrix. Second, if α is set such that $\alpha/N < 1$, then the IBP implements a bias towards using fewer features as there is a $(\alpha/N)^{K_+}$ term, which decreases as the number of features increases. Thus, it encodes a simplicity bias, capturing an important aspect of human perception and cognition [Chater and Vitanyi, 2003, Hochberg and McAlister, 1953, Lombrozo, 2007] and so, it also satisfies an aspect of Criterion 4 (prior expectations).

The latent states (features) generated by the IBP are independent *a priori* (unlike the Chinese restaurant process; Anderson 1990; Sanborn, Griffiths, & Navarro, 2010; Griffiths, Sanborn, Canini, Navarro, & Tenenbaum, 2011). However, both the feature ownership matrix and feature images inferred by the model are dependent given the observable properties of a set of objects (through the likelihood function). Thus, the features are dependent under the IBP *a posteriori* (i.e., with the information provided by observing the objects). The current feature ownership matrix is used to infer feature images directly from the raw sensory data. Based on the other objects presented with an object (i.e., its context), the feature images or feature ownership matrix could be different. In Chapter 4, we will show that the initial model infers different features for an object appropriate for its context, and so, it satisfies Criterion 3 (context sensitivity). Furthermore, we will explore in Chapter 4 how prior expectations can be incorporated into the model by using a feature image prior $P(Y)$ that favors adjacent pixels to share the same value. Combined with our previous discussion of how the framework has a simplicity bias and an exploration of how category information can be included in Chapter 6, the framework is able to incorporate perceptual and conceptual biases, and thus, it satisfies Criterion 4 (prior expectations).

In Chapter 5, we will demonstrate how to learn transformationally invariant features and incorporate various expectations about how the transformations of different features are related (Criterion 5: transformational invariance). In Chapter 6, we explore how to encode different types of knowledge and expectations about feature representations. After exploring how to include a proximity bias (as discussed previously), we develop two accounts for the effect of categorization on the learned features (Criterion 6: category diagnosticity). Afterwards, we tackle one issue with models in the computational framework, which is that the probability distribution over feature representations is *order-invariant*, meaning they

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

infer the same features regardless the order that objects are presented. This is not consistent with Criterion 7 (incremental learning) as there are reported effects of the order of object presentation on the features learned by people [Schyns and Rodet, 1997]. We address this issue by formulating a rational process model for feature learning (in the spirit of Sanborn et al., 2010). This model is an incremental form of the transformation-invariant model, and appropriately infers different features depending on the order that objects are presented.

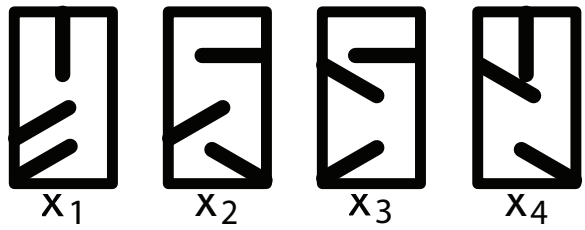
Chapter 4

Learning features with identical instantiations

In the last chapter, we established a computational framework for understanding how people infer feature representations. In this chapter, we explore how the initial model learns features depending on how the parts, which compose objects, are distributed over the observed set of objects. First, we use the model to re-analyze a previous perceptual learning experiment, where people learned unitized features after repeated visual search, in terms of how the parts composing objects are distributed over the objects. Based on these results, we next illustrate how the model predicts that people will infer different representations depending on how the parts composing objects are distributed over the observed objects. Then, we test and find robust support for the model predictions in four behavioral experiments, two of which use 2-D binary images, one of which uses 3-D rendered grayscale images, and one of which uses conceptual stimuli (animals). Finally, we consider and find evidence that the behavioral results cannot be explained using two classic dimensionality reduction techniques: Principal Component Analysis and Independent Component Analysis.

4.1 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 the initial model that we outlined in Chapter 3. As we discussed in Chapter 2, 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, after learning that the parts generating the observed objects covary in particular ways, participants represented each object as its own feature instead of as



x_1	1	1	1	0	0	0
x_2	0	1	0	1	0	1
x_3	0	0	1	1	1	0
x_4	1	0	0	0	1	1

Figure 4.1: Inferring representations for objects. Stimuli and feature ownership matrix from Shiffrin and Lightfoot (1997).

three separate features [Shiffrin and Lightfoot, 1997]. 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, we consider the possibility that they are ultimately traceable to the statistical structure of the stimulus context: i.e., how parts co-vary over the set of observed objects. Because the 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 the occurrence of differentiation and unitization 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 model.

In Experiment 1 of Shiffrin and Lightfoot (1997), participants were trained to find one of the objects shown in Figure 4.1 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 Figure 4.1, with $z_{ik} = 1$ if object i has feature k. After prolonged practice, human performance drastically improved, but 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 single-segment parts 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 (i.e., 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 model provides an answer: namely, 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 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 [Shiffrin and Lightfoot, 1997]. Figure 4.2 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 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 A 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 general statistical cues, people use perceptual constraints to infer features, such as proximity information [Goldstone, 2000], which are not included in this version of our model. In Chapter 6, 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.

4.2 Wholes and parts: Inferring features using distributional information

Our model predicts that 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. Figure 4.3a shows the base (on left) and

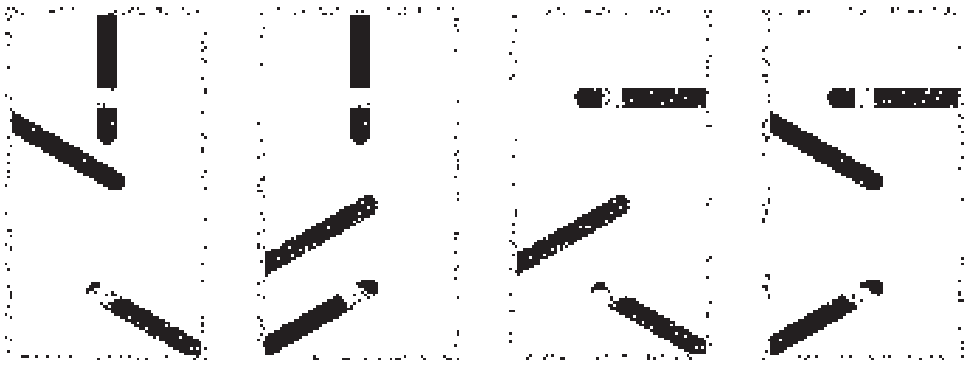


Figure 4.2: Inferring feature representations using distributional information from Shiffrin and Lightfoot (1997). The four objects are learned as features; however, the features inferred by the IBP with the independent feature image prior contain discontinuities where the objects overlap. (See Chapter 6 for how to include perceptual constraints into the model, which allow the model to infer more psychologically valid features). The model justifies the human perceptual system's unitization of the objects as features

the set of six parts used in the simulations. Figure 4.3b is an artificially generated set of observed objects for which there is not enough statistical evidence to warrant differentiation. The feature ownership 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 (remember that the rows and columns of a feature ownership matrix correspond to objects and features, respectively). 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.

Figure 4.3c 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.[1] 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).

[1]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.

We applied the model to the *correlated* and *independent* object sets, using a noisy-OR likelihood (see Appendix A). Figures 4.3d and 4.3e 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 co-variation information contained in the context of the other objects presented with the four original objects. This suggests that co-variation 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 this chapter, we examine whether this prediction holds for human feature learning.

4.3 Experiment 1: Feature learning from binary images

The simulations in the previous section lead to the question of whether people use co-variation information to infer the features of objects, as our 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.

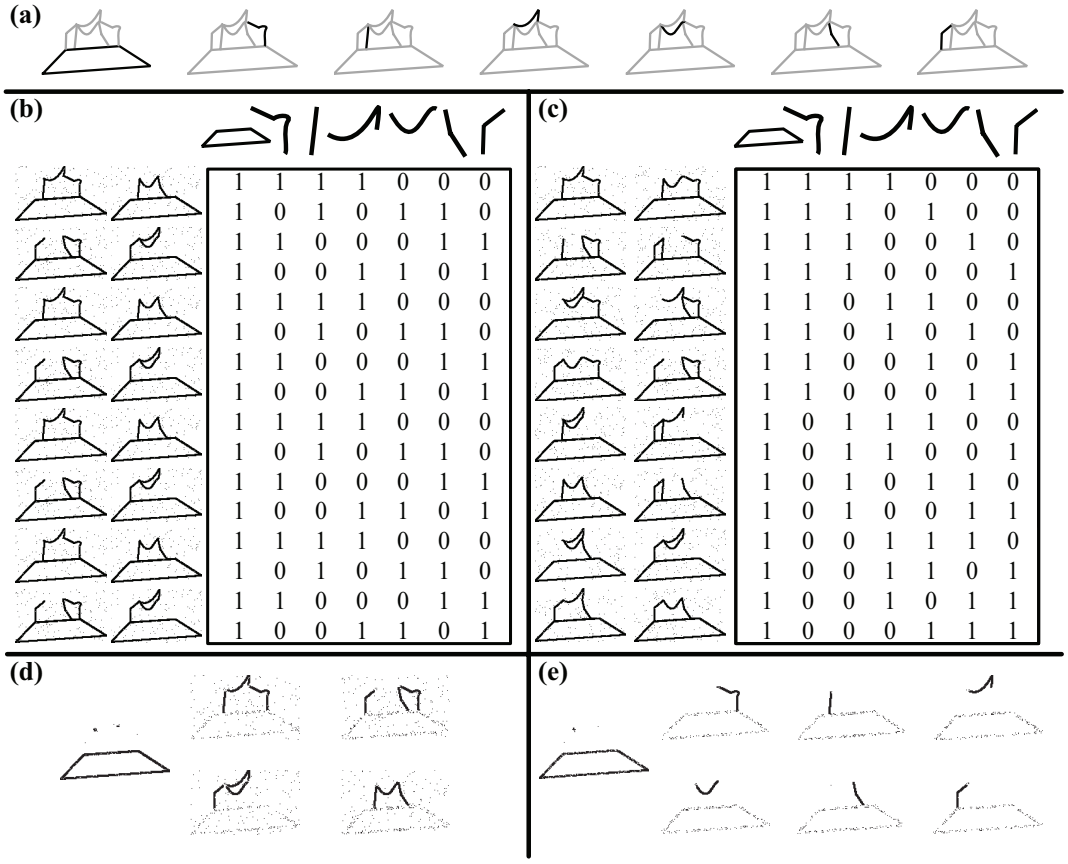


Figure 4.3: Inferring different feature representations depending on the distributional information. (a) The base (on left) and the six features used to generate both object sets. (b) - (c) The feature membership matrices and objects for the (b) *correlated* and (c) *independent* sets respectively. (d) - (e) The feature representations inferred by model for the (d) *correlated* and (e) *independent* sets, respectively.

4.3.1 Methods

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.

Stimuli

Figure 4.3a 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 perfectly co-vary. 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 with perfect certainty. The parts of the objects in the *correlated* sets had the same amount of correlation as those in Shiffrin and Lightfoot [Shiffrin and Lightfoot, 1997]. The correlated stimuli are shown in Figure 4.4a. 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 Figure 4.4b. 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 the each part in the *correlated* set and the *independent* set appeared with the same frequency, allowing us to control for familiarity 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*). Figure 4.5a-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.

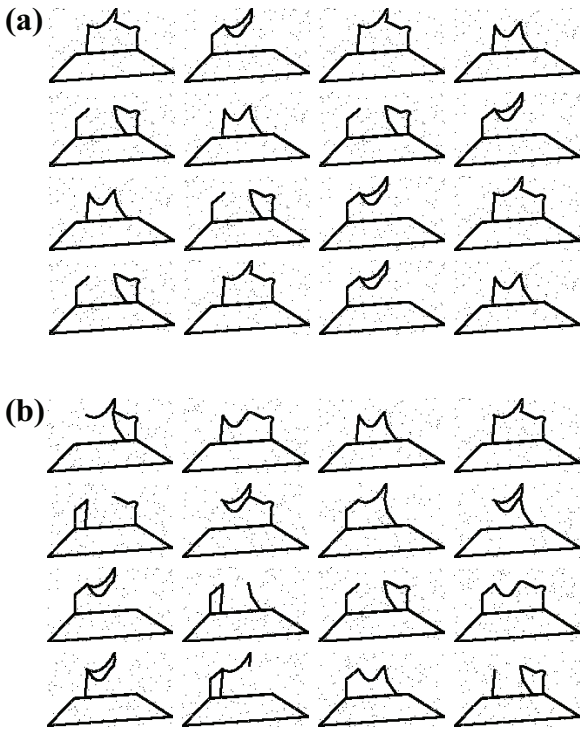


Figure 4.4: Object sets used for Experiment 1. (a) The *correlated* training set, consisting of four copies of four images of different objects. (b) The *independent* training set. These two sets share four objects. The four objects missing from the *independent* training set forms the *unseen* objects used in testing.

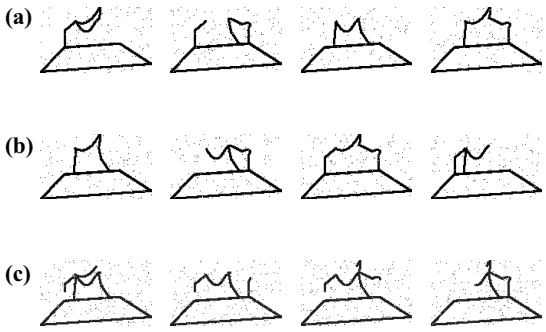


Figure 4.5: The three sets of test images. (a) *Seen* for Experiment 1 and *unseen* for Experiment 2. (b) *Unseen* for Experiment 1 and *seen* for Experiment 2. The *seen* and *unseen* objects are switched for Experiments 1 and 2, with the *unseen* objects for Experiment 1 being the objects that appear in the *correlated* set of Experiment 2. (c) *Shuffled parts* for Experiments 1 and 2.

Procedure

Participants were given the 16 images appropriate to their conditions on business cards (width 3.5 inches by height 2.5 inches) 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 minutes 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 set ("Rate from 0-10 how likely you believe the rover is to see this image on another part of the cave wall").

4.3.2 Results and Discussion

Figure 4.3.2 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 ($F(2, 48) = 40.72, p < 0.001$)[2]. 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 $F < 2$). Because there were no major effects of *test order*, we collapsed over this variable in the subsequent pre-planned analyses.

Supporting 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 ($t(13) = 8.3162, p < 0.001$). 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 Figure 4.3.2b 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 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 Appendix A

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

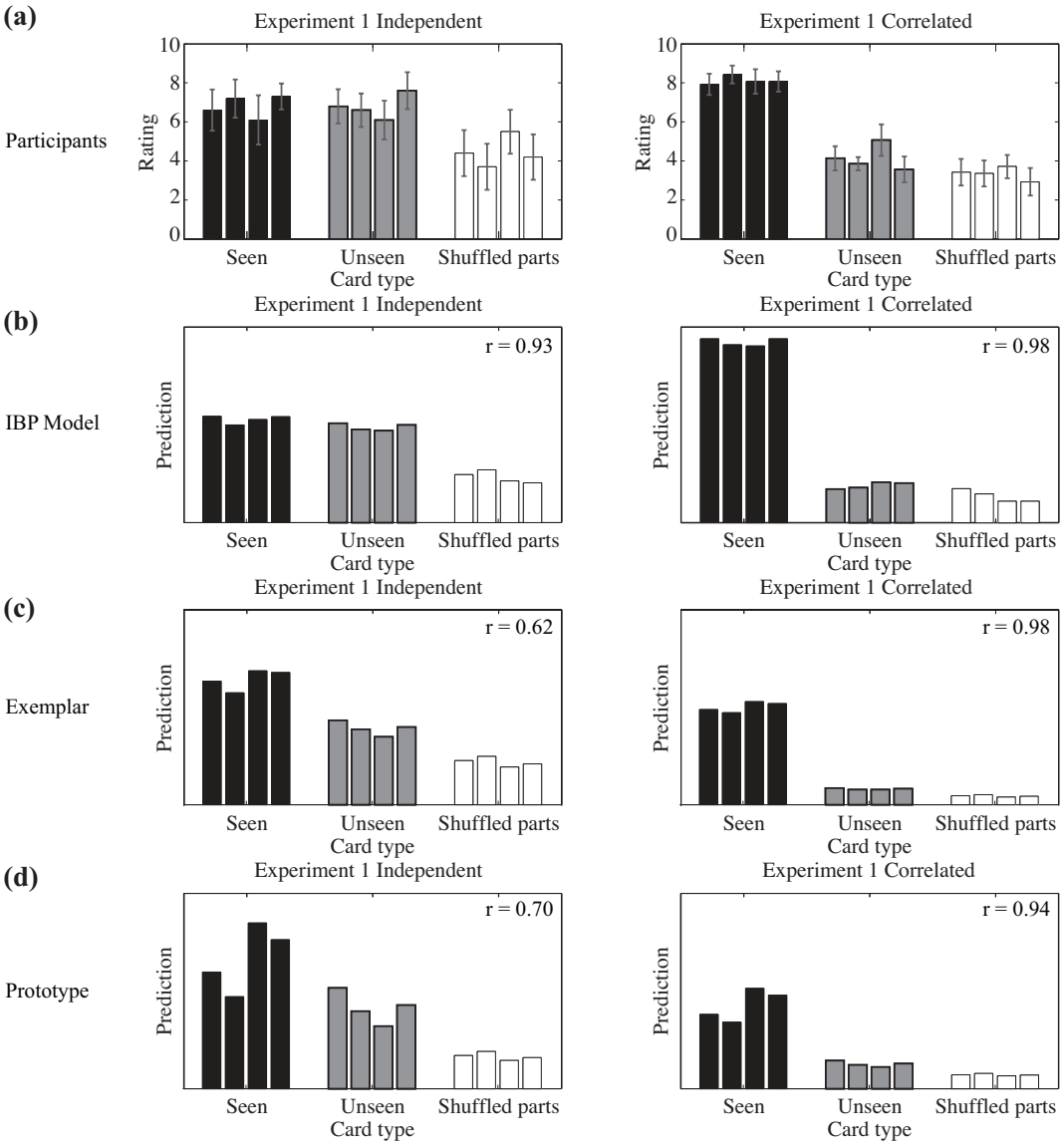


Figure 4.6: Results of Experiment 1 for *test images* in left to right order given by Figure 4.5. (a) The mean ratings participants made for test items as a function of training condition. Error bars show one standard error. (b) Predictions made by the IBP model as a function of training set. Notice 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.

for details). Because 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 Appendix A 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 of 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.[3] For the prototype model, we predicted generalization based on the distance to the average of the seen objects (as in Reed, 1972).

Figure 4.3.2cd 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.[4] 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

[3]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.

[4]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.

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 judgments made by the human participants in the two conditions varying in co-variation information. The exemplar model 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.

4.4 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.

4.4.1 Methods

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.

Stimuli

Like Experiment 1, Figure 4.3a 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 Figure 4.7ab. As the *correlated* set of this experiment is the same as the *unseen* test set used in Experiment 1, Figures 4.5bac form the *seen*, *unseen*, and *shuffled parts* test sets for Experiment 2, respectively. Otherwise the stimuli were identical to Experiment 1.

Procedure

The procedure was identical to Experiment 1.

4.4.2 Results and Discussion

As shown in Figure 4.8, 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 ($F(2, 48) = 21.381, p < 0.001$). Importantly, participants in each *distribution type* group rated the *test types* differently as is shown by a two-way interaction ($F(2, 48) = 4.136, p < 0.05$). 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* ($F(2, 48) = 4.251, p < 0.05$). However, the latter effect is irrelevant to the question of whether people use distributional information because 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). As there were no theoretically substantive effects of *test order*, we collapsed over this condition in the subsequent pre-planned analyses.

Supporting 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 ($t(13) = 5.2893, p < 0.001$). Participants in the *correlated* condition were more likely to generalize to the *seen* images than those in the *independent* condition ($t(26) = 2.1617, p < 0.05$). 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 is evident in Figure 4.8a, 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 – and so, it may be an exemplar effect).

