Within-Category Feature Correlations and the Curse of Dimensionality

This thesis is submitted in partial fulfilment of the

Honours degree of Bachelor of Psychology

School of Psychology

University of Adelaide

Word Count: 10,322

Table of Contents

List of Tables

List of Figures

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the School to restrict access for a period of time.

X

October, 2017

Acknowledgements

I would like to thank my supervisor, Dr. Amy Perfors. Your guidance throughout this year has been invaluable. I have learnt a great deal in this short amount of time thanks to your vast knowledge and feedback. I would also like to thank my family and my partner. Thanks for putting up with all my work when you would have preferred that we were "Netflix and chilling".

Abstract

Within the domain of category learning, the curse of dimensionality states that as categories acquire more features, the size of the feature space and thus the number of examples necessary to adequately learn the category grows rapidly. As a result, category learning should be a highly difficult task. However, people learn to classify categories with ease. The primary aim of the current study was to address how people overcome this problem by determining if they were attuned to information about which features were correlated with each other, and whether or not they used this information as an indication that those features were relevant for categorisation. In theory, this would then allow for the detection of natural category family resemblance structure in highly dimensional environments, and subsequently, allow human learners to overcome the curse of dimensionality. In addition to this primary aim, and under the assumption that people were attuned to this correlational structure, a number of secondary hypotheses were also proposed. These hypotheses assessed both feature and correlation learning, evaluated relative category learning improvement as a function of dimensionality between conditions, as well as compared human performance to a Naïve Bayes model that was inherently incapable of detecting any correlational structure. The results of the current study suggest that people do not utilise within-category feature correlational structure as a heuristic for category predictive feature detection. However, these findings were contingent on a number of methodological shortcomings present in the current study. Directions for future research are proposed that provide clear methodological changes in category structure which may mitigate the lack of support found for the proposed hypotheses.

1. Introduction

People do not experience the world as a structure-less and unpredictable wave of stimuli but as a systemised and somewhat predictable set of events. We have both categories (the class of objects in the real world) and concepts (mental representations) to thank for this coherence. Structures that act as heuristics, allowing us to take advantage of the knowledge that we have gained from prior experience (Chin-Parker & Ross, 2002). Any time a person wants to make a categorisation decision, make a prediction about a missing feature, or communicate an idea, knowledge of category membership is essential to effectively accomplish these actions. This knowledge includes not just being able to assign labels to categories, but also knowing how the features relate to the category (Ross & Spalding, 1994). For instance, knowing that balls are spherical makes it easier to decide that a novel spherical item may also be a ball.

Categories are a way of carving up the rich correlational structure of the world. For example, *dogs* are a useful category because things that have four legs, fur, wagging tails and bark, form a coherent set. Because they capture this correlational structure, categories and concepts thus act as a cognitive shortcut, allowing human learners to make predictions in the face of limited cognitive resources and constant sensory overload; they are thus a means of achieving cognitive economy (Rosch 1999; Garner, 1974).

How do categories accomplish this? One possibility is that their utility is derived from the way that they capture both within and between category feature correlations in the world. For instance, it may be that the category *train* is useful because trains tend to be long and loud (within-category correlations) but also because they, unlike cars, run on tracks and hold hundreds of people (between-category correlations). Although it has been well established that categories tend to capture between-category feature correlations (Ahn, 1998; Ahn, Kim, Lassaline, & Dennis, 2000; Ahn & Medin, 1992; Medin, Wattenmaker, & Hampson, 1987; Rehder & Burnett, 2004; Rosch & Mervis, 1975; Waldmann, Holyoak, & Fratianne, 1995), the

representation of within-category feature correlations is still seen as contentious by some (Chin-Parker & Ross, 2002; Rehder & Grittiths, 2005). In spite of this, a relatively large body of evidence has shown support for the representation and use of such correlations in both feature inference and classification tasks (Anderson & Finchman, 1996; Crawford, Hayes-Roth & Hayes-Roth, 1977; Huttenlocher, & Hedges, 2006; Hammond, McClelland, & Mumpower, 1980; Klayman, 1988; Malt & Smith, 1984; Medin, Altom, Edelson & Frecko, 1982; Wattenmaker, 1991).

The implications of such correlational structure is important, as representational assumptions made by several models assume class conditional independence between the features of an exemplar (Anderson 1990, 1991; Anderson & Matessa, Collins & Loftus, 1975; McCloskey & Glucksberg, 1979; Reed, 1972; Smith, Shoben, & Rips, 1974; Tversky, 1977), meaning that the occurrence of one feature, such as the presence of a tail on a dog, does not affect the probability of another feature occurring, such as a dog barking, thus, each feature exists independently of the other (Russell & Peter, 2002). Even know these assumptions are often made on the basis of mathematical and computational tractability, they may not be analogous to how people perceive world structure, a question that has yet to be addressed in the literature. If this question is to be resolved, it that may be able to account for some of the disparities seen between computational and human model performance in categorisation tasks. A problem that this thesis intends to address.