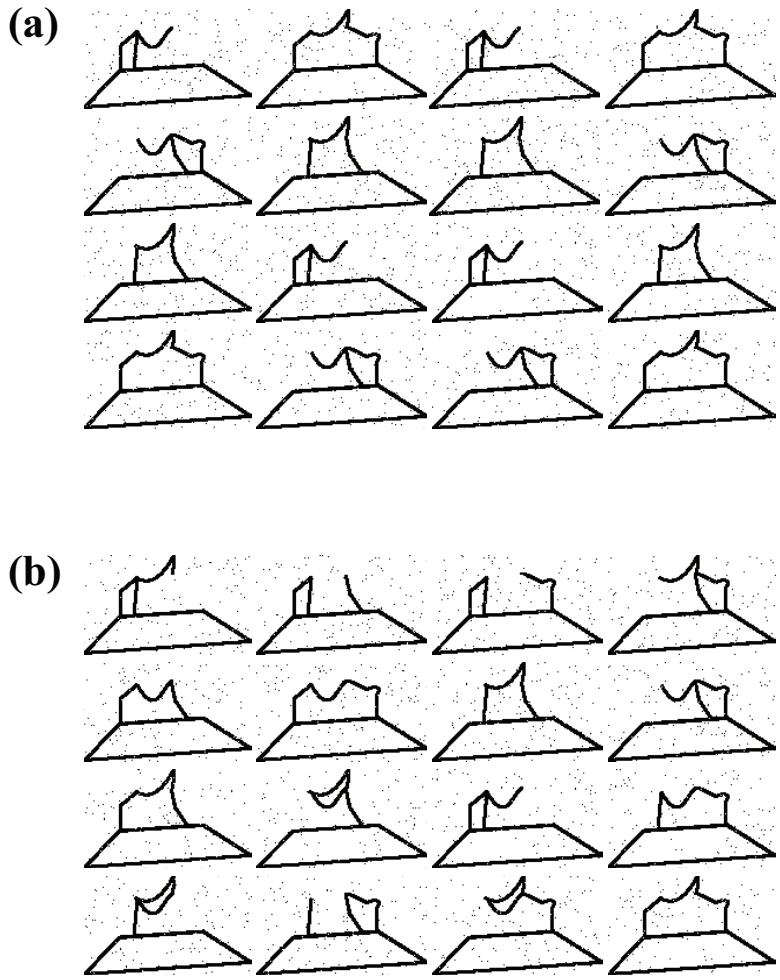


Figure 4.7: Object sets used for Experiment 2. (a) The *correlated* training set, consisting of sixteen images of four objects, each with independent noise. (b) The *independent* training set. These two sets share four objects. The four objects missing from the *independent* training set forms the *unseen* objects used in testing.

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).

Figure 4.8b 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).

Figures 4.8cd 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 fail to 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).

4.5 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 effects 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 model, which can work with grayscale images as input representations.

4.5.1 Methods

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 and 50 participants in the *independent* condition.

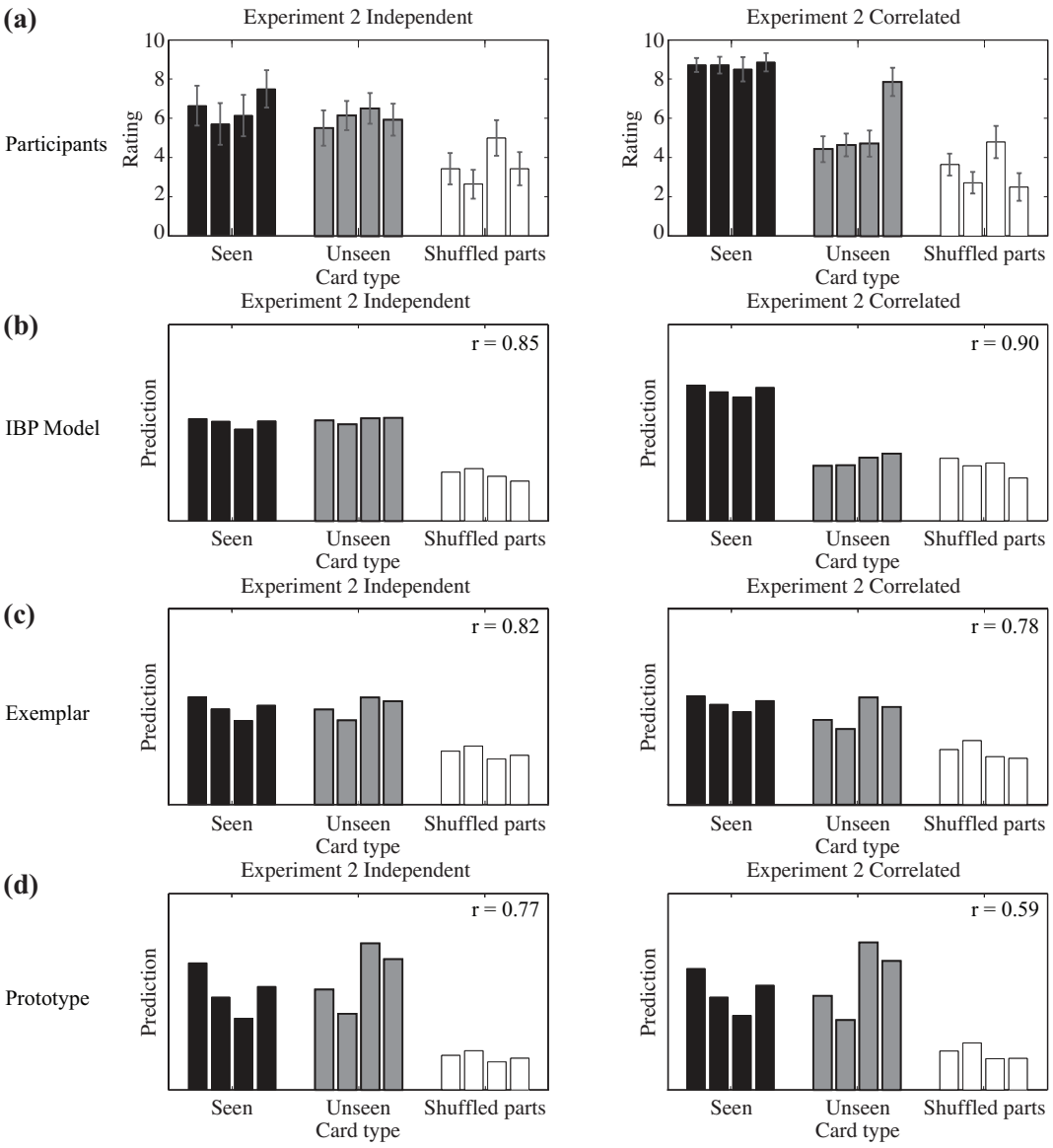


Figure 4.8: Experiment 2 results for *test images* in left to right order given by Figure 4.5. (a) The mean ratings participants made for test items as a function of training condition. Error bars show one standard error. (b) Predictions made by the IBP model as a function of training set. Notice the 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.

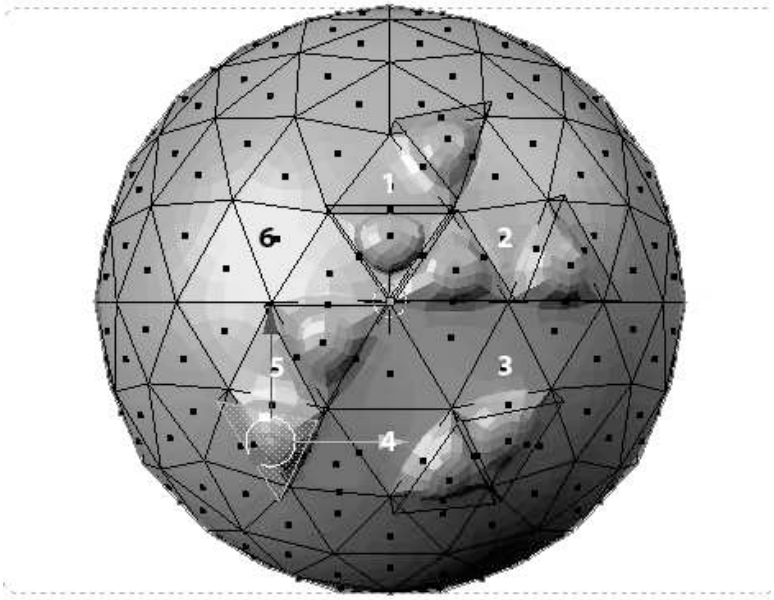


Figure 4.9: Images used as stimuli in Experiment 3. Each number corresponds to the location of a part – a small extrusion on the surface of the sphere. Each object has three parts, which were assigned in the same manner as Experiments 1 and 2.

Figure 4.10: 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.

Stimuli

Figure 4.9 shows the objects used as stimuli, and the location of the six parts used in the experiment. Each part is an extrusion on the surface of the sphere (motivated by the stimuli used in Schyns and 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/). Figure 4.10 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 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.

Procedure

Participants were given the sixteen images appropriate to their conditions on cards (width 4.25 inches by height 5.5 inches) randomly in front of them and 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 minutes 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").

4.5.2 Results and Discussion

Figure 4.11a 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 ($t(46) = 3.55, p < 0.001$). Figure 4.11b gives the 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 because it rules out a potential alternate explanation that participants in the *independent* condition are simply rating everything higher.

4.6 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

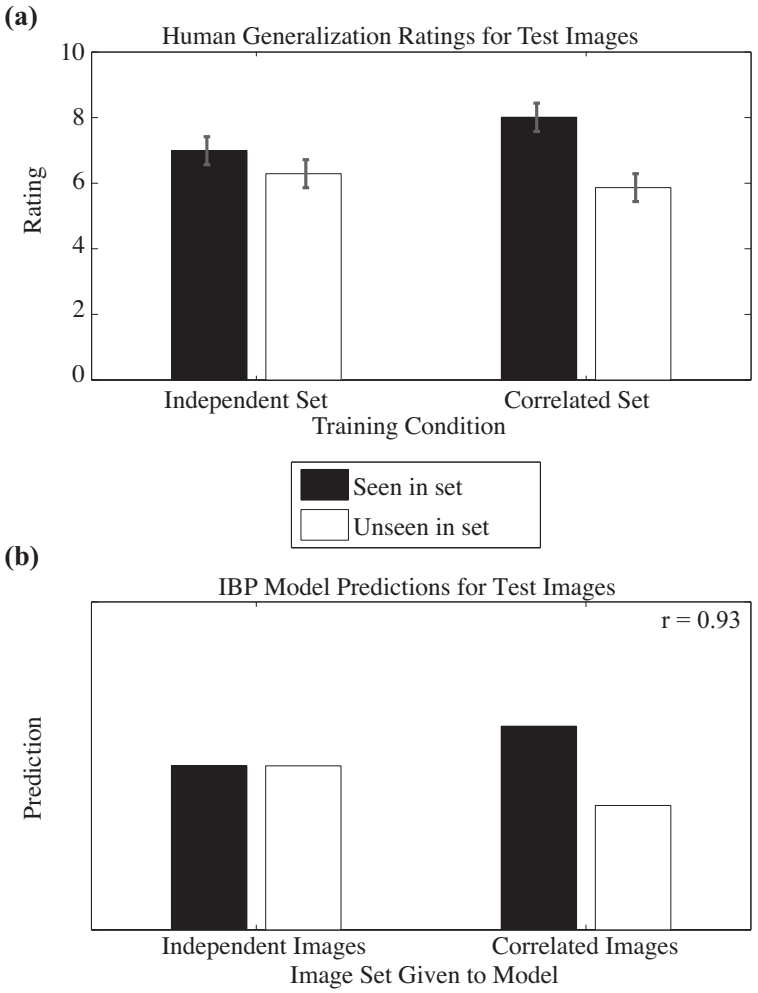


Figure 4.11: Results of Experiment 3. (a) The mean ratings participants made for test items as a function of training condition. Error bars show one standard error. (b) Predictions made by the model as a function of training set. Notice the close qualitative correspondence to human performance. Replicating 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.

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.

4.6.1 Methods

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.

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.

Procedure

Participants were given the sixteen cards appropriate to their condition (width 3.5 inches by 2 inches) 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 minutes 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").

4.6.2 Results and Discussion

Figure 4.12a 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 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 therefore promotes 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.

Figure 4.12b 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).

4.7 Comparison with machine learning methods

We can compare the features learned by the two dimensionality reduction techniques that we mentioned in Chapter 2, Principal Component Analysis (PCA) and Independent Component

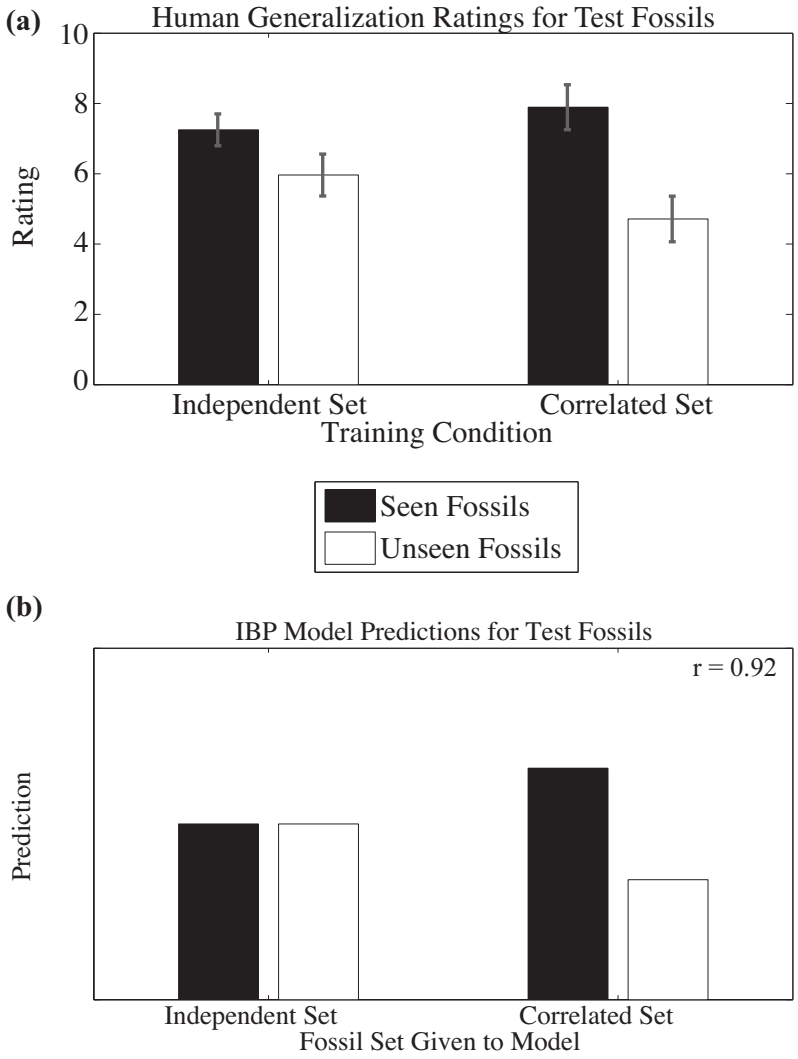


Figure 4.12: Results of Experiment 4. (a) The mean ratings participants made for test items as a function of training condition. Error bars show one standard error. (b) Predictions made by the model as a function of training set. Notice the close qualitative correspondence to human performance. Although participants in both conditions differentiate between the *seen* and *unseen* objects, participants in the *correlated* condition differentiate more than those in the *independent* condition.

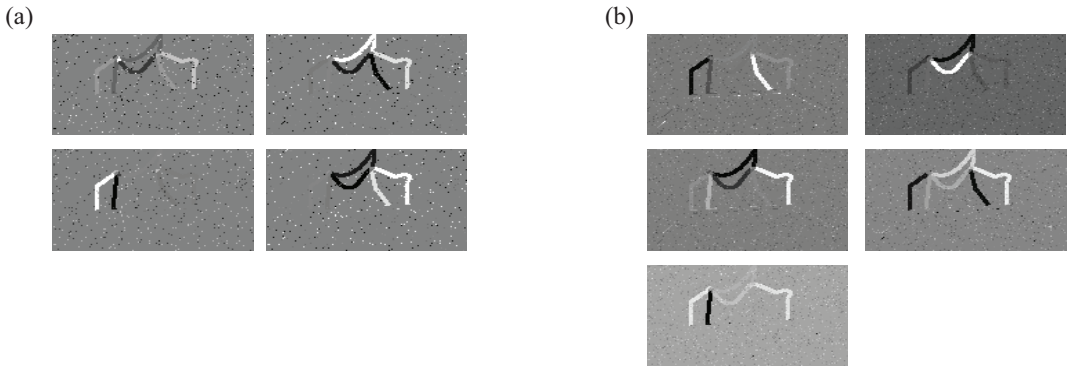


Figure 4.13: Features inferred by Principal Component Analysis (PCA) when given objects whose parts either co-vary or vary independently. The features are presented as a color map, where negative weights are colored black and positive weights are colored gray. On left, features inferred by PCA given 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.

Analysis (ICA), with the features learned by our 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.[5] 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).

Figure 4.13 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 Figure 4.7) 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 Figure 4.13a 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 Figure 4.13b). 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. Figures 4.15ab show the best-fitting predictions from PCA given each object set found in the same manner as for the other models

[5]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.

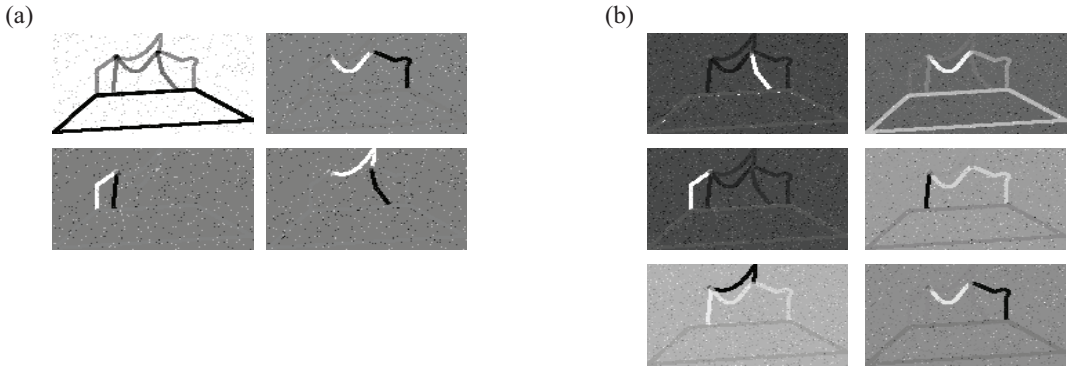


Figure 4.14: Features inferred by Independent Component Analysis (ICA) when given objects whose parts either co-vary or vary independently. The features are presented as a color map, where negative weights are colored black and positive weights are colored gray. On left, features inferred by ICA given 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.

($\gamma = 0.3082$). 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 (by about 0.1) for the *independent* condition of Experiment 1 and *correlated* condition of Experiment 2.

Figure 4.14 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 Figure 4.14a) 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). Figures 4.15cd show the best-fitting predictions from ICA given each object set found in the same manner as for the

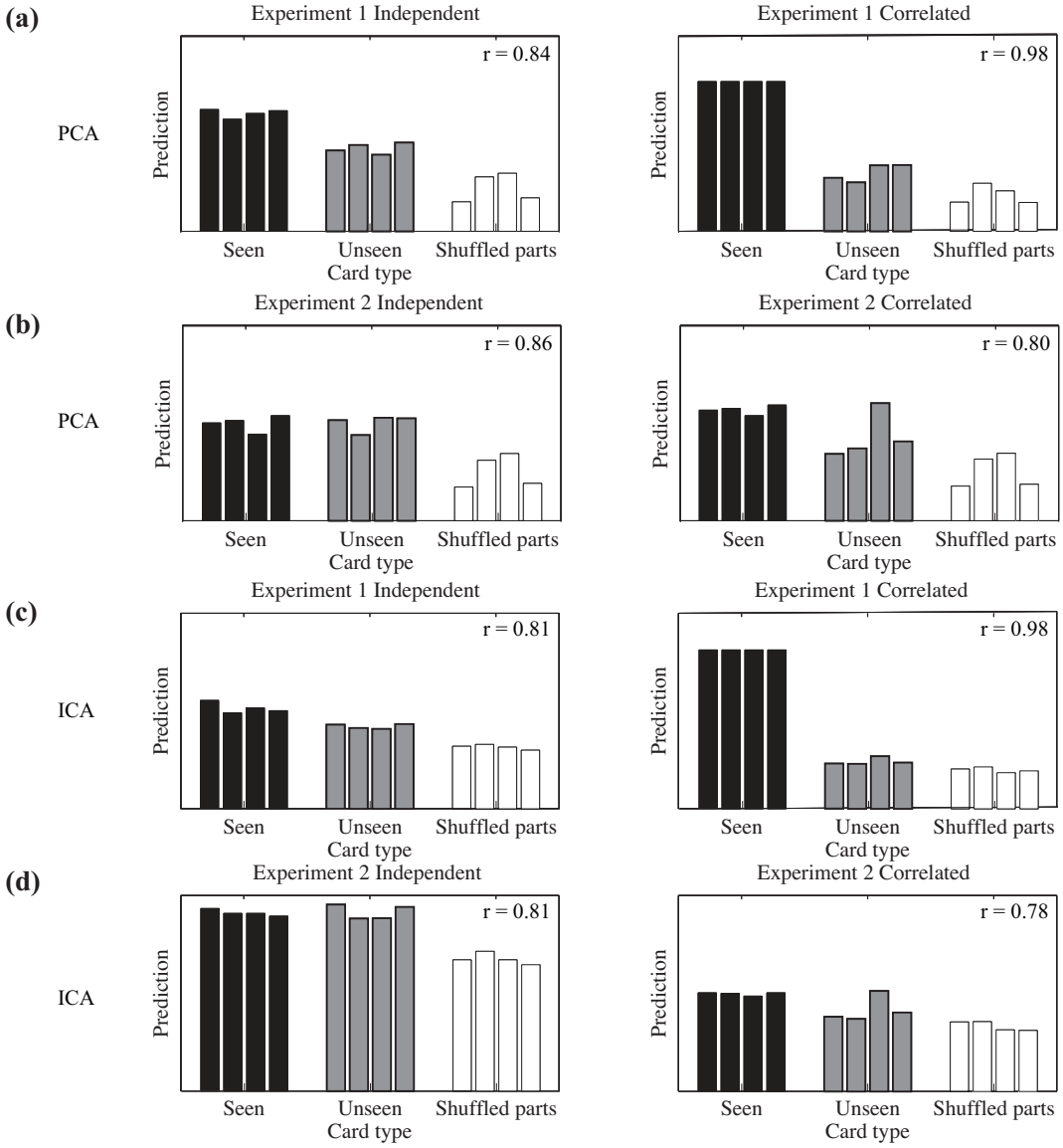


Figure 4.15: The best-fitting predictions of Principal Component Analysis (PCA) (a) and (b) to human results of Experiments 1 and 2 respectively and the best-fitting predictions of Independent Component Analysis (ICA) (c) and (d) to human results of Experiments 1 and 2. PCA predicts a weaker effect of *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.

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) described a method for visual feature learning that infers features by picking those that provide the information about 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, because 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.

Chapter 5

Learning features that transform

In this chapter, we address how to learn transformation-invariant features (Criterion 5). In our environment, stimuli do not occur identically every time they appear in raw sensory data. For example, imagine that on your desk, next to your computer, is your coffee mug. The position of the image of the coffee mug on your retina varies whenever your eye moves. Or, if you move your head towards the coffee mug while fixating on it, the coffee mug's image on your retina changes as well. Although your retinal image changes in each of these cases, objects and their properties in the environment do not change (they are invariant).

Fortunately the retinal image varies according to predictable rules or *transformations*. When your eye moves to the left, the retinal image before the saccade is the same as the retinal image after the saccade after each point is shifted to the right by some amount (see Figure 5.1 for a possible set of psychologically relevant transformations). Additionally, imagine that you move closer to the coffee mug. When you move closer, the mug's image scales to be larger, but otherwise it is the same. In both of these cases, given the original image and knowledge of the transformation, the resulting image is perfectly predictable. Thus, we can tell objects are equivalent when we rotate our heads even though the retinal image has been rotated [Rock, 1973] or when we navigate the world and the image of objects is translated and scaled in different ways [Palmer, 1983].

5.1 Extending the model to include transformations

Although we showed in Chapter 4 that the initial model is very successful at explaining human feature learning when those features occur identically on every presentation, features frequently occur differently in our sensory inputs across presentations. This presents a challenge for the usefulness of the initial model because it cannot learn features that occur differently across appearances. Currently, every transformed occurrence of a feature would be treated as a distinct and unique feature even though they are really the same feature after taking into account a simple transformation. Thus, the previously discussed model is

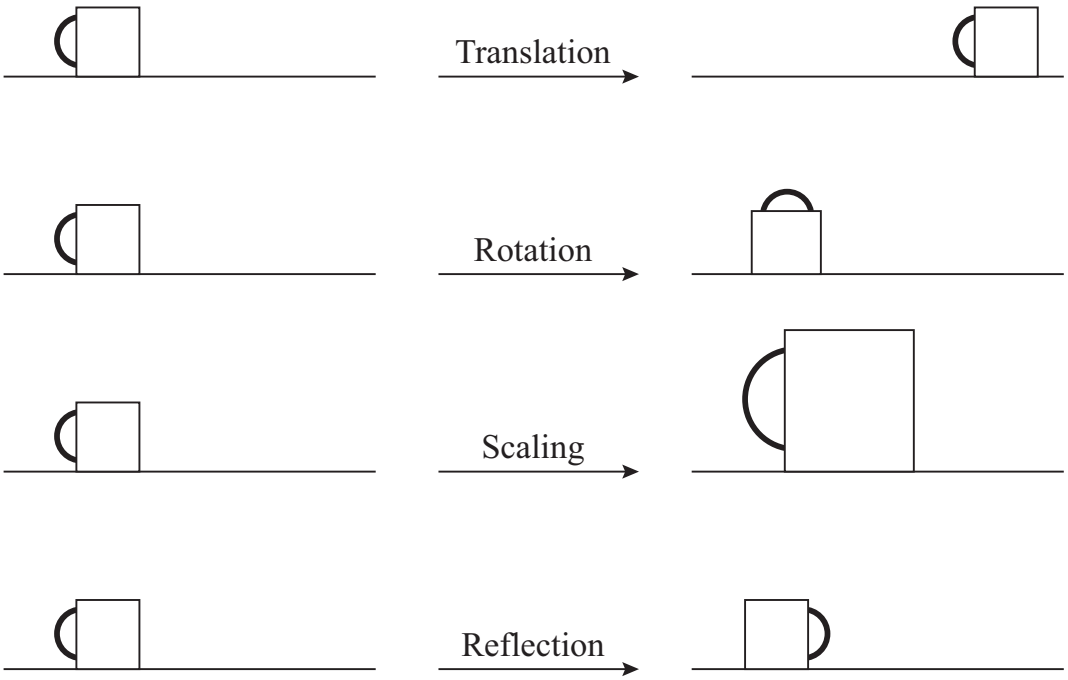


Figure 5.1: A possible set of transformations used by the human perceptual system (adapted from Palmer, 1983).

not sufficient to explain human feature learning.

The images of objects and features change, but in predictable ways. After an eye movement, the new retinal image is not completely different, but a simple transformation (a translation) from the previous retinal image. Therefore, if we include in the computational framework the possibility of features being transformed, we could learn features that are invariant to transformations in appearance. This can be done by adding an extra step to the culinary metaphor for the IBP discussed above.

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

Figure 5.2 illustrates an example feature representation generated by the extended culi-

nary metaphor, where we have defined the set of possible transformations to be right translations.[1] As before, Figures 5.2a-c show how customers enter the restaurant drawing dishes, resulting in a corresponding feature ownership matrix, and associate feature images with each dish. Unlike the initial model, the extended model draws a transformation for every object and feature (even if the object does not use the feature) as shown in Figure 5.2d. This results in the reconstructed objects shown in Figure 5.2e.

This new process is called the transformed IBP (tIBP; see the Appendix C for more technical details). Although the model incorporates the set of *a priori* possible transformations, what the features are and how the features are transformed in each object are not known *a priori*. In other words, the model learns both a set of features and for each object how the learned features are transformed. By including the different transformations hypothesized by psychologists (like those in Figure 5.1), the tIBP can overcome the limitations of the IBP and, like people, learn features that are invariant over transformations typical of our environment.

By including some of the transformation types in Figure 5.1, the tIBP is able to learn transformation-invariant features, and predict novel contextual effects. First, we demonstrate the difference in features learned by the IBP and the tIBP in a classic machine learning task: learning features from images containing horizontal and vertical bars in random locations [Ghahramani, 1995, Rumelhart and Zipser, 1985, Spratling, 2006]. Next, we model previous work showing that people learn spatially invariant features [Fiser and Aslin, 2001] using the tIBP. Afterwards, we test behaviorally a novel context effect that is predicted by our model.

5.2 Models learning invariant vs. "variant" features

A classic problem in machine learning is learning features from images composed of a horizontal and vertical line in random locations [Ghahramani, 1995, Rumelhart and Zipser, 1985, Spratling, 2006]. For example, Ghahramani (1995) applied his model, which reconstructs a separate feature for the horizontal and vertical line in each position. Though this is a good solution, it is not ideal. Ideally, a horizonal and vertical line should be inferred as features. With these features, the model will be able to generalize to images containing horizontal and vertical lines in novel locations. Figures 5.3ab show the feature representations inferred by the tIBP and IBP, respectively, when given 100 images composed of a vertical and horizontal bar each occurring in the image with probability 0.8 in a position drawn uniformly at random over the image. Not surprisingly (since the IBP assumes features occur identically in every image), the IBP model infers a feature for different positions of the lines (similar to the solution of Ghahramani, 1995) and some of the objects themselves. The tIBP model infers a horizontal and vertical bar that occur in different positions – the ideal solution.

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

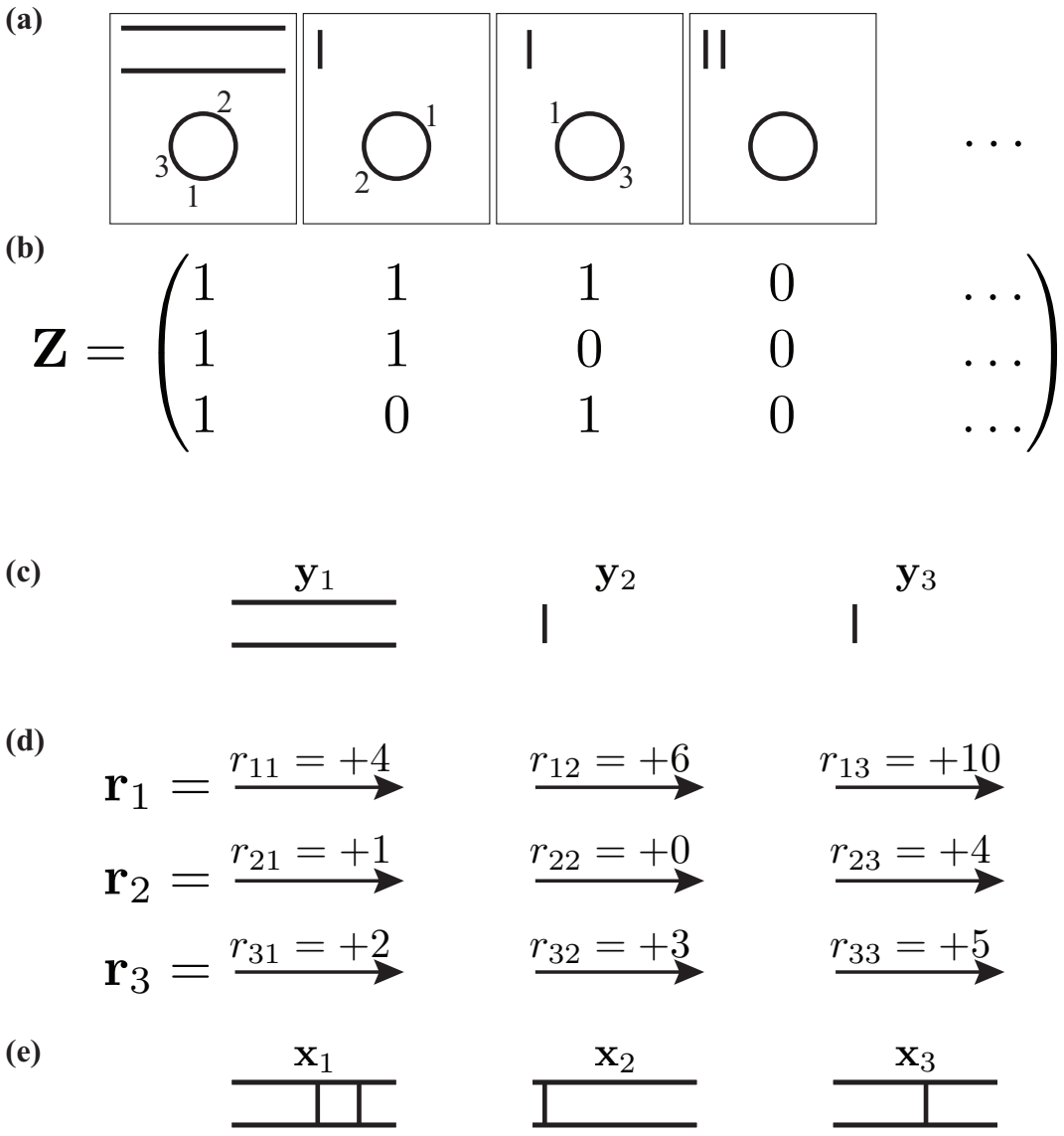


Figure 5.2: An illustration of the relation between the culinary metaphor, the transformed Indian buffet process, and the model. Note that the only difference between this and the illustration of the Indian buffet process is the "spice" (transformation) that each customer (object) draws each time she takes a dish (feature). (a) The culinary metaphor for the transformed Indian buffet process. The numbers are customers and the circles are dishes. A number adjacent to a table represents its corresponding object taking that feature. A feature image is generated for each dish, which appears above the circle for each dish. (b) The equivalent feature ownership matrix represented by the culinary metaphor above. (c) The feature images generated from the feature image prior for each feature. (d) The transformations drawn (uniformly over a set of translations to the right) for each customer for each dish. The number corresponds to how many pixels the feature image is shifted to the right when it appears in a reconstructed object. (e) The reconstructed objects using $\mathbf{Z}$, $\mathbf{Y}$, and $\mathbf{R} = [\mathbf{r}_1^T, \mathbf{r}_2^T, \mathbf{r}_3^T]$ defined in (b) and (c) respectively.

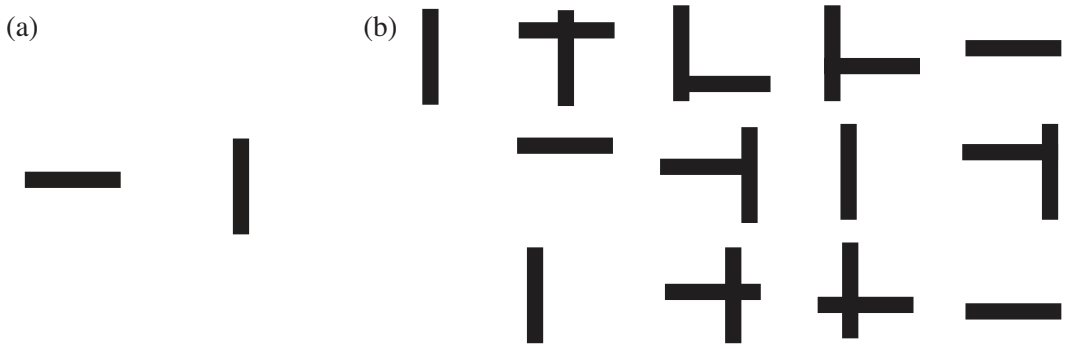


Figure 5.3: Comparing the features learned by two models defined from our mathematical framework. (a) The features learned by the tIBP and (b) IBP given images containing a vertical and horizontal bar that randomly occur in random positions. Note that the tIBP is able to learn a vertical and horizontal bar due to the set of pre-defined transformations it is given *a priori*, but the IBP must create a different feature each time it observes a bar in a novel position.

5.3 Learning spatially invariant features

People form representations of objects, even when the object features occur in locations that were never previously observed. To see if statistical learning could be a potential mechanism for how people learn higher-order visual representations, Fiser and Aslin (2001) investigated whether people could learn the basic units of scenes from merely observing them. In their first two experiments, they showed participants scenes that were composed of three "base pairs" out of a set of six possible base pairs, where a base pair is a pair of images that always co-occurred in a particular spatial arrangement (see Figure 5.4a for an example set of six base pairs, and Figure 5.4b for an example scene). After observing 144 scenes, participants judged base pairs in novel locations (a never observed image) as more familiar than a pair of parts that "accidently" occurred together (but were not a base pair). Based on similar results in a series of experiments, Fiser and Aslin (2001) suggested that participants learned the base pairs as translation-invariant units (or features) from passively viewing scenes.

We now illustrate that the tIBP can infer translation-invariant features comprised of complex parts, as participants did in Fiser and Aslin (2001). We did not use the exact images as Fiser and Aslin (2001) because they were too large (1200 by 900 pixels), and simply reducing their resolution to a tractable size rendered the parts indistinguishable. Therefore, we preserved the same statistical structure by recoding each part to be 3 by 3 pixels, and then, formed the same configurations that they used in their images. This resulted in scenes that were 9 by 9 pixel. Figure 5.4a shows the parts grouped into their base pairs, Figure 5.4b shows an example scene comprised of these parts, one of the 144 scenes that the model observed, and Figure 5.4 (c) shows the features the tIBP inferred given the 144 scenes (see Appendix C for details). Not only does the model reconstruct the base pairs

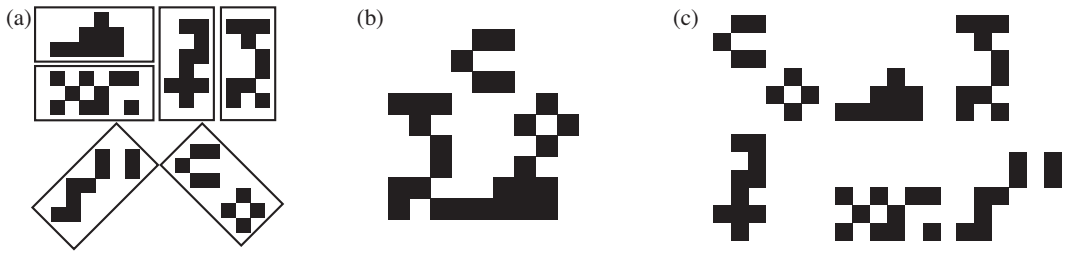


Figure 5.4: Learning spatially invariant features. (a) The twelve parts that were grouped into base pairs, with each base pairs in a rectangle. (b) An example scene. (c) The features inferred by the tIBP model. The tIBP infers the base pairs as features.

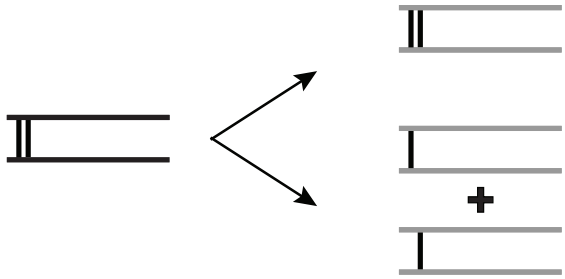


Figure 5.5: Does the image have one feature containing two vertical bars or two features each of which is a vertical bar? Allowing transformations introduces new ambiguities for learning features.

that were used to generate the images, but it also learns that the base pairs can occur in any location. We next compared the model's solution with people's familiarity judgments from Fiser and Aslin (2001) and found the same pattern. After observing the 144 scenes, the probability of observing a base pair in a new location (a novel scene) was larger than observing two parts that occurred in the same location together in a previously observed scene. This was true for all six base pairs.

5.4 Experiment 1: One or two features?

Having transformable features raises an interesting new problem: If components of an image can be mapped on to one another using one or more transformations, when should the components be perceived as different features, as opposed to different instantiations of the same feature? This problem is illustrated in Figure 5.5, where an object containing two vertical bars can be represented as (a) having a single feature comprised of two vertical bars (the *unitized* feature), or (b) having two features, each comprised of a single vertical bar with their own translation (the *separate* features). An intuitive heuristic for solving

this problem is to choose the feature representation with the fewest number of features that can adequately encode the observed objects. Figure 5.6 shows how this same object can be represented with different features, depending on the structure of the other objects in the set. In the unitized set (Figure 5.6a), all of the objects can be represented by the single, two-bar feature, whereas the objects in the separate set (Figure 5.6b) require at least two, single-bar features that translate independently. Although the objects in Figure 5.6a can be represented with two single-bar features as in Figure 5.6b, that representation seems less good because it requires more features, and it would be a surprising coincidence that the two vertical bars are always the same distance if they varied independently.

The other objects expected to be in a set of objects depends on how those objects are represented. When the objects are represented by the *unitized* feature (two vertical bars), new objects with two vertical bars that same distance apart are expected. These objects are also expected, when the *separate* features are used. However, any object with two vertical bars (*New Separate*) can only be represented using the *separate* feature representation. Thus, if participants generalize object set membership to the *New Separate*, we can infer that they represent the objects using the *separate* feature representation (and if they do not, they are using the *unitized* feature representation).

5.4.1 Methods

Participants and Stimuli

We recruited 40 participants through Amazon Mechanical Turk, which were randomly assigned to observe either the *unitized* (Figure 5.6a) or *separate* object set (Figure 5.6b). Unfortunately, three participants did not complete the task, and so there were 18 participants in the *unitized* condition and 19 participants in the *separate* conditions. The test images were organized into the following image types: *Seen Both* (an image in both the *unitized* and *separate* sets), *Seen Unit* (an image only in the *unitized* set), *Seen Sep* (an image only in the *separate* set), *New Unit* (an image that can be represented using either the *unitized* or *separate* feature sets), and four control images, which are shown in Figure 5.7.