Another issue that arises if people are tracking and relying on information about features and feature correlations when learning and using categories, is that this leaves them susceptible to the curse of dimensionality. The curse of dimensionality is a well-studied phenomenon in statistics, machine learning, and computer science, and is based on the fact that additional features increase processing complexity at an exponential rate (Bellman, 1961; Donoho, 2000; Keogh & Mueen, 2011; Verleysen & Francois, 2005). For instance, within

computer science and machine learning, as the number of features in a model grows larger, the examples from the training data grow sparser relative to the feature space; as a result, exponentially more training data is required for each additional feature added to the model (Verleysen & François, 2005; Keogh & Mueen, 2011).

In the field of cognitive science, a similar problem arises in the domain of category learning. As we consider categories with increasing numbers of features, the size of the feature space grows rapidly. In the real world, most categories have many available features for categorisation (Rosch, 1973). In theory, this implies that the curse of dimensionality should make the acquisition of real-world categories a highly difficult learning problem. Yet people can acquire real-world categories with no apparent difficulty. Often this is achieved off of only limited experience with category exemplars. How do we explain this apparent contradiction?

The focus of the present study is on addressing one aspect of this question. It evaluates the hypothesis that people are able to use within-category feature correlational structure to identify important features amidst increasing dimensionality, thus allowing them to overcome the curse of dimensionality. This chapter begins by establishing some important definitions, and then explaining what is currently known about how people use and learn between-category and within-category feature information. This is then tied into a discussion on the curse of dimensionality and its role within human category learning. Finally, the chapter culminates in a brief look at the aims and hypotheses guiding the current study.

1.1.1. Foundations: conceptual representations and features

Before addressing these open questions, it is important to be clear about what we mean by *conceptual representations* and *features*. Conceptual representations are a fundamental explanatory device used by cognitive psychologists to explain human behaviour (Austerweil

& Griffihs, 2013). A core tenet of cognitive psychology states that an individual's reaction to a stimulus is defined by his or her representation of the stimulus and not the metaphysical state of the stimulus itself (Chomsky, 1959; Neisser, 1967). Representations have been notoriously hard to define (Cummins 1989; Markman, 1998), but a sufficient starting point for our purposes is provided by Palmer (1978, p. 262) who wrote: "A representation is something that stands in place for something else". Within cognitive science, the something that stands in place is the mental representation, and the something else is the real-world category. Building upon this, Markman (1998) defined concepts as mental representations that are used to compartmentalise the world for the purpose of analysis. These compartmentalisations can refer to objects, events, or ideas and can be used for reasoning, prediction, and communication. Although these two definitions are somewhat vague, their functional utility is sufficient for the aims of this thesis.

Features are a form of conceptual representation that are commonly used in psychological theories (Markman, 1998; Palmer, 1999; Tversky, 1977). They are rudimentary units that can be simple, such as the occurrence of a vertical line at a specific location, or more complex, such as the unitisation of numerous elements to form a feature. They can be discrete (binary - such as present or absent), have a countable set of values (a dotted or dashed line) or be continuous (e.g. the length of a line). The response to a given stimulus is produced by first translating the values of the features for the input and then making a decision on the basis of the resultant feature values. If two people are presented with the same stimulus, yet represent the stimulus with different feature values, then their response to the stimulus may vary (Austerweil & Griffiths, 2013).

A number of criteria for human feature representation can be derived from the literature. Firstly, research suggests that feature representation is flexible and adaptive, meaning that feature representation is not hardwired but is rather adaptive to the changing environment (Gerganov, Landy, & Roberts, 2008; Goldstone, 2003; Goldstone, Hoffman & Richards,

1984). It is also known that human learners have a number of perceptual and conceptual expectations that they use to infer features for object representation. For example, the features that people use to represent changing stimuli are reweighted according to experience and context (Garner, 2014; Gibson, 1969; Goldmeier 1972; Goldstone, 2003). People also infer features that are "simpler" (Austerweil & Griffiths, 2011; Chater & Vitanyi, 2003; Hochberg & McCalister, 1953), coherent with background knowledge of an objects function (Lin & Murphy, 1997), are consistent with previously learned categories (Pevtzow & Goldstone, 1994; Schyns & Murphy. 1994), and based on Gestalt Pragnnz (perceptual "goodness") (Palmer, 1977).

Other work has demonstrated that people learn features incrementally (Schyns & Rodent, 1997), that categorization training promotes inference of new feature representations (Goldstone, 2000; Goldstone & Steyvers, 2001; Lin & Murphy, 1997), and that people are able to recognise that two images can have the same feature, even if the feature occurs differently in each image (Palmer, 1983; Rock, 1973, Rust & Stocker, 2010). Overall, much is known about how individual features are learned and represented – but categories consist of multiple features that often have relationships to one another. What is known about this?