Procedure

Participants were told that they were given images that the Mars rover took while exploring a Martian cave (an adapted version of the cover story used in Experiments 1 and 2 of Chapter 4). After observing the eight objects appropriate to their condition, they were asked to rate, from 0 to 6, how likely the Mars rover was to encounter nine other objects (presented in a random order) in a new portion of the Martian cave (the test images).

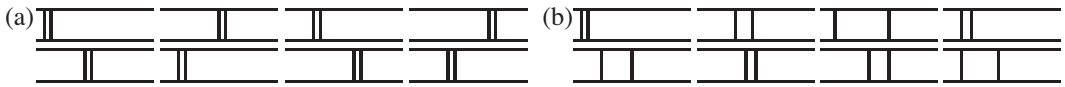


Figure 5.6: Resolving ambiguity when learning features invariant over translations by inferring the smallest feature representation that encodes the whole set of objects. (a) The *unitized* set. These objects were made by translating the unitized feature. (b) The *separate* set. These objects were made by independently translating the two separate features. The number of times each vertical bar is presented is equal over the two object sets.

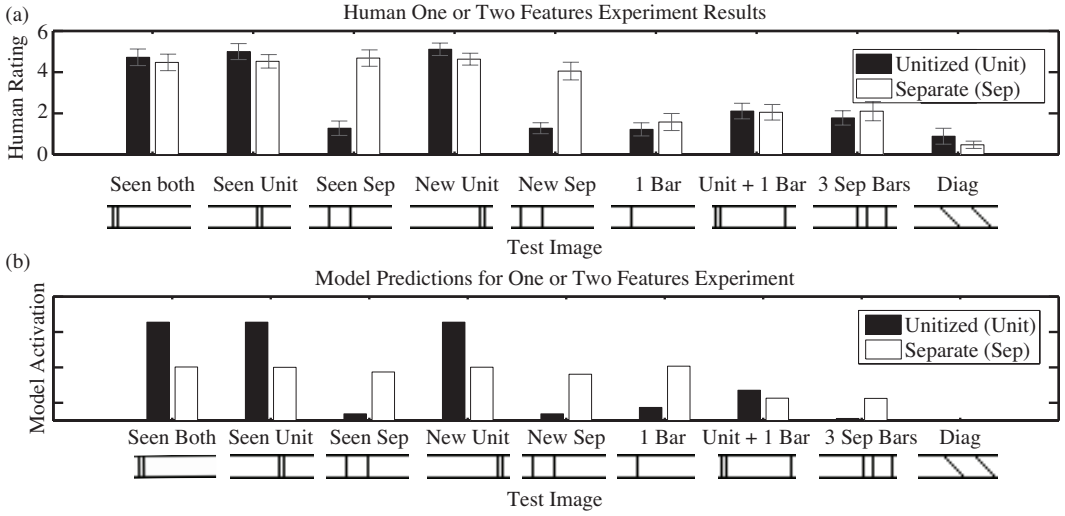


Figure 5.7: Human judgments and model results for learning features invariant over translations depending on the object set. (a) Human judgments. The *unitized* group only rated those images with two vertical bars close together highly. The *separate* group rate any image with two vertical bars highly. (b) The results of the tIBP model.

5.4.2 Results and Discussion

The average participant ratings on the test images are shown in Figure 5.7a. The *separate* group generalized to the *Seen Sep* ($t(35) = 6.40, p < 0.001$) and *New Sep* ($t(35) = 5.43, p < 0.001$) more than the *unitized* group, but otherwise, the two groups generalized in the same manner. This supports our predictions that the *separate* group will use the *separate* feature representation, whereas the *unitized* group will use the *unitized* feature representation (even though the *separate* feature representation is also consistent with the objects they observed). As both groups generalize to the appropriate new test images, they inferred translation-invariant features. The predictions made by the tIBP given either the *separate* or *unitized* images are shown in Figure 5.7b (see the Appendix C for simulation details).[2] The quantitative fit of the model is quite good, as the Spearman's rank order correlation between the human results and the model's predictions is 0.85 (using only one parameter to fit to human responses). Disregarding the *1 Bar* test image, the qualitative fit is quite good as well.[3] The model incorrectly predicts that the *separate* group of participants should generalize the most to the *1 Bar* test image, as it only has one feature and the model prefers images with fewer features. These results indicate that while inferring the appropriate representation for a set of objects, people also infer other expectations about the set of objects (e.g., the number of features per object). This is an interesting phenomenon for future research, which we discuss further in Chapter 8.

[2] A form of the tIBP that learned relations about how features in the same image are transformed might produce similar results without forming a unitized feature given the *unitized* images. This is an intriguing idea, which may be difficult to distinguish from the formation of a unitized feature.

[3] By qualitative fit, we mean the model explains how participants change their responses between test objects within a condition. What is considered the highest activation for a test objects is different between conditions. This is most likely because the *separate* representation can represent many more possible objects, and thus, each valid object gets less weight.

Chapter 6

Prior expectations in feature learning

In this chapter, we investigate how prior expectations about the types of features that should be inferred can be included into the computational framework (Criterion 4: Prior expectations), including categorization information into the framework (Criterion 6: Category diagnosticity), and how the features inferred by a model in the framework can be affected by the order that objects are presented to it (Criterion 7: Incremental learning). First, we demonstrate that a bias towards contiguous feature images can be incorporated in the framework by changing the prior on feature images. Second, we present and critique two possible ways to incorporate category information into the framework. Third, we then explore how the perceptual system assumes the transformations of features in the set of objects are dependent. Previous results by Smith (2005) suggest that people generalize the direction of one type of transformation (translation) to another type of transformation (scaling). Fourth, we demonstrate that people and the tIBP learn which types of transformations are allowed from the observed set of objects. In both cases, the model explains human behavior as a prior assumption of how feature transformations are dependent. Finally, we conclude the section by illustrating that the features inferred by the model may or may not be influenced by the order that objects are presented to it, depending on the machine learning algorithm that is used for inference.

6.1 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 and Cooper, 1991, Braunstein et al., 1989, Palmer, 1999]. One consequence of this simplifying assumption was seen in the features inferred from the stimuli

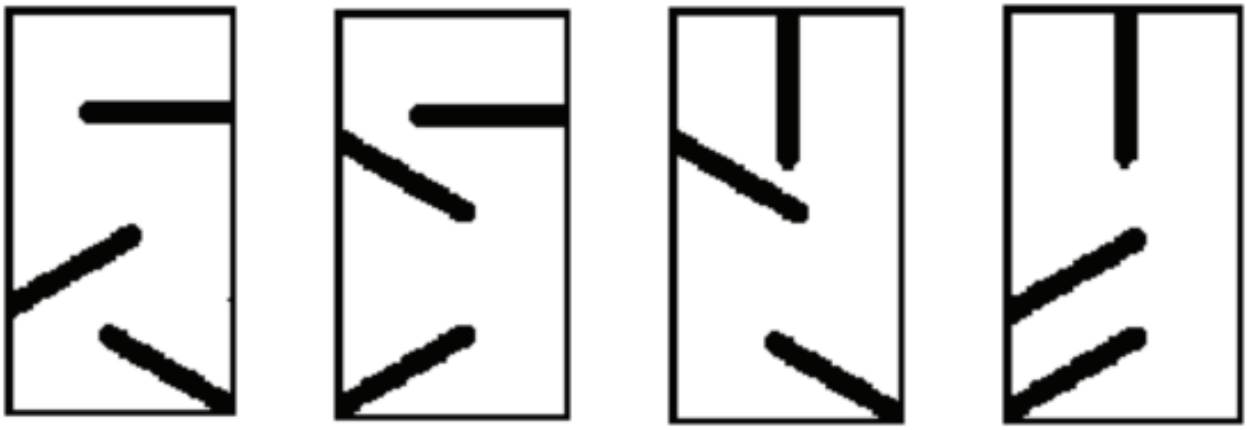


Figure 6.1: Features learned by the model given the objects from Experiment 1 of Shiffrin and Lightfoot (1997) using the proximity bias feature image prior. Note that the speckled holes have been filled in using proximity information resulting in more psychologically plausible features.

used by Shiffrin and Lightfoot (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. In Chapter 2, Equation 3.4 defined one possible method for doing this with the noisy-OR likelihood.

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 Figure 4.2. 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 Figure 6.1). 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.[1] 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. 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

[1] We thank Rob Goldstone for this suggestion.

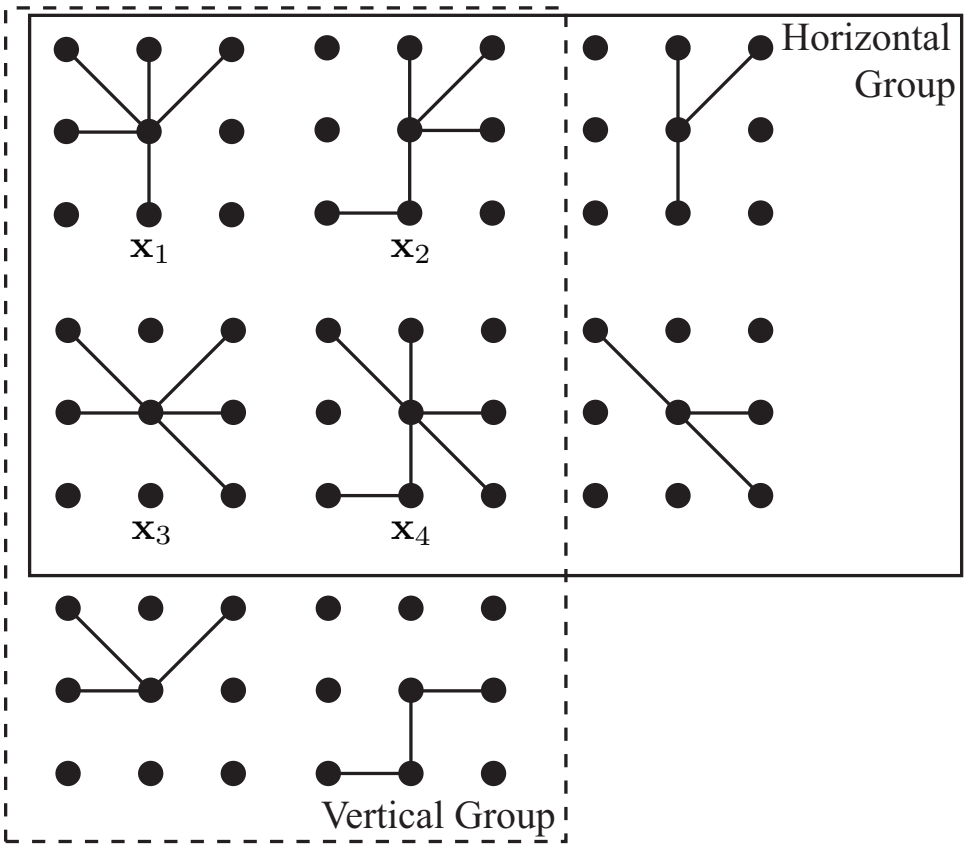


Figure 6.2: Effect of category learning on feature representations (adapted from Goldstone, 2003). Participants taught a horizontal categorization rule ($\mathbf{x}_1$ and $\mathbf{x}_2$ in one category, $\mathbf{x}_3$ and $\mathbf{x}_4$ in another) learn the features on the right; whereas those taught a vertical categorization rule ($\mathbf{x}_1$ and $\mathbf{x}_3$ in one category, $\mathbf{x}_2$ and $\mathbf{x}_4$ in another) learn the features on the bottom of the figure.

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.

6.2 Using categories to infer features

In this section we explore Criterion 6 (Category diagnosticity), which is that the categorization of an object set should be able to affect the features people use to represent the objects [Pevtzow and Goldstone, 1994, Schyns and Murphy, 1994]. In Pevtzow and Gold-

stone (1994), participants learned to categorize four objects, $\mathbf{x}_1$ to $\mathbf{x}_4$, as shown in Figure 6.2. In the *horizontal* condition, participants learned a "horizontal" categorization scheme, where objects $\mathbf{x}_1$ and $\mathbf{x}_2$ formed one category, and objects $\mathbf{x}_3$ and $\mathbf{x}_4$ formed another category. Conversely, in the *vertical* condition, participants learned a "vertical" categorization scheme, where objects $\mathbf{x}_1$ and $\mathbf{x}_3$ formed one category, and objects $\mathbf{x}_2$ and $\mathbf{x}_4$ formed another category. After participants finished the category learning phase of the experiment, they showed enhanced processing for the part that was diagnostic for the learned categorization scheme (specific to their condition), but there was no enhanced processing for the part that was not diagnostic for categorization. For example, participants who learned the "horizontal" categorization scheme showed enhanced processing for the two diagnostic parts (shown in the right of Figure 6.2), but not for the non-diagnostic parts (shown in the bottom of Figure 6.2), and vice versa for the participants in the "vertical" categorization scheme.

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

We now investigate two potential methods for incorporating category information into the feature learning framework: (1) appending the category label to the sensory information as a sort of observable property (similar to how category information is encoded by the rational categorization model; Anderson, 1990), and (2) assuming each category has its own feature representation (or IBP), but that the IBPs are coupled such that the features come from some common source (similar to how clusters are shared across categories in a hierarchical form of the rational categorization model; Griffiths et al., 2011). When the first technique is used (see Appendix E for simulation details), the model only learns features to encode the diagnostic parts for the given categorization scheme, as shown in Figures 6.3ab. Although the model successfully infers different features depending on the categorization scheme, it does so at the expense of accurately encoding each object (it represents each object with a single feature, the diagnostic part).

Another approach for learning features that incorporates categorization information is to have a separate IBP for each category that shares a common source of features (i.e., the base distribution that generates the feature images for each IBP is the same). To do so, we change the culinary metaphor of each IBP such that new features are generated by a common source. Rather than assuming the images of new features are generated independently at random, the new feature images for each category's IBP are generated from a shared Chinese restaurant process (CRP; Pitman, 2002). According to the CRP, customers (features) enter a restaurant with an infinite number of tables (each of which can hold an infinite number of

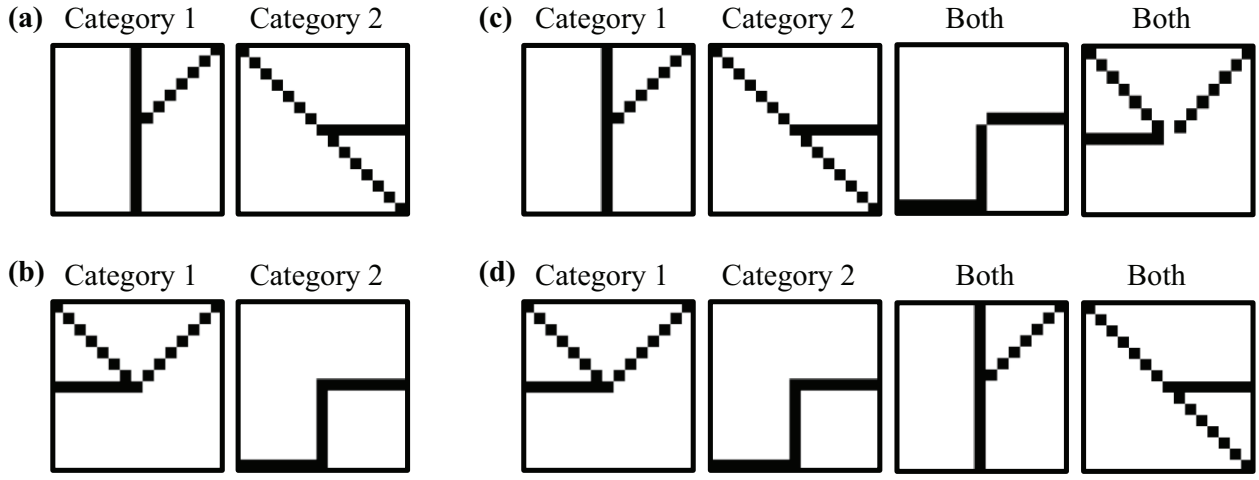


Figure 6.3: The effect of category information on the inferred feature representations given the objects from Pevtzow and Goldstone (1994), while varying the categorization scheme. (a-b) Features learned by the model that encodes category information by appending it to the sensory data using the horizontal and vertical categorization schemes, respectively. (c-d) Features learned by the extended model that has a separate IBP for each category, but couples the source of new features, using the horizontal and vertical categorization schemes, respectively.

customers) sequentially and sit at a table, which serves a single dish (feature image).[2] When the first customer enters the restaurant, all of the tables are empty. She therefore sits at the first table and orders a dish for the table. As more customers enter the restaurant, they each sit at a table according to the following rule: Assume there are K tables with at least one customer and the number of customers sitting at table k is n_k, the i-th customer sits at table $k \leq K$ with probability $\frac{n_k}{\beta+i-1}$ and starts a new table with probability $\frac{\beta}{\beta+i-1}$ (and order a new dish for the table), where β is a parameter that controls the probability of creating a new table (i.e., how likely are two features to share the same image?). Unlike the IBP where each customer can take multiple dishes, customers in the CRP only take the one dish that is being served at their table.

In fact, this is a novel model, which we call the Indian buffet franchise (IBF), due to the analogous manner that clusters are shared across categories in the Chinese restaurant franchise [Teh et al., 2004]. According to the IBF, each category has its own feature rep-

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

resentation and so the feature representation of category a is the combination of feature ownership matrix $\mathbf{Z}^{(a)}$ and feature image matrix $\mathbf{Y}^{(a)}$. Although the feature ownership matrices are generated independent of each other from the standard IBP ($\mathbf{Z}^{(a)} \sim \text{IBP}(\alpha)$), the feature images are generated from a common source $\mathbf{Y}^{(0)}$, according to the rules of the Chinese restaurant process (i.e., $\mathbf{Y}^{(a)}|\mathbf{Y}^{(0)} \sim \text{CRP}(\beta)$). The new dishes are generated from the feature image prior (so, the independent Bernoulli prior for our current purposes).

Figure 6.4 illustrates how this model infers features after observing the first two objects of each category that have been categorized according to the horizontal scheme of Pevtzow and Goldstone (1994).[3] Let $\mathbf{x}_1^{(a)}$ denote the first object in category a, $\mathbf{y}_k^{(a)}$ to be the k-th feature image of restaurant a (where $a = 0$ denotes the shared CRP, and $a > 0$ denotes category a's IBP), and $n_k^{(a)}$ to be the number of customers who have taken dish k in restaurant a. To start, the participant observes the first object in category 1, $\mathbf{x}_1^{(1)}$ enters category 1's empty IBP and draws Poisson(α) new dishes (which happens to be two for this example). The two new dishes enter the shared CRP sequentially (in any order), which is empty to start. The first customer, $\mathbf{y}_1^{(1)}$ creates a new table and orders a feature image, $\mathbf{y}_1^{(0)}$, which is sampled from the feature image prior (the product of independent coin flips as before). This is the first dish that trickles back down to category 1's IBP and so, $\mathbf{y}_1^{(1)} = \mathbf{y}_1^{(0)}$, which is now served in category 1's IBP. Next, the second customer entering the shared CRP, $\mathbf{y}_2^{(1)}$, sits at the first table with probability $\frac{1}{1+\beta}$ or starts a new table with probability $\frac{\beta}{1+\beta}$. In this example, she happens to start a new table, order its dish $\mathbf{y}_2^{(0)}$, and this trickles back down to category 1's IBP and so, $\mathbf{y}_2^{(1)} = \mathbf{y}_2^{(0)}$. Now, the participant observes the next object, which happens to be the first object of category 2. As this is the first object of category 2, it enters an empty IBP, and draws Poisson(α) new dishes (which happens to be two for this example). The two new features for category 2 become customers of the shared CRP and enter it to sample their feature images. When $\mathbf{y}_1^{(2)}$ enters the CRP, there are two tables with one customer, and so it sits at table 1 with probability $\frac{1}{2+\beta}$, table 2 with probability $\frac{1}{2+\beta}$, and starts a new table with probability $\frac{\beta}{2+\beta}$. It happens to sit at the second table, and so, $\mathbf{y}_1^{(2)} = \mathbf{y}_2^{(0)}$. Next, the second new feature of category 2's IBP enters the CRP, and sits at table 1, 2, or a new table with probabilities $\frac{1}{3+\beta}$, $\frac{2}{3+\beta}$, and $\frac{\beta}{3+\beta}$ respectively. It happens to create a new table, orders new dish $\mathbf{y}_3^{(0)}$ for the table, and brings it back to category 2's IBP (meaning that $\mathbf{y}_2^{(2)} = \mathbf{y}_3^{(0)}$). This process continues as more objects are observed for each category, and Figure 6.4 shows the result after observing two new objects (one that is in category 1 and another that is in category 2). We refer the reader interested in the simulation details to Appendix E.