1.2 Correlations between and within features

As discussed earlier, features can have different relationships to each other depending on whether those correlations occur between categories or within categories (or both). This section considers each possibility in turn.

1.2.1. Between-category correlations

A large body of evidence supports the idea that people represent and use between-category feature correlations when reasoning about and learning natural categories (Rehder &

Burnett, 2004). This usage often revolves around making causal inferences about those features. Imagine having an encounter with an unfamiliar bird, and having to infer whether it is likely that the bird will fly. In a situation such as this, people might represent feature A and B of one type (e.g., size features, like feather size or wing size) and then infer the probability of feature C) of another type (e.g., ability features, like the ability to fly). Within the framework provided by Rehder & Burnett (2004), features A and B are effects, and feature C is a cause. Their work suggests if there is an observable effect feature present, then people are able to infer the existence of an unobservable cause feature. Moreover, they perform this inference based on prior knowledge of between-category feature correlations (e.g., wings and feathers are correlated with flying).

Additional research further supports the idea that people are sensitive to between-category feature correlations. For example, categories tend to form around clusters of casually related features (Ahn & Medin, 1992; Medin, Wattenmaker, & Hampson, 1987), with Rosch & Mervis (1975) emphasising the inter-feature correlations that are obtained between categories, thus defining the clusters of features. Indeed, supervised category learning is dependent on the casual relations that exist between a category's features (Waldmann, Holyoak, & Fratianne, 1995). Moreover, it has also been found that inter-feature causal relations have an influence on how items are classified (Ahn, 1998; Ahn, Kim, Lassaline, & Dennis, 2000).

As demonstrated, the presence of between-category feature correlations has been well established in the literature. A smaller but substantial body of evidence also supports the representation of within-category feature correlations by human learners, yet, many computational models still assume the independence of features. We turn to this issue now.

1.2.2. Within-category correlations

According to Chin-Parker & Ross (2002), the ability of human learners to access and represent within-category feature correlations is fundamental to their use of categories. One reason for this is that the knowledge of how features are related allows people to make specific predictions about the features of category members. Another reason is that if people use within-category correlations, their categories better reflect the relational structure of the real world. Just as between-category correlations afford us the ability to distinguish between categories, such as wings and flying, within-category correlations may help us to distinguish between subcategories, such as large and small wings. A final reason is that within-category feature correlations can signal the presence of a relationship between features, which may form a foundation for causal and explanatory grounding. For example, if one was to know that a novel object had the feature *beak,* they may be able to infer that it will most likely also have the feature *wings,* and thus, may fit within the *bird* category and be able to fly.

Since the 1970's, evidence that people represent within-category feature correlations has been found for both natural and artificial categories. For instance, Rosch, Mervis, Gray, Johnson, & Boyes-Braem (1976) found that features of natural categories tend to cluster together rather than being independent. For example, animals with feathers have wings and beaks, whilst animals with fur do not. The non-independence of features at this level of abstraction is fundamental to the creation of categories since we tend to make our major category divisions based upon the clusters of these properties (Malt & Smith, 1984). If these intuitions are correct, knowledge about the combinations of properties will influence typicality judgements in ways that cannot be accounted for if one presumes that stimulus features are conditionally independent of each other. Early evidence in support of this view came from Medin, Altom, Edelson & Frecko (1982), who found that novel exemplars were more likely to be classified as category members when they preserved correlations in the training stimuli than

when they broke them. Consistent with this, Neumann (1974) found that subjects not only encode the individual properties of training stimuli, but the property pairs that also occur, whilst Hayes-Roth & Hayes-Roth (1977), found that property combinations as well as individual properties, were also encoded.

More recent research has explicitly explored the representation of within-category feature correlations. Malt & Smith (1984) found that features of category members occur in systematic relationship to one another rather than existing independently in natural categories. Wattenmaker (1991) showed that participants can learn within-category correlational structures in implicit learning conditions. Moreover, people appear to use this correlational structure to make predictions, as has been demonstrated in multiple-cue probability learning (Hammond, McClelland, & Mumpower, 1980; Klayman, 1988). Anderson & Finchman (1996) found that participants both learnt and used within-category feature correlations among continuous dimensions; unlike previous research, these results could not be attributed to remembering specific instances or to breaking official categories down into subcategories. Further research has found that within-category feature correlations influence various cognitive tasks, including classifying novel items and inferring missing features (Crawford, Huttenlocher, & Hedges, 2006).

Despite the overwhelming evidence that people robustly use and learn within-category feature correlations, many computational models and theories that are used today still assume the independence of features. This includes the Rational Theory of Categorisation (Anderson 1990, 1991; Anderson & Matessa, 1992) as well as prototype theories which measure an instance in terms of its distance from a single average prototype, thus not allowing for the possibility that certain patterns of features may be correlated (Reed, 1972). Indeed, a central assumption of the family resemblance model and related models is that features are both

independent and additive (Collins & Loftus, 1975; McCloskey & Glucksberg, 1979; Smith, Shoben, & Rips, 1974; Tversky, 1977).

Although most researchers acknowledge that the assumption of class-conditional feature independence is made for reasons of mathematical and computational tractability rather than because it is necessarily true, this divergence between the empirical literature and the computational and theoretical accounts is striking and worth noting. It is especially interesting in light of the fact that computational and statistical models are known to struggle with certain issues involved in category learning that people do not. This thesis considers the possibility that people may overcome these issues by using the class-conditional feature information denied to most models. We focus on one particular issue, the *curse of dimensionality*.

1.3 The curse of dimensionality

The curse of dimensionality is a problem that arises in statistics, machine learning, and computer science (Bellman, 1961; Donoho, 2000; Keogh & Mueen, 2011; Verleysen & Francois, 2005). Intuitively, it centres on the fact that each additional feature or dimension in a model adds exponentially to the processing complexity of the model. Within the domain of category learning, as categories acquire more features, the size of the feature space and thus the number of examples necessary to adequately learn the category grows rapidly. As Vong et al., (2016) explains, for entities with N independent binary features, there are 2^N possible examples and 2 to the 2^N possible ways of assembling these objects into 2 distinct categories. As a result, the number of potential categories grows at a double-exponential rate compared to the number of independent features of a category. Therefore, even for modest values of *N*, category learning should be a demanding computational task (Searcy & Shafto, 2016).

This difficulty is exacerbated because, in most real-world category learning scenarios, features can take on more than two values. For example, if items can have 16 possible features

with five possible values each, there are $1.5x10^{11}$ possible exemplars. Learning how to classify all of these possible exemplars is clearly an acutely difficult learning problem. To make matters worse, most real-world categories have many available features for categorisation (Rosch, 1973). Taken altogether, the implication is that the curse of dimensionality should make the acquisition of real-world categories a highly difficult learning problem. Yet people can acquire real-world categories with no apparent difficulty. Often this is achieved off of only limited experience with category exemplars. How do we explain this apparent contradiction?

Studies investigating this question have yielded conflicting results. Some found that additional features impair learning (Edgell, Castellan, Roe, Barnes, Ng, Bright, & Ford, (1996), others found that they facilitate learning (Hoffman, Harris, & Murphy, 2008; Hoffman & Murphy, 2006), whilst others found that they have no effect on learning at all (Minda & Smith, 2001). Vong et al., (2016) observed that each of these studies differed in the category structure that was being learnt and suggested that variations in category structure may explain these discrepancies. For instance, Edgell et al. (1996) used many features, yet only a few were predictive of category membership, which may explain why more features were not useful. In contrast, when all features were somewhat predictive, especially if they were not perfectly correlated with each other, the additional features were beneficial or at least not harmful (Hoffman et al., 2008; Hoffman & Murphy, 2006; Minda & Smith, 2001). This observation is interesting given the fact that most real-world categories have precisely that sort of family resemblance structure (Murphy, 2004; Rosch & Mervis, 1975).

Consistent with these findings, Vong et al. (2016) found that category learning was not hurt by additional features if the categories followed a family resemblance structure. However, if categories were more rule-based (meaning only one of the many features was predictive of category membership), people were affected by the curse of dimensionality – categories with more features were not learned as well. Vong et al. (2016) argued that the observed pattern of

performance reflected capacity limitations that prevented people from using more than a few features at a time. As a result, when only one or a few features were predictive (as in rule-based categories), people struggled to find the useful ones. However, when all of the features were predictive this was not as much of a problem.

The fact that people struggled to find features in the rule-based categories is interesting because it dovetails with another open question in cognitive science: how do people know what the features are, anyway? While it's true that family resemblance categories contain many correlated features (like that birds who fly have wings and feathers), they also contain many properties that are uncorrelated and irrelevant to category membership (like that birds live 150 million kilometres from the sun and are smaller than houses). How do people identify which features are which? One implication of this previous work is that if people are sensitive to feature correlations, they might use that as a basis for identifying important features amidst noisy and irrelevant features in highly dimensional environments: perhaps important features are precisely those that correlate with one another. Identifying those features and focusing on them alone may then enable human learners to overcome the curse of dimensionality. This thesis explores this hypothesis, as described in the final section.

1.4 Aims and hypotheses

The current study investigates whether people use within-category feature correlations as a heuristic for identifying which features are important in environments with increasing dimensionality. To do this, we will present human learners with a classification task containing two different conditions. Participants will be shown a set of amoebas and will have to classify them as either Bivimia's or Lorifens.

The first condition will be an uncorrelated condition, it will start with four category predictive features that will predict the amoeba's category type 75% of the time. Noise features

will be added in increments of 4 up until 8 are present, resulting in a maximum of 12 features on each amoeba. The noise features will predict the amoeba's category type 50% of the time. The second condition will be a correlated condition that follows the same structure as the uncorrelated condition, yet the category predictive features will not only predict category membership 75% of the time, they will also predict each other 100% of the time. Thus, not only containing between-category feature correlations between the category predictive features and the categories, but within-category predictive feature relations also.