Figure 6.3cd show that this model overcomes the limitations of the first technique in that it forms a full representation of each object, not just the diagnostic parts. The features diagnostic for categorization are not shared between each category's IBP and thus, they still

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

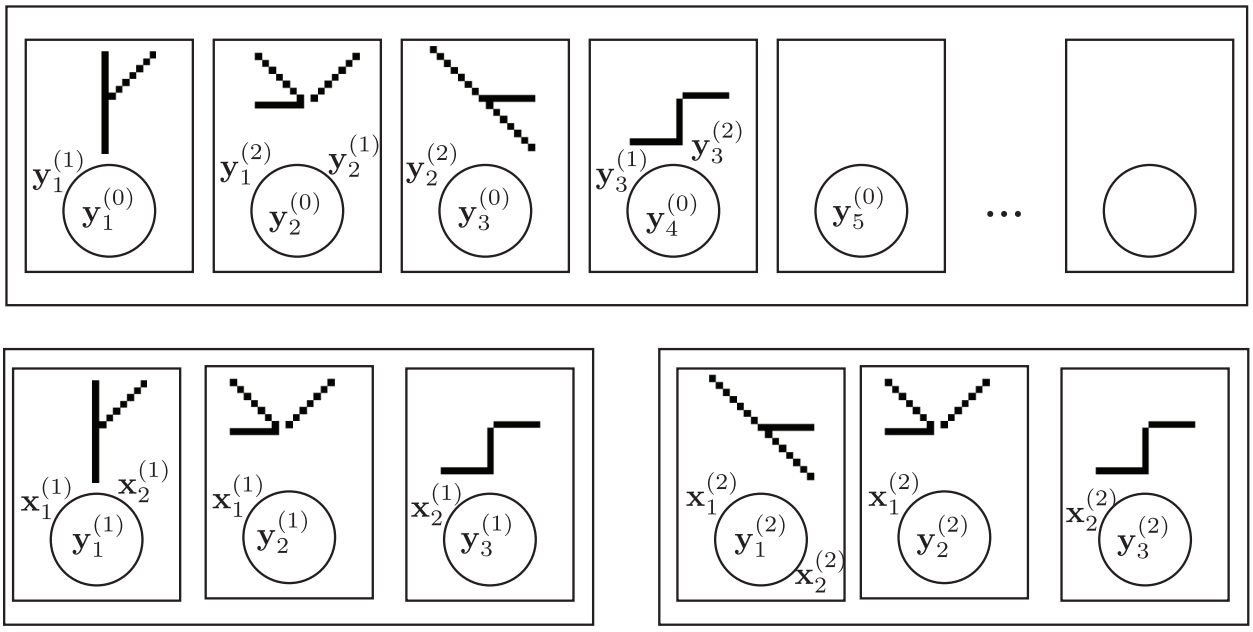


Figure 6.4: The Indian buffet franchise after observing two objects from the two categories under the horizontal categorization scheme of Pevtzow and Goldstone (1994). See text for details.

have special status in this model. Which features are not shared between the category's IBP depends on the categorization scheme of the object set given to the model. As categorization affects the features inferred by the model, Criterion 6 (Category diagnosticity) is satisfied.

6.3 Beyond transformation independence: translations affecting scalings

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

Another two types of transformations that are coupled are translations in depth and scalings of a single moving object. Imagine fixating on an object while it moves towards you. As the object translates over time towards you, its image on your retina grows (scales equally in both dimensions). In a similar vein, participants (children around two and a half years

old) in Experiment 3 of Smith (2005) watched an experimenter move a new object called a "zup" horizontally or vertically (and then the participant actively moved the object the same way). If translation and scaling transformations are coupled (e.g., because the object is translating fast and so, roughly speaking, a"blur" in the same direction occurs; McCarthy, Cordeiro, and Caplovitz, 2012), then participants should be more likely to believe an object scaled in the same direction as the motion is a "zup" than in the opposite direction of the motion. This is exactly the result that Smith (2005) reported, as shown in Figure 6.5a.

Although in its simplest form the tIBP assumes that the transformations occur independently, it is easy to extend it to incorporate dependencies between transformations. If we assume the translation and scale transformations are both normally distributed with their own means, but share the same covariance, then we would expect any variation in one transformation to be the similar to the variation in the other transformation. Thus, if we observe an object translate horizontally, but not vertically, then the model will expect variation in horizontal, but not vertical, scalings (and vice versa if it observes vertical translations). Figure 6.5b shows that the tIBP predictions with this assumption (see the Appendix D for more details) is able to leverage the information learned about translations when observing scalings to predict judgments qualitatively similar to those made by participants in Experiment 3 of Smith (2005). Additionally, the quantitative fit between the model predictions and human results is quite good with a Pearson's product-moment correlation of $r = 0.90$ (and Spearman's rank order correlation of $r = 0.94$).

6.4 Experiment 1: Learning which transformations apply

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

This phenomenon raises the question of whether people can infer which transformations are permissible for a particular feature. We addressed this question by testing whether participants' generalization behavior changed depending on feature transformations that were present in the observed set of objects.

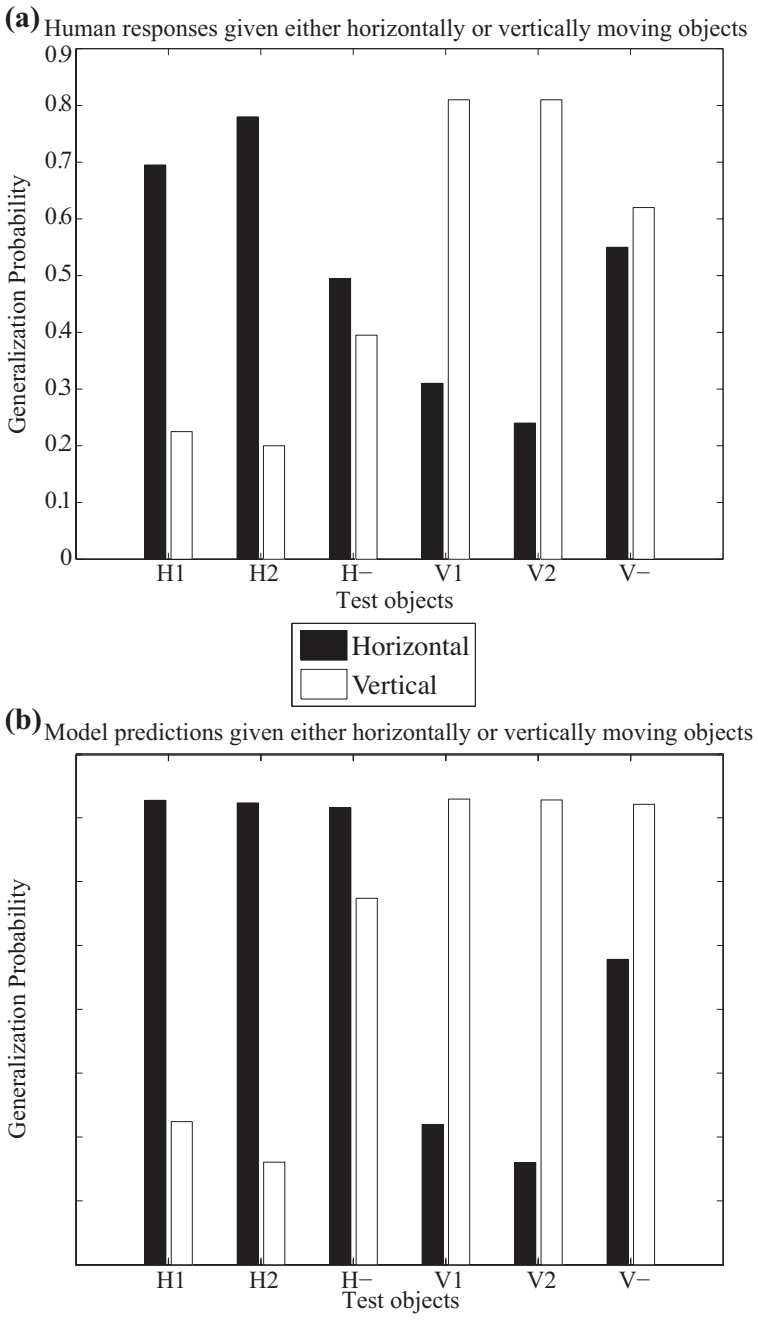


Figure 6.5: Sharing information learned about one type of transformation (translations) with another type of transformation (scalings). (a) Predictions from the transformed Indian Buffet Process model and (b) Human responses from Experiment 3 of Smith (2005).

6.4.1 Methods

Participants and Stimuli

Two groups of 20 online participants recruited from Amazon Mechanical Turk were presented with either the *rotation* (Figure 6.6a) or *size* set (Figure 6.6b). The test images are shown in Figure 6.6c.

Procedure

The same cover story as the One or Two Features Experiment was used. Participants were presented with either the *rotation* (Figure 6.6a) or *size* set (Figure 6.6b). Participants were then asked to rate (0-6 scale) how strongly five new objects appeared to belong to the observed set: *Same Both* (the object that is observed by both groups of participants), *Same Rot* (the last object of the *rotation* set), *Same Size* (the last object of the *size* set), *New Rot* and *New Size* (Figure 6.6c).

6.4.2 Results and Discussion

As expected, participants in the *rotation* condition generalize more to the *New Rot* object than the *size* condition ($t(38) = 4.44, p < 0.001$) and vice versa for the *New Size* object ($t(38) = 5.34, p < 0.001$). Supporting our hypothesis, people infer the appropriate set of transformations (a subset of all transformations) that features are allowed to use for a class of objects.

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

This extension to the tIBP predicts the *New Rotation* object when given the *rotation* set and predicts the *New Size* object when given the *size* set — effectively learning the appropriate type of invariance for a given object class. Figure 6.7b shows the model predictions (see Appendix D for simulation details). The model exhibits nearly the same behavior as the participants (Spearman's rank order correlation of 0.68), except for rating the *Same Size* square when given the *rotation set*, which is due to the images looking identical after downsampling.

One interpretation of these results is that we have learned that a square is not orientation-invariant because it is called a new name when rotated 45 degrees (a diamond). This provides a novel potential explanation for why some shapes are orientation-invariant and other are not:

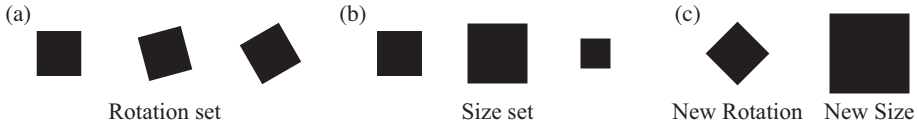


Figure 6.6: Stimuli for investigating how different types of invariance are learned for different object classes. (a) The *rotation* training images. (b) The *size* training images. (c) The two images used to test whether people inferred that rotation transformations were allowed (*New Rotation*) or that size transformations were allowed (*New Sizes*).

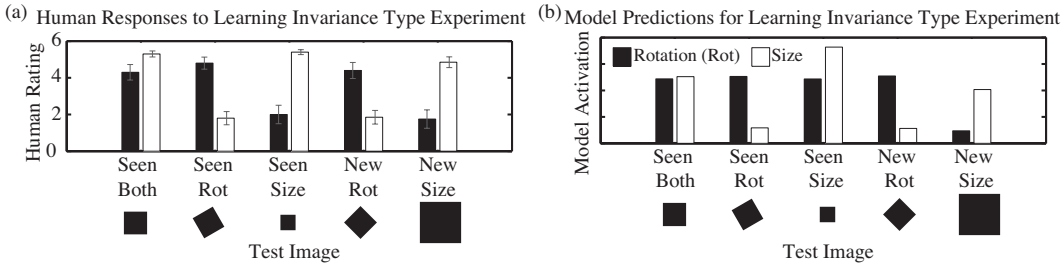


Figure 6.7: Learning the type of invariance. (a) Responses of human participants. (b) Model results.

A shape is assumed to be orientation-invariant unless there is reason from some observation that its identity changes when it is rotated (e.g., it is called a new name). Thus the image of a rotated square is not a square because we have observed others call it a diamond.[4]

6.5 Capturing incremental feature learning

In this section we address Criterion 7 (Incremental learning), which states that a computational model of human feature learning should be sensitive to the order that objects are presented to it. A common assumption (and criticism, see for example Kruschke, 2006; Jones & Love, 2011; and Trueblood & Busemeyer, 2011) of Bayesian models is that the probability of a sequence of objects is *exchangeable*, meaning that the model assigns the same probability to any order of the objects.[5] As human category and feature learning are sensitive to the

[4]Note that we are not suggesting that squares and diamonds are observed from specific-viewpoints more often (than for example, a pentagon). Conversely, we are suggesting that pentagons are not given different labels when observed from different viewpoints, but squares and diamonds are given different labels when observed from different viewpoints. Although we do not contend the differences between squares and diamonds may be due to other more perceptual factors (this may be where the difference in labelling came from in the first place), we are interested in whether people can learn which transformations apply based on how labels are applied to transformed images of the feature.

[5]This is a valid criticism for Bayesian models that assume that the order in which information is observed does not matter. Although this is true of many Bayesian models of cognition, this assumption is not necessary

order that information is presented [Anderson, 1990, Schyns and Rodet, 1997], order effects are a challenge for Bayesian models that assume exchangeability. This includes the models in our framework, which also assume exchangeability and are thus unable to produce order effects. We consider how to address this limitation in this section after first describing the order effects reported by Schyns and Rodet (1997).

In Experiment 2 of Schyns and Rodet (1997), participants learned to categorize three different types of "Martian cells" (circles with differently shaped blobs inside). Two of the blobs, a and b, were diagnostic for determining category membership. Figure 6.8a shows three different categories using two parts that we constructed in the same style as Schyns and Rodet (1997): A (the cell contained the blob a), B (the cell contained the blob b), and AB (the cell contained the blobs a and b connected in a particular spatial arrangement). They trained two different groups of participants to learn the three categories in one of two orders: $A \rightarrow B \rightarrow AB$ (shown in Figure 6.8b) or $AB \rightarrow A \rightarrow B$ (shown in Figure 6.8c). Their hypothesis was that those participants who learned to categorize AB first would learn a single feature that was the conjunction of a and b in a particular spatial arrangement and so they would not extend category membership to a new Martian cell containing a and b in a new spatial arrangement. On the other hand, those that learned AB last would already have learned the features a and b and thus would think of the category AB as the conjunction of two pre-existing features a and b that were learned to perform the two previous categorizations. Thus, if it is simply the conjunction of these two features, the two features should be allowed to be in any spatial arrangement and so they should extend membership to a new Martian cell containing a and b in a new spatial arrangement. The results of the experiment supported their hypothesis; participants that learn AB last extended category membership for AB to a novel Martian cell with a and b in new spatial arrangements (shown in Figure 6.8d) more often than the participants who learned A and B first (shown in Figure 6.8e). Thus, the order in which data are presented to participants affects the features they inferred because those who learned AB first inferred the features a, b and ab and those who learned AB last inferred the features a and b.

So far, we have focused on understanding how people infer features to represent objects from the computational level. We have not discussed the actual processes that might implement the computational solution (an algorithmic-level explanation). One method for approximating the optimal solution is Gibbs sampling [Geman and Geman, 1984], which was

to formulate Bayesian models. In fact, it is usually made for pragmatic reasons (computational ease) and not due to a strong theoretical commitment that agents should be indifferent to the order of observing information. We thank Michael Lee for pointing this out. Exchangeability is a weaker assumption, which is related to the common assumption that a sequence of objects are independently and identically distributed. An independently and identically distributed sequence of objects is exchangeable, but an exchangeable sequence of objects is not necessarily identical and independently distributed. For example, the probability of a sequence of objects in our mathematical framework is exchangeable, but it is only independent and identically distributed given the feature representation of the objects. See Bernardo and Smith (1994) for more details on the distinction between objects being exchangeable or independently and identically distributed.

used to generate the model predictions for the simulations previously reported. Although Gibbs sampling is an effective algorithm for approximating complex probability distributions and psychologically plausible in some situations [Sanborn et al., 2010], it requires all objects to be observed before approximation commences. Thus, it must start from scratch whenever it is given a new object. As shown in Figures 6.8bc, the features learned by a Gibbs sampler are the same regardless of the object presentation order (see Appendix F for simulation details).

A different approach would be to investigate more psychologically valid forms of approximating the probability distributions defined by our Bayesian models. Inferring the distributions defined by the models is intractable, and so, even machine learning methods can only approximate them. Although some approximation algorithms do not show order effects (like Gibbs sampling), there are some classes of statistically motivated approximation techniques (those yielding arbitrarily good precision of the probability distributions with increasing resources) that do introduce order effects.

This is the idea behind rational process models based on *particle filtering*, which explain order effects as an artifact of using a statistically motivated approximation to a Bayesian model [Sanborn et al., 2010]. A rational process model for the IBP that is an *incremental* learner (given objects sequentially) has already been derived [Wood and Griffiths, 2007], based on a statistical algorithm known as a particle filter [Gordon et al., 1993]. As Sanborn et al.(2010) used a particle filter to explain order effects in categorization, it is plausible that a rational process model for the IBP can explain the feature learning order effects found by Schyns and Rodet (1997). The incremental form of the IBP uses a number of "particles" for approximation, where each particle is a feature representation inferred for the currently observed objects. The probability of a feature representation is given by the proportion of particles containing that feature representation. Each time a new object is given to the incremental algorithm, the representations in each particle are updated to account for the new object. The feature representation in each particle is inferred sequentially as each object is observed using the previously inferred feature representations. Importantly, the feature representation for an object within a particle does not change once features have been inferred for it.[6] This property is the most important difference between the previously discussed Gibbs sampler and the particle filter: Each particle in the particle filter infers features for an object once, only when it is first observed.

Because each particle in the particle filter only infers features for objects once, it is weakly dependent on the object presentation order (especially when only a small number of particles are used). This allows it to account for the order effects observed by Schyns and Rodet (1997). Consider the condition where participants learn AB, A, and then B. When a particle first observes the AB objects, it infers a single feature, ab because it can encode

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

all of the observed objects. Next, the particle encounters the A objects. As the feature it currently posses, ab, cannot explain all of the objects in A (they do not have the b part), the particle captures the objects by inferring the a part as a feature. By an analogous argument, the particle infers the b part as a feature to explain the objects in B that it observes next. As the particle represents the AB objects with a single feature that is the conjunction of the ab parts in a particular spatial arrangement, it cannot not represent the test objects that have a and b in new spatial arrangements with its ab feature. Thus, in this case, it would not extend the AB category membership to these objects.

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

By extending the incremental form of the IBP by Wood and Griffiths (2007) to the tIBP (the tIBP is necessary because a, b, ab have independent translations for each object they are in), we can account for the order effects of Schyns and Rodet (1997).[7] Figure 6.8bd shows that the features inferred and predictions made by the tIBP using a Gibbs sampler (Gibbs) and a particle filter (PF) are the same as those made by participants in Schyns and Rodet (1997). However, as shown in Figure 6.8ce, the tIBP using Gibbs sampling is insensitive to the presentation order and incorrectly predicts that only a and b are learned as features and thus predicts that participants in both conditions should generalize AB to objects with a and b in a new spatial arrangement. Following our previous discussion, the features learned by the particle filter depend on the order that objects are presented (in the same manner as participants) and so, it makes the same predictions as people by learning incrementally (see Appendix F for more details). Using particle filtering to perform inference rather than Gibbs sampling results in a model that learns features incrementally (in that the order of object presentation matters), and so it satisfies Criterion 7 (Incremental learning).

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

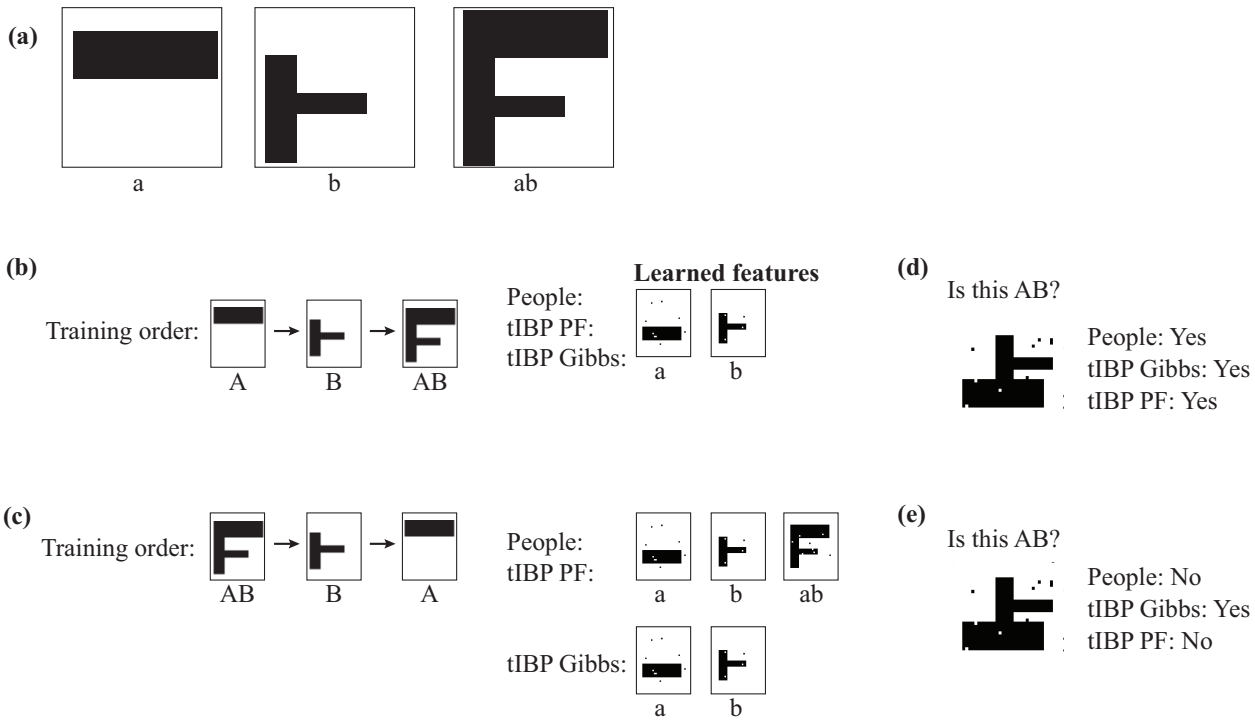


Figure 6.8: Effects of order on feature learning. The order of observing objects from categories affects the features inferred and generalization judgments of participants in the experiments of Schyns & Rodet (1997) and the tIBP using particle filtering (PF), but not the tIBP using Gibbs sampling (Gibbs). (a) Three parts used to construct the stimuli of Schyns & Rodet (1997). (b) - (c) The features learned after observing objects from categories in order the AB last order and the AB first order, respectively. (d) - (e) Categorizing objects with parts a and b in a new spatial arrangement as AB after observing objects from categories in order the AB last order and the AB first order, respectively.

Chapter 7

Theoretical implications and conclusions

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

Furthermore, we have demonstrated that the proposed computational framework can satisfy the seven criteria we established for a computational explanation of human feature learning. First, we showed in Chapter 3 that Criterion 1 (sensory primitives) and Criterion 2 (unlimited features) are satisfied by setting up the computational problem as a problem of matrix factorization and using the IBP as our prior on the feature ownership matrix. Second, we demonstrated in Chapter 4 that the simplest model in the framework infers different features for an image depending on the other images presented in its context (satisfying Criterion 3: Context sensitivity). Furthermore, we showed that prior expectations of simplicity (representations with fewer features are better) in Chapter 2 and proximity (adjacent pixels are likely to share the same value) can be included into the model in Chapter 6, satisfying Criterion 4 (Prior expectations).

Although the initially presented models in our computational framework satisfy the first four criteria, they did not satisfy the other criteria. We therefore defined new models that do so. We defined a novel model, the transformed Indian buffet process, which inferred transformation-invariant features in Chapter 5 and incorporated prior knowledge as to how transformations are dependent on each other to explain novel and previous experimental results in Chapter 6 (satisfying Criterion 5: Transformational invariance). Then in Chapter 6, we explored two methods for including categorization information into the framework, and defined a novel process, the Indian buffet franchise, to properly account for previous human

feature learning results (satisfying Criterion 6: Category diagnosticity). Finally in Chapter 6, we also defined an incremental learner (as a method of inferring the features for the tIBP), which learned different features depending on the the order that objects were presented to it in the same manner as people (satisfying Criterion 7: Incremental learning). To the best of our knowledge, this is the first computational framework capable of satisfying all seven criteria.

In this chapter, we critique the theoretical implications of the proposed framework of human representation learning, discuss future directions and conclude. First, we interpret representations in Bayesian models in terms of traditional views of psychological representations and Marr's levels. Second, we explore issues of mimicry between representations [Anderson, 1978], and explain our justification for inferring the representations participants use in our experiments when (in most cases) we did not make any explicit commitments about how participants processed those representations. Third, we discuss the relation between Bayesian modeling and Behaviorism as they are both similar in their focus on the understanding the input-output mapping that is human behavior. Afterwards, we discuss implications of the Bayesian models for perceptual theories of object representations. Finally, we consider directions for future research and conclude.

7.1 Marr's levels and the interpretation of representations in Bayesian models

In this dissertation, we have been exploring the representations inferred by models expressed at the computational level. Although the representations posited at the algorithmic level are given psychological importance, the representations used at the computational level traditionally are not. This is because computational level solutions are typically used to explain why an agent well-adapted to her environment should act in a particular way, regardless of the actual algorithm she uses to act that way. Our analysis of human feature learning gives psychological importance to the representations learned by our mathematical model. What justification do we have in giving psychological importance to these representations?

One cannot specify a problem or formulate a solution at the computational level, without first choosing a representation for the problem. The representation chosen at the computational level affects the solution identified by the model and thus, the behavior of an agent (and in turn, also the algorithmic and implementational explanations). To see this, consider the computational problem of generalization (which underlies many cognitive problems, such as word learning, categorization, and property induction). As argued by Shepard (1987), Bayesian inference solves the problem of generalization at the computational level and matches human (and pigeon) generalization behavior in a wide range of perceptual domains. However, the solution provided by Bayesian inference depends on how properties in the domain are represented. For example, consider generalization over two dimensions

(such as circles varying in their size and the orientation of a radial line) where properties are assumed to be intervals of the dimensions (contiguous rectangular regions). The ideal solution depends on whether the rectangular regions are aligned with the axes or indifferent to the axes (any orientation of rectangular region is allowed).

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

These examples suggest that Bayesian models can be psychologically relevant for understanding how people infer context-sensitive feature representations. For example, consider the debate between fixed and flexible feature representations discussed earlier. As our work is a computational level analysis, it does not provide a definitive answer to the debate. This is because it cannot distinguish between the processes used to infer different features for the same object (whether people are truly learning new features or simply re-weighting pre-existing features) because both are valid cognitive processes that might approximate the computational level solution. However, this does not imply that the analysis leaves us where we started. We believe that our results weakly favor the feature learning interpretation. This is because people must be re-weighting the features used to represent an object remarkably often to capture the changes in generalization behavior from the same object in different contexts, if the feature re-weighting process is the correct process-level description. Regardless of whether our results are ultimately due to a feature learning or re-weighting process, the key principles governing feature learning that we have identified through our approach establish important criteria for evaluating different algorithmic-level accounts of how feature representations are inferred in future research.

Despite this argument, the representations in Bayesian models may not be what many psychologists have typically thought of as psychological representations, since it is unclear whether or not they are being explicitly manipulated at the process level. This is one source of the recent controversy about the utility of Bayesian models [Jones and Love, 2011]. Our perspective is that the representations identified by the mathematical models should be interpreted as those that are the most useful for determining generalization behavior in a particular situation (which can be simple line segments or more complex configurations depending on the context). This is theoretically similar to the "intermediate representations"

posited by Biederman and colleagues [Biederman, 1987, Biederman and Cooper, 1991] or borrowing a term from the categorization literature, "basic level features" [Pomerantz, 2003]. However, further work is needed to explore precisely in what way representations in Bayesian models are explanatory and how their representations relate to the representations typically posited by psychologists.

7.2 Mimicry of different representations

Another potential issue for our argument that representations at the computational level have psychological importance is that it can be difficult from experimental data to distinguish between representations without the processes acting on the representations [Anderson, 1978]. Anderson's argument can be summarized as follows: If two representations preserve the same distinctions to external stimuli, then a function exists mapping one representation to the other representation (and vice versa). After applying the mapping function, the process of the second representation can be applied to the first (as it now is in the format of the second representation) and the first representation will act identically in all behavioral tasks to the second representation. Thus, we cannot distinguish between two representations without specifying processes acting on each representation.

At first glance, this argument may seem to cast doubt on the usefulness of the presented research. If we cannot distinguish between representations because one could be mimicking the other, then how can we talk about some participants inferring one representation in one condition and other participants inferring a different representation in a different condition? There are three main rebuttals to Anderson's argument: different distinctions, simplicity, and optimality (or rationality). We now discuss these rebuttals.

One assumption of Anderson's argument is that both representations preserve the same distinctions in the external world. Otherwise, for one of the representations (the one that is coarser), a function cannot be found mapping it to the other representation (the one that makes more distinctions). For example, consider two representations of the same object: a *holistic* representation, which consists of single feature that contains the whole of the object and a *part-based* representation that combine to reproduce the object (as in Experiment 1 of Chapter 4). The *holistic* representation cannot be mapped into the *part-based* representation; however, the *part-based* representation can be mapped into the *holistic* representation. Thus no matter what process a cognitive model using the *holistic* representation uses, if human behavior depends on one part of the object (e.g., all objects with one of the parts are members of a category), then people must be using the *part-based* representation rather than the *holistic* representation. However, the *part-based* representation can always be mapped onto the *holistic* representation, so if human behavior is in accordance with a *holistic* representation, we can always mimic it with a *part-based* representation using Anderson's argument. Although the different distinctions reply can enable us to distinguish when a finer-grained representation is needed instead of a coarser representation, it can never rule

out a finer-grained representation in favor of a coarser one.

Even if two possible representations make the same distinctions (or if we want to argue in favor of a coarser representation), using assumptions about the simplicity of the process can distinguish between different representations. If two representations explain the observed objects equally, but one representation uses fewer features and so in a sense, it is simpler than the other, we assume that the simpler representation is used. This is not a controversial assumption. In fact, psychologists in many domains use this principle [Feldman, 2000, Pomerantz and Kubovy, 1986] and it has been proposed as a principle that unifies all cognitive processes [Chater and Vitanyi, 2003]. Additionally, one way of interpreting the results of the experiments reviewed in this article is that people infer the simplest (in the sense of fewest number of features) feature representation that adequately reconstructs the observed set of objects. Thus, despite the issue of mimicry between representations without specifying a process, we can argue for one representation over another via an appeal to the simplicity of the representation and process acting on the representation.

Another constraint is to assume that the mind uses representations that are optimal with respect to its environment. Thus, we distinguish between psychological representations by selecting the one with maximal probability given the environmental statistics. This assumption can be controversial [Danks, 2008], but has a long and productive history in psychology [Brunswik, 1956, Gibson, 1966, Hecht et al., 1942] going by a number of names such as "rational analysis" [Anderson, 1990, Chater and Oaksford, 1999], Bayesian modeling [Griffiths et al., 2010, Tenenbaum et al., 2011] or ideal observer modeling [Geisler, 2003]. As our computational level models are Bayesian models, the representations they infer are justified using the optimality constraint. Finally, due to the strong mathematical connections between optimality and simplicity, these may actually be two names for the same constraint (though this interpretation is controversial; see Chater 1996; Feldman 2009; van der Helm 2000, 2011).

Finally, neuroscience may provide another possible source of constraints for distinguishing between representations making the same distinctions. For example, perhaps we should favor representations that are more consistent with our current knowledge of early visual processing from neuroscience, such as line segments [Hubel and Wiesel, 1962] or Gabor patches [DeValois and DeValois, 1988]. Although these results from neuroscience are undoubtably important for understanding early visual processing (e.g., Poggio, Edelman and colleagues have a long history of developing visual processing models based on neuroscience results; see Serre & Poggio, 2010 for an accessible discussion of their results), their relevance for understanding representations used in higher-level psychological processing is unclear [Uttal, 1971]. How can configural superiority effects (where wholes are recognized faster than their parts) be understood in terms of line segments or Gabor patches as basic units [Pomerantz et al., 1977, Pomerantz, 1978, Pomerantz, 2003]? This is not meant as an argument against ever using neuroscientific evidence to understand psychological representations, but rather to caution against inappropriately constraining psychological theories by broadly generalizing neuroscience results without considering the behavioral evidence in tandem.

7.3 Bayesian modeling and Behaviorism

From a theoretical perspective, Bayesian modeling and Behaviorism have an interesting and unclear relationship. In Chapter 1, we noted that there are similarities between these two theoretical frameworks as proponents of both frameworks are interested in describing the input-output mapping that produces human behavior in terms of the structure of environmental inputs. In fact, without any additional theoretical guidance, the essential idea of Bayesian modeling is to explain psychological behavior in terms of optimal adaptation to the structure of the environment. This is precisely the sort of explanations that Behaviorism sought [Jones and Love, 2011, Skinner, 1977, Watson, 1913]! So, is Bayesian modeling a modern form of Behaviorism dressed in fancy mathematics?

In addition to being a valid critique of Bayesian modeling without any additional theoretical commitments or assumptions, it is in fact a critique of any computational paradigm that is universal, or in other words is able to capture any input-output mapping, without any further commitments or assumptions (e.g., Connectionism or the Rules and Symbols approach). Each computational framework provides a language for specifying how inputs are mapped to outputs, and how this mapping changes (or does not change) with experience. The goal of a computational framework is to provide a simple description of human behavior in a domain, extract some key principles or regularities that underlie how people respond to different inputs, and exploit the principles and regularities in new situations. When the description of human behavior is too complex or does little more than recapitulate human behavior without being generalizable to new domains, the computational explanation should be rejected and a new one sought. In other words, computational models (Bayesian models included) are useful to the extent that they provide "interesting" and "useful" descriptions of human behavior. For example, if the only task that human behavior is rational with respect to is the task of producing the observed behavior, then clearly, we have not produced a useful model of human behavior. However, has any Bayesian modeler (or any other type of modeler) ever proposed a model that merely recapitulates human behavior as an explanation [Chater et al., 2011]?

The representations and assumptions about how those representations are used to transform inputs into outputs are the key differences between a Behaviorist and a computational cognitive explanation of behavior. As we discussed earlier, computational representations are not like those typically used in psychology. However, they are still meaningful nonetheless as it is impossible to solve a problem without some sort of computational representation.

In a sense, exploring what sort of representation in a model is necessary to capture a set of observations and relating this to the sort of representations people use has been the goal of this dissertation: Given that we fix the mapping from representations to outputs, what representation should be used to capture the set of observations? Fortunately, we were able to specify a framework that infers psychologically valid representations without needing to add more and more assumptions to the model, which would have simply lead to a redescription of the desired representations in terms of the prior assumptions. And

when the psychological and computational representations differed, it lead to an exploration of the different inductive biases people must be using (from a computational perspective). This in turn led to more novel predictions about how people infer representations. Thus, from at least a methodological perspective, the computational framework has been useful for discovering new phenomena about how people infer representations that might not have discovered otherwise.

7.4 Connection to perceptual theories of object representation

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

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

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

On the other hand, our model also decomposes objects into multiple parts (called features) and so it is a feature model as well. This is a desirable property as there is strong phenomenal [Hoffman and Richards, 1985, Palmer, 1977] and empirical [Biederman and Cooper, 1991, Braunstein et al., 1989, Palmer, 1977] evidence that people represent objects using features. Although our proposal does not yet include relations between features or feature hierarchies (though we are pursuing future work in this direction),

we have already demonstrated how more structure can be included into our object representations (beyond just a set of features). The last three demonstrations of the model included structure and relations between the transformations of the features. This incorporates realistic prior expectations of how the features of objects are transformed into our model and thus our model already has some of the characteristics of structural description representations. Thus, our model integrates the strengths of three of the most popular proposals for object representation in perception.

7.5 Future directions

Although we have focused on inferring feature representations for objects given visual images, it should be straightforward to extend our approach to inferring representations for other modalities given their raw sensory data. For example, Yildirim and Jacobs (2012) explored an interesting application of our approach to inferring representations with multimodal features. They demonstrated that a simple extension of the IBP can infer a set of features from visual and auditory sensory data. Thus, it is clear that our approach is not limited to the visual domain, but rather, we have been using the visual images as an effective test bed for exploring more and more powerful extensions to the basic IBP model.

In the future, it would be interesting to explore how to extend our framework to incorporate prior expectations about the goodness of different feature images (Gestalt constraints such as good continuation), more detailed category information, and to incorporate relations between objects and their features, and explore what kinds of features are learned from natural scenes. Including Gestalt constraints into our framework amounts to defining a more elaborate prior distribution on the types of feature images allowed, $P(\mathbf{Y})$. We demonstrated one method that we can use to learn more psychologically valid features: defining a prior on feature images that favors adjacent pixels to have the same value. Although this probability distribution does not fully capture human perceptual expectations, it is straightforward to improve the psychological validity of the inferred features by incorporating more complex probability distributions used in computer vision (e.g., Sudderth & Jordan, 2011). In fact, our framework provides a natural method for evaluating different computational proposals of perceptual expectations and to investigate the precise nature of these expectations. Another future direction is to explore the sorts of features inferred by the model from natural scenes. Unfortunately, learning features from natural scenes is not computationally feasible using the inference algorithms described in this dissertation. However, there have been exciting recent developments in machine learning that improve the efficiency of feature inference for the tIBP to allow it to work on images taken from video games (e.g., Mario Bros.), surveillance cameras, and the Microsoft Kinect sensor [Hu et al., 2012, Joho et al., 2011]. Potentially, these new algorithms could be used to see what kinds of features are learned by the tIBP when given a database of natural scenes.

Machine learning research has already extended the IBP to deal with relations between

the observed objects [Miller et al., 2009], and it is likely that similar techniques could be used to incorporate relations between features (e.g., ON TOP OF). Also, dynamic features that evolve over time could be learned using a time series form of the IBP [Fox et al., 2009, Williamson et al., 2010]. Finally, we are interested in learning features that can occur more than once in an image (where each instantiation has its own transformation parameter). This could be accomplished by extending the tIBP in the same vein as the Poisson-Gamma model, an extension of the IBP where that there can be multiple instances of a feature in the same object [Titsias, 2008]. It is plausible that a simple extension to this model could allow the IBP to learn the expected number of times a feature should be instantiated in an object, and therefore, be able to learn that there should be two features per object given the *separate* test objects.

7.6 Conclusions

Our capability to form complex representations about our environment and adapt these representations to changes in our environment enables us to navigate and manipulate our environment successfully. As the best artificial systems for solving most cognitive problems still pale in comparison to human ability, our success remains a puzzle. Inferring a representation is an inductive problem, and as such, there are always multiple possible solutions to the problem. Thus, people must use some extra constraints to disambiguate the solutions. In this dissertation, we explored computational models that formalize prior constraints and integrate them with observations. Importantly, the solution given by the models depend on the context, which starts to provide an explanation for how people are able to adapt their representations to differences in the environment.

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

Furthermore, we have demonstrated that models constructed from the proposed computational framework satisfy the seven criteria (sensory primitives, unlimited features, context sensitivity, prior expectations, transformational invariance, category diagnosticity, and incremental learning) we established for a computational explanation of human feature learning. To the best of our knowledge, this is the first computational framework capable of satisfying all seven criteria. Although this is an important step towards understanding the problem of how people construct representations of the world, we have only started to address the simplest case of this problem. For example, we have not addressed how we construct hierarchical, relational, and other types of representations that have more complex structure.

Thus, there is still a lot of work to be done on developing a formal understanding of how we construct representations of the world. It is our hope that it will be possible to define new models in the computational framework that we presented, which can address these issues.

In addition to providing and empirically testing explanations for how people learn representations, the presented computational framework and results serve an important role for Bayesian models in psychology. Bayesian models have been used to explain a range of different cognitive phenomena [Chater et al., 2006, Chater and Oaksford, 2008, Griffiths et al., 2010, Tenenbaum et al., 2011], but the representations used in the models are often hand-picked by the modeler and usually specific to the particular investigated phenomenon. However, people cannot rely upon a modeler to formulate their representations, so how do people form representations in a novel domain or context? We presented a unified answer to this question: a representation is used for a set of observed stimuli depending on how well it explains the set of observed stimuli and its prior probability. Although we only have begun to sketch the foundation of this answer and are still a long way from explaining how art theorists learn to represent Jackson Pollock paintings, the computational framework provides an important step towards the larger goal of explaining how people learn to construct complex representations based on their knowledge and experiences in the world.

Bibliography

[Abdi et al., 1998] Abdi, H., Valentin, D., and Edelman, B. G. (1998). Eigenfeatures as intermediate-level representations: The case for PCA models. *Brain and Behavioral Sciences*, 21:17–18.

[Anderson, 1978] Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, 85(4):249–277.

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

[Anderson and Bower, 1972] Anderson, J. R. and Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, 79(2):97–123.

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

[Attneave, 1950] Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology*, 63:546–554.

[Attneave, 1954] Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61:183–193.

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

[Bell and Sejnowski, 1995] Bell, A. and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1004–1034.

[Bernardo and Smith, 1994] Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian theory*. Wiley, New York.

[Biederman, 1987] Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147.

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

[Biederman and Gerhardstein, 1993] Biederman, I. and Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, 12:1506–1514.

[Biederman and Shiffrar, 1987] Biederman, I. and Shiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(4):640–645.

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

[Blodgett, 1929] Blodgett, H. C. (1929). The effect of the introduction of reward upon the maze performance of rats. *University of California Publications in Psychology*, 4:113–134.

[Boring, 1961] Boring, E. G. (1961). Fechner: Inadvertent founder of psychophysics. *Psychometrika*, 26(1):3–8.