The main hypothesis is that human category learning will be improved when within-category feature correlations are present in the experimental stimuli, resulting in a higher number of correct classifications in the correlated condition. Thus, demonstrating the utilisation of within-category correlational structure by human learners in environments with increasing dimensionality.

A number of secondary hypotheses will also be tested. The second hypothesis states that the relative improvement in category learning will be greater when there are more potential features, because the difficulty of finding the useful feature will also be greater. It, therefore, makes sense that upon the identification of this predictive feature, the relative improvement in category learning will be greater compared to when there are less potential features.

The following hypothesis states that people will learn to identify important features better if they are correlated. For example, when feature x is 75% predictive of category membership and 100% predictive of other category predictive features, inter-correlated category predictive features will become more salient amidst increasing levels of noise features (dimensionality) within the experimental stimuli. Thus, allowing for better learning of category predictive features when inter-correlations exist between them.

In accordance with and contingent upon support for the previous hypotheses, the final hypothesis states that behaviour in this task cannot be accounted for by a model that assumes class-conditional feature independence.

2. Method

2.1 Participants

300 participants (174 males, 125 females, and 1 other) were recruited from Amazon Mechanical Turk. Amazon Mechanical Turk is an online service which allows people to complete online tasks that involve human intelligence. These tasks are known as Human Intelligence Tasks (HITs), and have been shown to be beneficial in conducting behavioural research due to the low cost of administration and diverse participant pools available (Paolacci, Chandler, & Ipeirotis, 2010; Mason & Suri, 2011). The mean age of participants was 35.8 years (SD: 10.6, range 19 to 68. All participants were 18 years of age or older and were predominantly from the United States (96%).[1] Participants were paid $3.50USD for their participation in the 15-20-minute study, which was coded in JavaScript.[2]

2.2 Design

Participants were presented with a supervised category learning problem in which they had to classify a novel stimulus (that looked similar to an amoeba) as either a Bivimia or Lorifen. Each amoeba had a circular base with a set of binary features (legs). Example stimuli are shown in Figure 3.

The two experimental conditions (CORRELATED and UNCORRELATED) manipulated the correlational **structure** of the categories in different ways. In both conditions, there were four binary features *f* that each predicted membership in category *c* 75% of the time: $P(c|f) = 0.75$. In the UNCORRELATED condition (N=137), the features of the amoeba were independent from each other, such that (once the category was known) knowing what one was gave no information about what the others were (i.e., $P(f_1|c, f_2, f_3, f_4) = P(f_1|c)$). In the CORRELATED

[1] Frequency data for participants by their country can be found in Appendix A.
[2] A/Prof Perfors coded the experiment, building on previous code from Wai Keen Vong.

condition (N=163) the features of the amoeba were 100% predictive of each other, such that $P(f_1|f_2,f_3,f_4) = 1.0$. This is shown in Figures 1 and 2.

Feature Structure: Uncorrelated Condition

	Object 1	**Object 2**	**Object 3**	**Object 4**
Feature 1	1	1	1	0
Feature 2	1	1	0	1
Feature 3	1	0	1	1
Feature 4	0	1	1	1
	Object 5	**Object 6**	**Object 7**	**Object 8**
Feature 1	0	0	0	1
Feature 2	0	0	1	0
Feature 3	0	1	0	0
Feature 4	1	0	0	0

BIVIMIAS

LORIFENS

Figure 1. Category structure for Bivimias and Lorifens in the UNCORRELATED condition. Four sample objects are shown of each category *c*, and each feature is 75% predictive of category membership. However, the feature values are class-conditionally independent of one another. This means that once the category is identified, knowing the value of one feature yields no additional information about the identify of any other feature of the object.

Feature Structure: Correlated Condition

BIVIMIAS	**Object 1**	**Object 2**	**Object 3**	**Object 4**
Feature 1	1	0	1	1
Feature 2	1	0	1	1
Feature 3	1	0	1	1
Feature 4	1	0	1	1
LORIFENS	**Object 5**	**Object 6**	**Object 7**	**Object 8**
Feature 1	0	0	1	0
Feature 2	0	0	1	0
Feature 3	0	0	1	0
Feature 4	0	0	1	0

Figure 2. Category structure for Bivimias and Lorifens in the CORRELATED condition. Four sample objects are shown of each category *c*, and each feature is 75% predictive of category membership. However, features are also 100% predictive of each other: knowing what one is means that you could successfully guess all of the other features of an object.

In addition to these two conditions, we varied how many additional irrelevant features there were across three levels of **dimensionality** (0, 4, and 8), which reflected the number of irrelevant binary features on the stimuli in addition to the four predictive ones described above. Summing the irrelevant and predictive features bring the total feature count for each **dimensionality** condition to 4, 8 and 12 (see Figure 3 below).

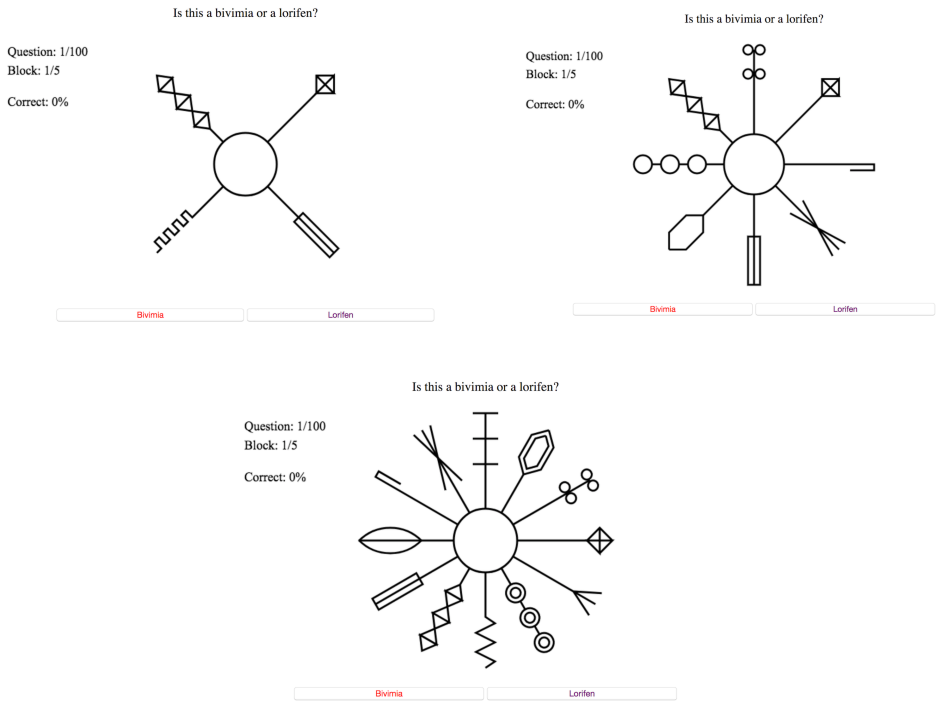

Figure 3. Example stimuli for each of the three-**dimensionality** conditions (4-FEATURE, 8-FEATURE, and 12-FEATURE, from left to right). The specific binary features (corresponding to the legs of the amoeba) randomly varied across participants.

For the lower-dimensionality conditions, the displayed features were a randomly selected subset of the features used in the 12-FEATURE condition. The position of each feature on the amoeba was randomised differently for each participant.

2.3 Procedure

Participants completed the study online through the Amazon Mechanical Turk platform. Once the task was selected and accepted, participants were asked to complete a few demographic questions regarding their gender, age, and country. Participants were then presented with instructions in which they were told that they needed to classify two types of amoebas known as Bivimias and Lorifens. Following this, in order to assess their command of English and understanding of the instructions, participants were quizzed on what they had just read. This was done to assess whether they understood the instructions and had a sufficient understanding of the English language. If any questions were answered incorrectly, the

participant was directed back to the instructions page and was required to read the instructions again until all questions were answered correctly. At this point, the participants began the experiment. Appendix B shows the full instructions and the quiz questions.

Part I of the experiment was a standard category-learning experimental design of 5 blocks of 20 learning trials each (100 total trials). On each trial, participants were shown either a Bivimia or a Lorifen and were asked to classify it accordingly (see Figure 4). Participants answered by clicking on a button and received feedback indicating whether or not they were correct. It was displayed for 500ms and consisted of a message ("Incorrect." vs "Correct!"), the correct category label, and a change to the colour of the circular base of the amoeba to indicate category membership (red for Bivimias and purple for Lorifens). Before the next trial was displayed, a blank screen was shown for one second. At the end of each block of twenty trials, people were given a summary of their performance (i.e., accuracy over each block so far).

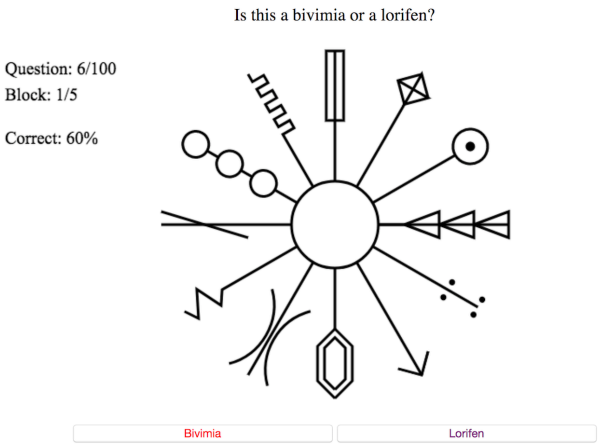

Figure 4. Sample trial from part I of the main category learning experiment (12-FEATURE).