[Brady and Oliva, 2008] Brady, T. F. and Oliva, A. (2008). Statistical learning using real-world scenes: Extracting categorical regularities without conscious intent. *Psychological Science*, 19(7):678–685.

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

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

[Brunswik, 1952] Brunswik, E. (1952). *The conceptual framework of psychology*. University of Chicago Press, Chicago.

[Brunswik, 1956] Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press, Berkeley, CA.

[Bülthoff and Edelman, 1992] Bülthoff, H. H. and Edelman, S. E. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, 89(1):60–64.

[Cantril et al., 1949] Cantril, H., Jr., A. A., Hastorf, A. H., and Ittelson, W. H. (1949). Psychology and scientific research. *Science*, 110:461–464, 491–497, 517–522.

[Carreira-Perpinán, 1997] Carreira-Perpinán, M. A. (1997). A review of dimension reduction techniques. *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, pages 1–69.

[Chater, 1996] Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, 103:566–581.

[Chater et al., 2011] Chater, N., Goodman, N., Griffiths, T. L., Kemp, C., Oaksford, M., and Tenenbaum, J. B. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. *Behavioral and Brain Sciences*, 34(4):194–196.

[Chater and Oaksford, 1999] Chater, N. and Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science*, 3:57–65.

[Chater and Oaksford, 2008] Chater, N. and Oaksford, M. (2008). The probabilistic mind: prospects for a Bayesian cognitive science. In *The probabilistic mind*, pages 3–31. Oxford University Press, Oxford, England.

[Chater et al., 2006] Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Special issue on "Probabilistic models of cognition". *Trends in Cognitive Sciences*, 10(7).

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

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

[Chomsky, 1957] Chomsky, N. (1957). *Syntactic structues*. Mouton, The Hague.

[Chomsky, 1959] Chomsky, N. (1959). A review of B.F Skinner's Verbal Behavior. *Language*, 31:26–58.

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

[Danks, 2008] Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In Chater, N. and Oaksford, M., editors, *The probabilistic mind: Prospects for rational models of cognition*, pages 59–75. Oxford University Press, Oxford.

[Day and Goldstone, 2012] Day, S. B. and Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist*, 47(3):153–176.

[DeValois and DeValois, 1988] DeValois, R. L. and DeValois, K. K. (1988). *Spatial Vision*. Oxford University Press, New York.

[Doshi-Velez et al., 2009] Doshi-Velez, F., Miller, K. T., Van Gael, J., and Teh, Y. W. (2009). Variational inference for the Indian buffet process. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pages 137–144, Clearwater Beach, FL. Journal of Machine Learning Research.

[Ebbinghaus, 1913] Ebbinghaus, H. (1885/1913). *Memory: a contribution to expeirmental psychology*. Teachers College, Columbia University Press, New York.

[Edelman, 1999] Edelman, S. E. (1999). *Representation and Recognition in Vision*. MIT Press, Cambridge, MA.

[Edelman and Intrator, 1997] Edelman, S. E. and Intrator, N. (1997). Learning as extraction of low-dimensional representations. In *The Psychology of Learning and Motivation*, pages 353–380. Academic Press, San Diego.

[Einstein and Infeld, 1942] Einstein, A. and Infeld, L. (1942). *The evolution of physics*. Simon and Schuster, New York.

[Fahlman and Lebiere, 1990] Fahlman, S. E. and Lebiere, C. (1990). The cascade-correlation learning architecture. In Touretzky, D. S., editor, *Advances in Neural Information Processing 2*, pages 524–532, Los Altos, CA. Morgan Kaufmann.

[Feldman, 2000] Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, 407:630–633.

[Feldman, 2009] Feldman, J. (2009). Bayes and the simplicity principle in perception. *Psychological Review*, 116(4):875–887.

[Fiser and Aslin, 2001] Fiser, J. and Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 12(6).

[Fiser and Aslin, 2002a] Fiser, J. and Aslin, R. N. (2002a). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3):458–467.

[Fiser and Aslin, 2002b] Fiser, J. and Aslin, R. N. (2002b). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, 99(24):15822–15826.

[Fiser and Aslin, 2005] Fiser, J. and Aslin, R. N. (2005). Encoding multielement scenes: Statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General*, 134(4):521–537.

[Fox et al., 2009] Fox, E., Sudderth, E., Jordan, M. I., and Willsky, A. (2009). Sharing features among dynamical systems with beta processes. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 549–557.

[Friedman and Tukey, 1974] Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 100(9):881–890.

[Gardner, 1987] Gardner, H. (1987). *The mind's new science: A history of the cognitive revolution*. Basic Books, New York.

[Garner, 1974] Garner, W. R. (1974). *The Processing of Information and Structure*. Erlbaum, Maryland.

[Gauthier and Tarr, 1997] Gauthier, I. and Tarr, M. J. (1997). Becoming a "greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, 37:1673–1682.

[Gauthier et al., 1998] Gauthier, I., Williams, P., Tarr, M. J., and Tanaka, J. (1998). Training "greeble" experts: A framework for studying expert object recognition processes. *Vision Research*, 38:2401–2428.

[Geisler, 2003] Geisler, W. S. (2003). Ideal observer analysis. In Chalupa, L. and Werner, J., editors, *The visual neurosciences*, pages 825–837. MIT Press, Boston.

[Geman et al., 1992] Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias-variance dilemma. *Neural Computation*, 4:1–58.

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

[Getty et al., 1979] Getty, D. J., Swets, J. B., Swets, J. A., and Green, D. M. (1979). On the prediction of confusion matrices from similarity judgments. *Perception & Psychophysics*, 26:1–19.

[Ghahramani, 1995] Ghahramani, Z. (1995). Factorial learning and the EM algorithm. In *Advances in Neural Information Processing Systems*, volume 7, pages 617–624, Cambridge, MA. MIT Press.

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

[Gibson and Gibson, 1950] Gibson, E. J. and Gibson, J. J. (1950). The identifying response: a study of a neglected form of learning. *American Psychologist*, 7:276.

[Gibson, 1966] Gibson, J. J. (1966). *The senses considered as perceptual systems*. Houghton-Mifflin, Boston.

[Gibson and Gibson, 1955] Gibson, J. J. and Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment. *Psychological Review*, 62(1):32–41.

[Gluck and Bower, 1988] Gluck, M. A. and Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, 27:166–195.

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

[Goldstein et al., 2010] Goldstein, M. H., Waterfall, H. R., Lotem, A., Halpern, J. Y., Schwade, J. A., Onnis, L., and Edelman, S. (2010). General cognitive principles for learning structure in time and space. *Trends in Cognitive Sciences*, 14(6):249–258.

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

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

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

[Goldstone, 2003] Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In *Perceptual Organization in Vision: Behavioral and Neural Perspectives*, pages 233–278. Lawerence Erlbaum Associates, Mahwah, NJ.

[Goldstone et al., 2008] Goldstone, R. L., Gerganov, A., Landy, D., and Roberts, M. E. (2008). Learning to see and conceive. In *The new cognitive sciences (Part of the Vienna Series in Theoretical Biology)*, pages 163–188. MIT Press, Cambridge, MA.

[Goldstone and Steyvers, 2001] Goldstone, R. L. and Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, 130:116–139.

[Goldwater et al., 2009] Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112:21–54.

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

[Gordon et al., 1993] Gordon, N., Salmond, J., and Smith, A. (1993). A novel approach to non-linear/non-Gaussian Bayesian state estimation. *IEEE Proceedings on Radar and Signal Processing*, 140:107–113.

[Green and Richardson, 2001] Green, P. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28:355–377.

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

[Griffiths et al., 2010] Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364.

[Griffiths and Ghahramani, 2006] Griffiths, T. L. and Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In Schölkopf, B., Platt, J., and Hofmann, T., editors, *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA. MIT Press.

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

[Griffiths et al., 2008] Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). Bayesian models of cognition. In Sun, R., editor, *Cambridge handbook of computational cognitive modeling*, pages 59–100. Cambridge University Press, Cambridge.

[Griffiths et al., 2011] Griffiths, T. L., Sanborn, A. N., Canini, K. R., Navarro, D. J., and Tenenbaum, J. B. (2011). Nonparametric Bayesian models of category learning. In Pothos, E. M. and Wills, A. J., editors, *Formal approaches in categorization*, pages 173–198. Cambridge University Press, Cambridge, UK.

[Griffiths and Tenenbaum, 2005] Griffiths, T. L. and Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51:354–384.

[Grossberg, 1987] Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11:23–63.

[Hall, 1991] Hall, G. (1991). *Perceptual and Associative Learning*. Clarendon, Oxford.

[Hansen et al., 2005] Hansen, L. K., Ahrendt, P., and Larsen, J. (2005). Towards cognitive component analysis. In Pattern Recognition Society of Finland, Finnish Artificial Intelligence Society, F. C. L. S., editor, *AKRR'05 -International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*. Pattern Recognition Society of Finland, Finnish Artificial Intelligence Society, Finnish Cognitive Linguistics Society.

[Hebb, 1949] Hebb, D. O. (1949). *The Organization of Behavior: A Newuropsychological Theory*. John Wiley & Sons, NY, NY.

[Hebb, 1958] Hebb, D. O. (1958). *A textbook of psychology*. W. B. Saunders Company, Philadelphia, USA.

[Hecht et al., 1942] Hecht, S., Shlaer, S., and Pirenne, M. H. (1942). Energy, quanta, and vision. *The Journal of General Physiology*, 25:819–840.

[Helmholtz, 1866] Helmholtz, H. v. (1866). Concerning the perceptions in general. In Southall, J. P. C., editor, *Treatise on physiological optics*, volume 3. Dover, New York.

[Hinton and Salakhutdinov, 2006] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313:504–507.

[Hochberg and McAlister, 1953] Hochberg, J. and McAlister, E. (1953). A quantitative approach to figural "goodness". *Journal of Experimental Psychology*, 46(5):361–364.

[Hoffman and Richards, 1985] Hoffman, D. D. and Richards, W. A. (1985). Parts in recognition. *Cognition*, 18:65–96.

[Hu et al., 2012] Hu, Y., Zhai, K., Williamson, S., and Boyd-Graber, J. (2012). Modeling images using transformed Indian buffet processes. In *International Conference of Machine Learning*.

[Hubel and Wiesel, 1962] Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154.

[Hyvarinen, 1999] Hyvarinen, A. (1999). Fast and robust fixed-point algorithm for independent component analysis. In *IEEE Trans. Neural Networks*, volume 10, pages 626–634.

[Hyvarinen et al., 2001] Hyvarinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley, New York.

[Joho et al., 2011] Joho, D., Tipaldi, G. D., Engelhard, N., Stachniss, C., and Burgard, W. (2011). Unsupervised scene analysis and reconstruction using nonparametric Bayesian models. *Robotic and Autonomous Systems*, 59(5):319–328.

[Jolicoeur, 1985] Jolicoeur, P. (1985). The time to name disoriented natural objects. *Memory & Cognition*, 13(4):289–303.

[Jones and Love, 2011] Jones, M. and Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34:169–231.

[Jordan, 1986] Jordan, M. I. (1986). An introduction to linear algebra in parallel distributed processing. In *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations*, pages 365–421. MIT Press, Cambridge, MA.

[Jordan, 2010] Jordan, M. I. (2010). Bayesian nonparametric learning: Expressive priors for intelligent systems. In *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, pages 167–186. College Publications.

[Kahneman et al., 1982] Kahneman, D., Slovic, P., and Tversky, A., editors (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press, Cambridge.

[Kalish et al., 2010] Kalish, C. W., Rogers, T. T., Lang, J., and Zhu, X. (2010). Can semi-supervised learning explain incorrect beliefs about categories? *Cognition*, 120(1):106–118.

[Kanizsa, 1979] Kanizsa, G. (1979). *Organization in Vision: Essays on Gestalt Perception*. Praeger Publishers, New York, NY.

[Kemp and Tenenbaum, 2009] Kemp, C. and Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, 116(1):20–58.

[Kirkham et al., 2002] Kirkham, N. Z., Slemmer, J. A., and Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, 83:B35–B42.

[Kosslyn, 1995] Kosslyn, S. M. (1995). Mental imagery. In *Visual Cognition: An Invitation to Cognitive Science*, pages 267–296. MIT Press, Cambridge, MA.

[Krumhansl, 1978] Krumhansl, C. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, 85:450–463.

[Kruschke, 1992] Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44.

[Kruschke, 2006] Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, 113:677–699.

[Lee and Navarro, 2002] Lee, M. D. and Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, 9(1):43–58.

[Leeuwenberg, 1971] Leeuwenberg, E. L. J. (1971). A perceptual coding language for visual and auditory patterns. *American Journal of Psychology*, 84:307–349.

[Lin and Murphy, 1997] Lin, E. L. and 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, 2007] Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3):232–257.

[Love et al., 2004] Love, B. C., Medin, D. L., and Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111:309–332.

[Luce, 1959] Luce, R. D. (1959). *Individual choice behavior*. John Wiley, New York.

[Lunn, 1948] Lunn, J. H. (1948). Chick sexing. *American Scientist*, 36(2):280–287.

[Mach, 1959] Mach, E. (1914/1959). *The analysis of sensations*. Open Court, Chicago.

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

[Marr, 1982] Marr, D. (1982). *Vision*. W. H. Freeman, San Francisco, CA.

[Marr and Nishihara, 1978] Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London*, 200:269–294.

[McCarthy et al., 2012] McCarthy, J. D., Cordeiro, D., and Caplovitz, G. P. (2012). Local form-motion interactions influence global form perception. *Attention, Perception, & Psychophysics*, 74:816–823.

[McClelland and Rumelhart, 1981] McClelland, J. L. and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. an account of basic findings. *Psychological Review*, 88(5):375–407.

[Medin et al., 1982] Medin, D. L., Altom, M. W., Edelson, S. M., and Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(1):37–50.

[Medin et al., 1993] Medin, D. L., Goldstone, R. L., and Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100:254–278.

[Medin and Schaffer, 1978] Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85:207–238.

[Medin and Schwanenflugel, 1981] Medin, D. L. and Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 7:355–368.

[Mestre, 2005] Mestre, J. P. (2005). *Transfer of learning from a modern multidisciplinary perspective*. Information Age Pub Inc.

[Miller et al., 2009] Miller, K. T., Griffiths, T. L., and Jordan, M. I. (2009). Nonparametric latent feature models for link prediction. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1276–1284.

[Minka, 2001] Minka, T. (2001). Automatic choice of dimensionality for PCA. In *Advances in Neural Information Processing Systems 13*, pages 598–604. MIT Press, Cambridge, MA.

[Minsy and Papert, 1969] Minsy, M. L. and Papert, S. A. (1969). *Perceptrons*. MIT Press, Cambridge, MA.

[Misiak and Sexton, 1966] Misiak, H. and Sexton, V. S. (1966). *History of Psychology - An Overview*. Grune & Stratton, New York.

[Murphy and Medin, 1985] Murphy, G. L. and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92:289–316.

[Navarro and Perfors, 2010] Navarro, D. J. and Perfors, A. F. (2010). Similarity, feature discovery, and the size principle. *Acta Psychologica*, 133:256–268.

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

[Neisser, 1967] Neisser, U. (1967). *Cognitive Psychology*. Appleton-Century-Crofts, New York, NY.

[Nosofsky, 1984] Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10:104–114.

[Nosofsky, 1986] Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57.

[Nosofsky, 1991] Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, 17(1):3–27.

[Nosofsky, 2011] Nosofsky, R. M. (2011). The generalized context model: an exemplar model of classification. In Pothos, E. M. and Wills, A. J., editors, *Formal Approaches in Categorization*, pages 18–39. Cambridge University Press, Cambridge, UK.

[Oaksford and Chater, 1998] Oaksford, M. and Chater, N., editors (1998). *Rational models of cognition*. Oxford University Press, Oxford.

[Olshausen and Field, 1996] Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609.

[Orban et al., 2008] Orban, G., Fiser, J., Aslin, R. N., and Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, 105(7):2745–2750.

[Osherson et al., 1990] Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., and Shafir, E. (1990). Category-based induction. *Psychological Review*, 97(2):185–200.

[Palmer, 1975] Palmer, S. E. (1975). Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. In *Explorations in cognition*, pages 279–307. W. H. Freeman, San Francisco.

[Palmer, 1977] Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, 9:441–474.

[Palmer, 1978] Palmer, S. E. (1978). Fundamental aspects of cognitive representation. In *Cognition and categorization*, pages 250–303. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

[Palmer, 1983] Palmer, S. E. (1983). The psychology of perceptual organiziation: A transformational approach. In *Human and machine vision*, pages 269–339. Academic Press, New York.

[Palmer, 1989] Palmer, S. E. (1989). Reference frames in the perception of shape and orientation. In *Object perception: Structure and process*, pages 121–163. Erlbaum, Hillsdale, NJ.

[Palmer, 1999] Palmer, S. E. (1999). *Vision Science*. MIT Press, Cambridge, MA.

[Palmer and Kimchi, 1986] Palmer, S. E. and Kimchi, R. (1986). The information processing approach to cognition. In Knapp, T. J. and Robertson, L. C., editors, *Approaches to cognition: Contrasts and controversies*, pages 37–77. Erlabum, Hillsdale, NJ.

[Pearl, 1988] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, San Francisco, CA.

[Perruchet and Vinter, 1998] Perruchet, P. and Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, 39:246–263.

[Pevtzow and Goldstone, 1994] Pevtzow, R. and Goldstone, R. L. (1994). Categorization and the parsing of objects. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pages 712–722, Hillsdale, NJ. Lawrence Erlbaum Associates.

[Pitman, 2002] Pitman, J. (2002). Combinatorial stochastic processes. Notes for Saint Flour Summer School.

[Pitt, 2008] Pitt, D. (2008). Mental representation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition.

[Pomerantz, 1978] Pomerantz, J. R. (1978). Are complex visual features derived from simple ones? In *Formal theories of visual perception*, pages 217–229. Wiley, New York.

[Pomerantz, 2003] Pomerantz, J. R. (2003). Wholes, holes, and basic features in vision. *Trends in Cognitive Sciences*, 7(11):471–473.

[Pomerantz and Kubovy, 1986] Pomerantz, J. R. and Kubovy, M. (1986). Theoretical approaches to perceptual organization: Simplicity and likelihood principles. In Boff, K. R., Kaufman, L., and Thomas, J. P., editors, *Handbook of perception and human performance: Volume II Cognitive processes and performance*, pages 1–45. Wiley, New York.

[Pomerantz et al., 1977] Pomerantz, J. R., Sager, L. C., and Stoever, R. J. (1977). Perception of wholes and of their component parts: Some configural superiority effects. *Journal of Experimental Psychology: Human Perception and Performance*, 3(3):422–435.

[Posner and Keele, 1968] Posner, M. I. and Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77:353–363.

[Pothos and Bailey, 2009] Pothos, E. M. and Bailey, T. M. (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the generalized context model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(4):1062–1080.

[Pylyshyn, 2002] Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and Brain Sciences*, 25:157–238.

[Rasmussen and Ghahramani, 2001] Rasmussen, C. E. and Ghahramani, Z. (2001). Occam's razor. In *Advances in Neural Information Processing Systems 13*, pages 294–300. MIT Press, Cambridge, MA.

[Reed, 1972] Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3:393–407.

[Restle, 1959] Restle, F. (1959). A metric and an ordering on sets. *Psychometrika*, 24:207–220.

[Riesenhuber, 2009] Riesenhuber, M. (2009). Object categorization in man, monkey, and machine: Some answers and some open questions. In Dickinson, S. J., Leonardis, A., Schiele, B., and Tarr, M. J., editors, *Object Categorization: Computer and Human Vision Perspectives*, pages 216–240. Cambridge University Press, New York, NY.

[Rock, 1973] Rock, I. (1973). *Orientation and form.* Macmillan, New York.

[Roediger et al., 2001] Roediger, H. L., Watson, J. M., McDermott, K. B., and Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin and Review*, 8:385–407.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by back-propagating errors. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel distributed processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA.

[Rumelhart and McClelland, 1986] Rumelhart, D. E. and McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volumes 1 and 2.* MIT Press, Cambridge, MA.

[Rumelhart and Zipser, 1985] Rumelhart, D. E. and Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, 9:75–112.

[Russell, 1986] Russell, S. J. (1986). A quantitative analysis of analogy by similarity. In *Proceedings of the National Conference on Artificial Intelligence*, pages 284–288. AAAI, Philadelphia, PA.

[Rust and Stocker, 2010] Rust, N. C. and Stocker, A. A. (2010). Ambiguity and invariance: two fundamental challenges for visual procesing. *Current Opinion in Neurobiology*, 20:382–388.

[Saffran et al., 1996] Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science*, 274:1926–1928.

[Saffran et al., 1997] Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., and Barrueco, S. (1997). Incidental language learning listening (and learning) out of the corner of your ear. *Psychological Science*, 8(2):101–105.

[Sanborn et al., 2010] Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4):1144–1167.