Since Part I yields information only about people's categorisation accuracy, it is important to have additional measures that reflect what feature information the participants learned. Part II of the experiment, therefore, presented two tests (Isolated and Combined, described below).

To remove order effects, the order was randomised across participants so that some saw Isolated before Combined and some saw Combined before Isolated.

In the Isolated test, participants were presented with stimulus with only one feature, as in Figure 5, and had to classify it as either a Bivimia or a Lorifen based only on that feature. This test was designed to reveal whether or not the participants had learned the relationship between the category and the feature (i.e., $P(c|f)$) for each of the predictive features. There were eight Isolated test trials; each of the four category predictive features appeared twice in random order, once with one of the feature values and once with the other (e.g., once as  and once as  , though for simplicity we refer to each feature value as either 0 or 1). In the 8-FEATURE and 12-FEATURE conditions, there were also eight randomly-ordered additional test trials with eight of the irrelevant features. This was in order to stop people from noticing that the features from the other 8 trials were the category predictive features. The analysis presented in the results only includes the eight trials with category predictive features, since performance on the others is always at chance. Feedback was not given.

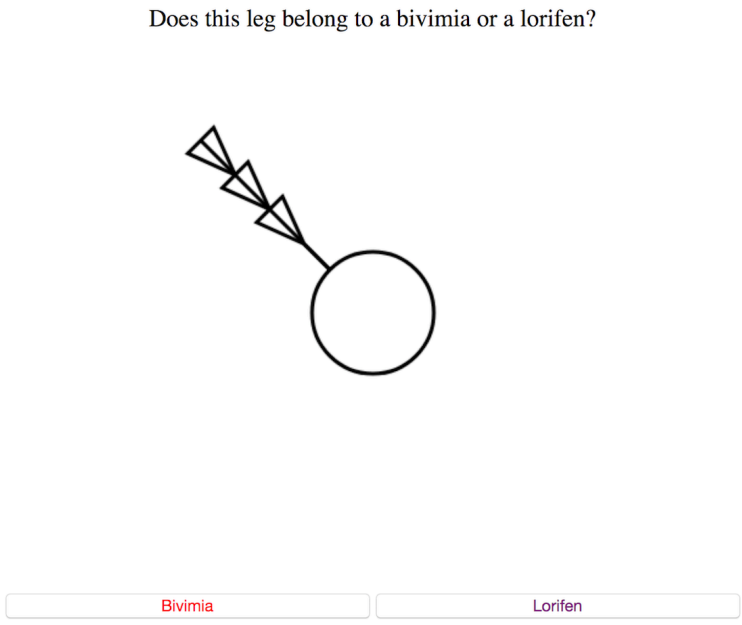

Figure 5. Example of a test trial in the Isolated test, which measured the extent to which participants had learned the predictiveness of individual features. Feedback was not given.

While the Isolated test aimed to measure people's knowledge of individual features, the Combined test was designed to reveal whether or not they had picked up on the within-category feature correlation information present in the CORRELATED condition. In it, people were shown stimuli with all features present except for one, as in Figure 6. They were then asked to predict the missing feature out of two possible options. If people in the CORRELATED condition had learned the correlation, their performance on this question could reach as high as 100%. Those in the UNCORRELATED condition, by contrast, were given no such information and thus could only achieve a maximum of 75% (since 75% is the class-conditional feature probability). As with the Isolated test, there were 8 test trials within which the missing feature was the category predictive feature and occurred twice. Also, as before, in the 8-FEATURE and 12-FEATURE conditions, there were eight additional trials with irrelevant features missing, which was designed to stop people from noticing that others were the category predictive features. The analysis in the results only includes the eight predictive features.

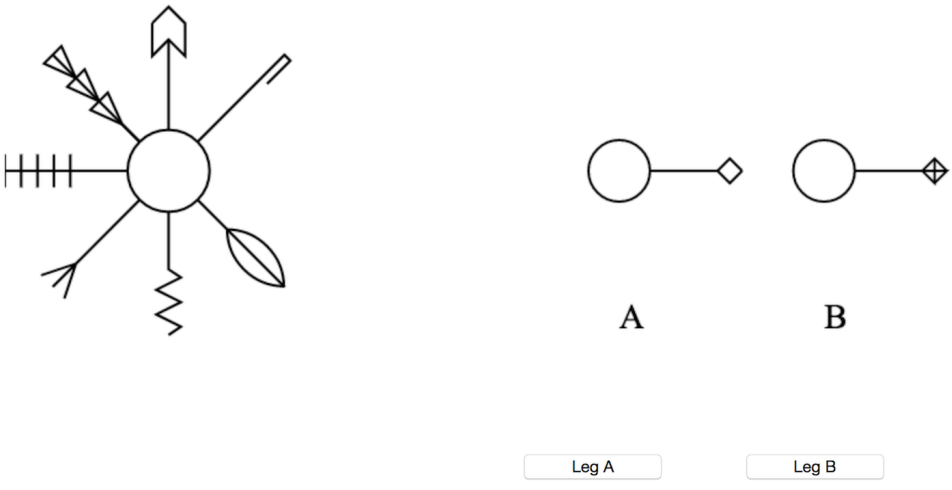

Figure 6. Example of a test trial in the Combined test, which measured the extent to which participants had learned the correlated feature information. Feedback was not given.