[Schyns et al., 1998] Schyns, P. G., Goldstone, R. L., and Thibaut, J. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, 21:1–54.

[Schyns and Murphy, 1994] Schyns, P. G. and Murphy, G. (1994). The ontogeny of part representation in object concepts. In *The Psychology of Learning and Motivation*, volume 31, pages 305–354. Academic Press, San Diego.

[Schyns and Rodet, 1997] Schyns, P. G. and Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23:681–696.

[Selfridge, 1955] Selfridge, O. G. (1955). Pattern recognition and modern computers. In *Proceedings of the March 1-3, 1955, Western Joint Computer Conference*, AFIPS '55 (Western), pages 91–93, New York, NY, USA. ACM.

[Selfridge and Neisser, 1960] Selfridge, O. G. and Neisser, U. (1960). Pattern recognition by machine. *Scientific American*, 203:60–68.

[Serre and Poggio, 2010] Serre, T. and Poggio, T. (2010). A neuromorphic approach to computer vision. *Communications of the ACM*, 53(10):54–61.

[Shepard and Arabie, 1979] Shepard, R. and Arabie, P. (1979). Additive clutering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86:87–123.

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

[Shepard et al., 1961] Shepard, R. N., Hovland, C. I., and Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, 75. 13, Whole No. 517.

[Shiffrin and Lightfoot, 1997] Shiffrin, R. M. and Lightfoot, N. (1997). Perceptual learning of alphanumeric-like characters. In *The Psychology of Learning and Motivation*, volume 36, pages 45–82. Academic Press, San Diego.

[Shultz and Sirois, 2008] Shultz, T. R. and Sirois, S. (2008). Computational models of developmental psychology. In Sun, R., editor, *The Cambridge handbook of computational psychology*, pages 451–476. Cambridge University Press, New York.

[Skinner, 1977] Skinner, B. F. (1977). Why I am not a cognitive psychologist. *Behaviorism*, 5:1–10.

[Smith, 2005] Smith, L. B. (2005). Shape: A developmental product. In Carlson, L. and van der Zee, E., editors, *Fucntional features in spatial concepts*, pages 235–255. Oxford University Press, Oxford, England.

[Solan et al., 2005] Solan, Z., Horn, D., Ruppin, E., and Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, 102(33):11629–11634.

[Sowden et al., 2000] Sowden, P. T., Davies, I. R. L., and Roling, P. (2000). Perceptual learning of the detection of features in X-ray images: A functional role for improvements in adults' visual sensitivity? *Journal of Experimental Psychology: Human Perception and Performance*, 26:379–390.

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

[Stromsten, 2002] Stromsten, S. B. (2002). *Classification learning from both classified and unclassified examples*. PhD thesis, Stanford University.

[Sudderth and Jordan, 2009] Sudderth, E. and Jordan, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1585–1592.

[Tanaka and Taylor, 1991] Tanaka, J. W. and Taylor, M. E. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 15:121–149.

[Tarr and Pinker, 1989] Tarr, M. J. and Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21(2):233–282.

[Tarr et al., 1998] Tarr, M. J., Williams, P., Hayward, W. G., and Gauthier, I. (1998). Three-dimensional object recognition is viewpoint-dependent. *Nature Neuroscience*, 1:275–277.

[Teh et al., 2004] Teh, Y., Jordan, M. I., Beal, M., and Blei, D. (2004). Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA.

[Tenenbaum, 2000] Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In Solla, S. A., Leen, T. K., and Muller, K.-R., editors, *Advances in Neural Information Processing Systems 12*, pages 59–65. MIT Press, Cambridge, MA.

[Tenenbaum and Griffiths, 2001] Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24:629–641.

[Tenenbaum et al., 2011] Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331:1279–1285.

[Titsias, 2008] Titsias, M. (2008). The infinite gamma-poisson feature model. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*, pages 1513–1520. MIT Press, Cambridge, MA.

[Tolman, 1948] Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208.

[Trueblood and Busemeyer, 2011] Trueblood, J. S. and Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. *Cognitive Science*, 35(8):1518–1552.

[Tversky, 1977] Tversky, A. (1977). Features of similarity. *Psychological Review*, 84:327–352.

[Ullman, 2007] Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences*, 11(2):58–64.

[Uttal, 1971] Uttal, W. R. (1971). The psychobiological silly season - or - what happens when neurophysiological data become psychological theories. *Journal of General Psychology*, 84:151–166.

[van der Helm, 2000] van der Helm, P. A. (2000). Simplicity versus likelihood in visual perception: From surprisals to precisals. *Psychological Bulletin*, 126(5):770–800.

[van der Helm, 2011] van der Helm, P. A. (2011). Bayesian confusions surrounding simplicity and likelihood in perceptual organization. *Acta Psychologica*, 138:337–346.

[Vandist et al., 2009] Vandist, K., De Schryver, M., and Rosseel, Y. (2009). Semisupervised category learning: The impact of feedback in learning the information-integration task. *Attention, Perception, & Psychophysics*, 71(2):328–341.

[Venkataraman, 2001] Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27:352–372.

[Watson, 1913] Watson, J. B. (1913). Psychology as the Behaviorist views it. *Psychological Review*, 20(2):158–177.

[Wertheimer, 1938] Wertheimer, M. (1923/1938). Laws of organization in perceptual forms. In Ellis, W., editor, *A Source Book of Gestalt Psychology*, pages 71–88. Routledge and Kegan Paul, London.

[Williamson et al., 2010] Williamson, S., Orbanz, P., and Ghahramani, Z. (2010). Dependent Indian buffet processes. In *Proceedings of the 13th Conference on Artificial Intelligence and Statistics*, pages 924–931.

[Winston, 1975] Winston, P. (1975). Learning structural descriptions form examples. In Winston, P., editor, *The psychology of computer vision*, pages 157–209. McGraw-Hill, New York, NY.

[Wisniewski and Medin, 1994] Wisniewski, E. J. and Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, 18:221–281.

[Wood and Griffiths, 2007] Wood, F. and Griffiths, T. L. (2007). Particle filtering for nonparametric Bayesian matrix factorization. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 1513–1520. MIT Press, Cambridge, MA.

[Wood et al., 2006] Wood, F., Griffiths, T. L., and Ghahramani, Z. (2006). A nonparametric Bayesian method for inferring hidden causes. In *Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence (UAI '06)*, pages 536–543.

[Xu and Tenenbaum, 2007] Xu, F. and Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2):245–272.

[Yildirim and Jacobs, 2012] Yildirim, I. and Jacobs, R. A. (2012). A rational analysis of the acquisition of multisensory representations. *Cognitive Science*, 36(2):305–332.

[Zaki and Nosofsky, 2007] Zaki, S. R. and Nosofsky, R. M. (2007). A high-distortion enhancement effect in the prototype-learning paradigm: Dramatic effects of category learning during test. *Memory & Cognition*, 35(8):2088–2096.

[Zeigenfuse and Lee, 2010] Zeigenfuse, M. D. and Lee, M. D. (2010). Finding the features that represent stimuli. *Acta Psychologica*, 133:283–295.

[Zhu, 1999] Zhu, S.-C. (1999). Embedding Gestalt laws in Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1170–1187.

[Zhu and Goldberg, 2009] Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.

[Zhu et al., 2007] Zhu, X., Rogers, T., Qian, R., and Kalish, C. (2007). Humans perform semi-supervised classification too. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, Menlo Park, CA. AAAI Press.

Appendix A

Inference for 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 Equation 3.3) for binary input data (Experiments 1, 2 and 4 of Chapter 4) and the linear-Gaussian likelihood function (given by Equation 3.5) for grayscale input data (Experiment 3 of Chapter 4). The feature image matrix (the image of each feature when instantiated) is $\mathbf{Y}$. To infer the feature images 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 images 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, Griffiths, and Ghahramani (2006) for the noisy-OR model and Griffiths and Ghahramani (2006; 2011) for the linear-Gaussian model. An alternative to Gibbs sampling is variational inference [Doshi-Velez et al., 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

$$P(\mathbf{Z}, \mathbf{Y}|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{Y}, \mathbf{Z})P(\mathbf{Y})P(\mathbf{Z}). \tag{A.1}$$

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 explicitly 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 and Ghahramani, 2006]

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

where σ_X^2 is the expected variance of the observed objects, $\mathbf{B}^T$ denotes the transpose of matrix $\mathbf{B}$, and $\mathrm{tr}(\mathbf{B})$ denotes the trace of the matrix $\mathbf{B}$ (the sum of its elements on the primary diagonal).[1] 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 and Ghahramani, 2006]:

$$E[\mathbf{Y}|\mathbf{X},\mathbf{Z}] = \left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_Y^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{X}. \tag{A.3}$$

For each experiment, we ran the sampler for at least 500 samples and took the inferred $\hat{\mathbf{Z}}$ (and $\hat{\mathbf{Y}}$) to be the mode of the distribution given by the Gibbs sampler after a burn-in of 50 iterations (removing the 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 $P(\mathbf{x}_{\mathrm{test}}|\mathbf{X})$, the probability of observing new objects $\mathbf{x}_{\mathrm{test}}$ after observing the object set $\mathbf{X}$:

$$P(\mathbf{x}_{\mathrm{test}}|\mathbf{X}) = \sum_{\mathbf{Z}_{\mathrm{test}},\mathbf{Z},\mathbf{Y}} P(\mathbf{x}_{\mathrm{test}}|\mathbf{Z}_{\mathrm{test}},\mathbf{Y})P(\mathbf{Z},\mathbf{Y}|\mathbf{X}) \tag{A.4}$$

$$\approx P(\mathbf{x}_{\mathrm{test}}|\hat{\mathbf{Z}}_{\mathrm{test}},\hat{\mathbf{Y}}) \tag{A.5}$$

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

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 $P(\mathbf{x}_{\mathrm{test}}|\hat{\mathbf{Z}}_{\mathrm{test}},\hat{\mathbf{Y}})^\gamma$, where γ is a parameter fit by minimizing the squared distance between

[1] 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.

the model's responses and the human data, and then renormalize over each type of object.[2] 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 of Chapter 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, Griffiths, and Ghahramani, 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 of Chapter 4), 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 small 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 of Chapter 4 simulations (each image in Experiments 1 and 2 of Chapter 4 is $86 \times 146 = 12556$-dimensional), the Gibbs sampler gets stuck in local minima and thus, simulated annealing [Geman and 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[3] 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 of Chapter 4, we found that the

[2]Due to the high dimensionality of the input data for Experiments 1, 2, and 3 of Chapter 4, the model predictions were expressed computationally as log probabilities. We transformed the log probabilities by $\exp\{\gamma|\log P(\mathbf{x}_{\text{test}}|\hat{\mathbf{Z}}_{\text{test}}, \hat{\mathbf{Y}})|\}$. This results in the same transformation on the probabilities defined previously.

[3]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 $P(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$.

Gibbs sampler converged to the reported feature representation by approximately iteration 400.

For Experiments 1 and 2 of Chapter 4, the Gibbs sampler was run for 1000 iterations and the best-fitting γ was 1.5677×10^3. For Experiment 3 of Chapter 4, 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 of Chapter 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 Chapter 6, 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 Equation 3.4 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 otherwise. The results reported in Chapter 4 use $\phi = 0.15$ and $\rho = 0.999$ (the other parameters were the same).

A.1 Modeling human responses using a prototype and exemplar model

To model the results of Experiments 1 and 2 of Chapter 4 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

$$d(\mathbf{x}_{\text{test}}, \mathbf{x}) = \sqrt{(\mathbf{x}_{\text{test}} - \mathbf{x})^T(\mathbf{x}_{\text{test}} - \mathbf{x})}, \tag{A.6}$$

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 ($\eta^p_{\mathbf{x}_{\text{test}},\mathbf{X}}$) for the prototype model was

$$\eta^p_{\mathbf{x}_{\text{test}},\mathbf{X}} = \exp\left\{-\kappa d(\mathbf{x}_{\text{test}}, \mathbf{x}_{\text{mean}})\right\} \tag{A.7}$$

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

$$\eta^e_{\mathbf{x}_{\text{test}},\mathbf{X}} = \sum_{\mathbf{x}\in\mathbf{X}} \exp\left\{-\kappa d(\mathbf{x}_{\text{test}}, \mathbf{x})\right\} \tag{A.8}$$

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 of Chapter 4, 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.2 Modeling human responses using principal component analysis and independent component analysis

To model the results of Experiments 1 and 2 of Chapter 4 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.[4] Then, we convert the average reconstruction error to the results from Experiments 1 and 2 of Chapter 4 using the same transformation as was used in the previous section for the IBP. For Experiments 1 and 2 of Chapter 4, we found that the best-fitting parameter value was $\gamma = 0.3082$ for PCA and $\gamma = 0.0726$ for ICA.

[4]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, Karhunen, and Oja (2001) for details.

Appendix B

The Transformed Indian Buffet Process

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

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

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

$$P(\mathbf{x}_{\text{test}}|\mathbf{X}) \approx P(\mathbf{x}_{\text{test}}|\hat{\mathbf{z}}_{\text{test}}, \hat{\mathbf{r}}_{\text{test}}(\hat{\mathbf{Y}})), \tag{B.1}$$

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

Except for the simulations reported in the beyond transformation independence and learning which transformations apply subsections, we used a uniform distribution on all possible translations as our prior on feature transformations ($r_{nk} \sim U(\{1, \ldots, D_1\} \times \{0, \ldots, D_2\})$,

where D_1 and D_2 are the size of the dimensions of the image and r_{nk} is a parameter specifying the translation to the right and down). We assumed the image was a torus, so any part of the image that are transformed to go beyond the dimensions are wrapped back to the beginning. We use Gibbs sampling for inference, drawing samples from $P(\mathbf{Z}, \mathbf{Y}, \mathbf{R}|\mathbf{X})$.

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

$$p(z_{nk}|\mathbf{Z}_{-(nk)}, \mathbf{R}_{-(nk)}, \mathbf{Y}, \mathbf{X}) = \sum_{r_{nk}} p(z_{nk}|\mathbf{Z}_{-(nk)}, \mathbf{R}, \mathbf{Y}, \mathbf{X}) p(r_{nk}) \tag{B.2}$$

$$\propto \sum_{r_{nk}} p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{r}_n(\mathbf{Y})) p(z_{nk}|\mathbf{Z}_{-(nk)}) p(r_{nk}) \tag{B.3}$$

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

$$p(r_{nk}|z_{nk} = 1, \mathbf{Z}_{-(nk)}, \mathbf{R}_{-(nk)}, \mathbf{Y}, \mathbf{X}) \propto p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{Y}, \mathbf{R}) p(r_{nk}) \tag{B.4}$$

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

Second, we sample the number of new features. To derive this sampler, we adapt a technique used by Wood et al. (2006) in the following derivation,

$$p(K_n^{\text{new}}|\mathbf{x}_n, \mathbf{Z}_{n,1:(K+K_n^{\text{new}})}, \mathbf{Y}, \mathbf{R}) \propto p(\mathbf{x}_n|\mathbf{Z}^{\text{new}}, \mathbf{Y}, K_n^{\text{new}}) P(K_n^{\text{new}}), \tag{B.5}$$

where $\mathbf{Z}^{\text{new}}$ includes the K_n^{new} extra columns for the new features (which is 1 for row n, but 0 otherwise). To compute the first term, $p(\mathbf{x}_n|\mathbf{Z}^{\text{new}}, \mathbf{Y}, K_n^{\text{new}})$, we need to sum over possible feature images and transformations. To simplify this calculation, we assume that the first time a feature is sampled, it is not transformed (this is justified because one of the transformations and the feature image are statistically unidentifiable). Without transformations, this is equivalent to summing over the possible feature images in the noisy-OR IBP, which was derived by Wood et al. (2006) to be

$$p(K_n^{\text{new}}|\mathbf{x}_n, \mathbf{Z}_{n,1:(K+K_n^{\text{new}})}, \mathbf{Y}, \mathbf{R}) \propto \frac{\alpha^{K_n^{\text{new}}} e^{-\alpha}}{K_n^{\text{new}}!} \left(1 - (1-\epsilon)(1-\lambda)^{\mathbf{z}_n \mathbf{r}_n(\mathbf{y}_d)} (1-p\lambda)^{K_n^{\text{new}}}\right) \tag{B.6}$$

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

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

$$P(\mathbf{x}_{N+1}|\mathbf{X}) \approx \frac{1}{T}\sum_{t=1}^{T} P(\mathbf{x}_{N+1}|\mathbf{Z}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{R}^{(t)}). \tag{B.7}$$

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

For the simulations described in Chapter 5, we use the feature representation from the last sample from the Gibbs sampler after 1000 iterations. The exponentiated Luce choice rule parameter was $\phi = 0.05$, and the parameter values were initialization to the following values: $\alpha = 0.8$, $\epsilon = 0.05$, $\lambda = 0.99$, and $\phi = 0.4$. Both the tIBP and IBP use these same parameter values for the simulations reported in the Learning Invariant vs. "Variant" Features section of Chapter 5.

Appendix C

Translations affecting scalings

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

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

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

The parameters $\alpha, \epsilon, \lambda$, and ϕ were set to $2, 0.005, 0.995$, and 0.25 respectively, and the Gibbs sampler was run for 1000 iterations. k_0 was set to 40 and μ_S was set to $[1.2, 1.2]^T$. μ_{T0} was set to $[25, 25]^T$ (where parts falling outside the dimensions of the images were shifted back to the beginning). The posterior distribution over r_{nk} and Σ is fairly broad. So rather than use the last value to form our predictions (as we do for the other simulations), the predictions were made by averaging over the samples using a burn-in of 200 and taking every 1 out of 10 samples (*thinning*). The exponentiated Luce rule with parameter $\phi = 0.031$ was used, except the unnormalized probabilities were not renormalized after exponentiating.

Appendix D

Learning Invariance Type using the tIBP

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

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

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

The parameters were set to $\alpha = 2$, $\epsilon = 0.01$, $\lambda = 0.99$, $\phi = 0.5$, and $\pi = 0.5$ (with $\phi = 0.005$ for the exponentiated Luce choice rule). The predictions shown were made after running the Gibbs sampler for 1000 iterations and the images were 38 by 38 pixels each.

Appendix E

Incorporating category information

In Chapter 6, we presented two methods for incorporating categorization information into our framework: as part of the observable properties of an object, and representing it directly through the Indian Buffet Franchise (IBF). For the first method, we appended $c = 35$ bits per category to each image. So, each image had $2c = 70$ extra bits. To encode that an image was in category 1, the first c extra bits were set to 1 and the last c bits were set to zero (and vice versa for category 2). Except for setting $\phi = 0.125$, we used the same parameter values as used by Appendix A: $\alpha = 2$, $\epsilon = 0.01$, and $\lambda = 0.99$. The Gibbs sampler (without using simulated annealing) converged to the features shown in Figure 6.3ab very quickly (within a few iterations, but the Gibbs sampler was still run for 2000 iterations).

Inference for the IBF proceeds as described by Wood et al. (2006), except that when the IBP creates new features, it draws the images for those features from the shared CRP. We refer the reader to Neal (2000) for a thorough description of inference for the CRP. Additionally, Gibbs updates for $y_{kd}^{(0)}$ are performed as described by Wood et al.(2006) except that all images (regardless of their category) that have a feature whose image is assigned to $\mathbf{y}_k^{(0)}$ should be included in the product over likelihood terms. Note that when a pixel of $y_{kd}^{(0)}$ is resampled, the corresponding $y_{k'd}^{(a)}$ should be set to $y_{kd}^{(0)}$ for any category's feature image k' that is assigned to feature image k of the CRP. The feature images of the individual categories otherwise do not get change from iteration to iteration of inference (i.e., they are deterministically set depending on the shared feature image matrix $\mathbf{Y}^{(0)}$). The results described in Chapter 6 use the same parameters as the other technique for incorporating category information, and $\beta = 2$.

Appendix F

Learning Features for the tIBP Incrementally

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

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

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

Intuitively, we have decomposed the posterior into three terms (described from left to right): the likelihood of $\mathbf{x}_n$ given the proposed feature representation (which is given by Equation 3.3), the probability of the proposed feature ownership matrix given the feature ownership matrix we had after $n-1$ objects (which is simply the restaurant sampling scheme for generating the IBP), and the posterior distribution of the feature representation after observing

$n-1$ objects. The last term is the posterior distribution over feature representations, but for $n-1$ objects instead of n objects, and thus we can sequentially update our posterior distribution given the next object using the posterior distribution given one fewer object.

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

For the simulations reported in Chapter 6, $T = 100$ and the parameters $\alpha, \epsilon, \lambda$, and ϕ were initialized to $2, 0.001, 0.999$ and 0.4 for both the Gibbs sampler (run for 1000 iterations) and particle filter, and the model was given five examples of each type of object. Each image was 30×30 pixels. The simulated annealing strategy was applied for the Gibbs sampler, which converged quickly to the two feature solution in both object orders. After observing all 15 objects, all particles contained the same feature representation.