2.4 Naïve Bayes Model

In addition to an analysis of the human data on its own, this thesis also compares human performance to the predictions of a computational model. The model, known as a Naïve Bayes classifier, is explained in more detail below. The most important element for our purposes is that it assumes class-conditional feature independence. This means that it cannot identify and utilise the within-category correlational structure present in the CORRELATED condition.

This modelling is valuable in two ways. Firstly, it provides a baseline for comparison of human performance. Since the model is unable to detect or use within-category feature correlations, but is otherwise ideal (with perfect memory and ability to track all features), it establishes the best performance possible for a learner who does not use correlated feature information. If participants perform better than the model, this is an indication that they are using such information. If they are, this finding would further challenge the independence assumptions made by many models used in cognitive science today (Anderson 1990, 1991; Anderson & Matessa, Collins & Loftus, 1975; McCloskey & Glucksberg, 1979; Reed, 1972; Smith, Shoben, & Rips, 1974; Tversky, 1977). Secondly, this modelling demonstrates the *pattern* of performance one might expect from a learner that can fully track and use all features. To the extent that people depart qualitatively from this pattern in all conditions (e.g., if they struggle more when there are more features), this may be an indication that people are unable to attend to or use all of the features. Although it is known that people are not ideal, the precise way that human cognition departs from optimal is still being worked out, so this model may help us to that end.

The Naïve Bayes classifier uses information about every feature of the stimulus to derive its category predictions. The model calculates the predictiveness of each of the *i*

features in each category c $\left(i.e., P(x_i|c)\right)$ based on all of the data so far. Using this, it computes the probability that the stimulus is in category c (in this case, the categories are either Bivimias or Lorifens), which it does by assuming that all categories are Gaussian. The model assumes class-conditional feature independence between the features of the stimuli, meaning that each feature is independently predictive of the category label and combined, as seen in Equation 1. The predicted classification from the model on each trial is the category with the highest probability, as seen in equation 2.

$$\boldsymbol{p(c|x) \propto \prod_{i=1}^{D} p(x_i|c)p(c)} \qquad \textbf{(1)}$$

$$\boldsymbol{\hat{c} = \arg\max_{c} p(c|x)} \qquad \textbf{(2)}$$

The Naïve Bayes model was implemented by the author using Python 3.5. The model was selected from Pythons Scikit learn library, an open source machine learning library for data mining and analysis (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, Duchesnay, 2011). The computational modelling was implemented in a Jupyter Notebook, an open-source web application that allows for the creation and sharing of documents that contain live code, equations, visualisations and explanatory text.[3]

The model was tested on binary category structure data that was directly analogous to the stimuli experienced by human learners.[4] Since humans saw 100 trials, the model was given 100 objects; separate runs corresponded to each of the **dimensionality** x **structure** conditions.

[3] Appendix D shows a portion of the model code in Jupyter.
[4] See Appendix C for the full simulated dataset.

After testing, model performance was measured by presenting the model with 100 novel objects and calculating classification accuracy on those items.

3. Results

3.1 Overview

Our main hypothesis was that human category learning would be improved when within-category feature correlations were present in the experimental stimuli, resulting in a higher number of correct classifications in the CORRELATED condition. A number of additional hypotheses were also examined. Hypothesis 2 originated from the intuition that when there are more features (i.e., category dimensionality is higher), it should be harder to identify the predictive ones. If a correlated feature structure helps with that, then there should be a higher *relative* improvement in category learning when the dimensionality of the category is higher; that is, the difference in performance between the UNCORRELATED and CORRELATED conditions should become larger with more features. Hypothesis 3 stated that in addition to learning the categories better, people would also learn to identify the predictive features better if they were correlated with each other. Finally, Hypothesis 4 stated that people's behaviour in the task could not be accounted for by a model that assumes class-conditional feature independence (specifically, we predicted that people would learn better in the CORRELATED condition but that the model would learn better in the UNCORRELATED condition).

3.2 Hypothesis 1: Did human category learning improve in the CORRELATED condition?

To test our main hypothesis that human category learning would be improved when within-category feature correlations were present in the experimental stimuli, a Welch *t*-test was conducted on the overall accuracy scores between the CORRELATED and UNCORRELATED

conditions.[5] Findings did not support this hypothesis, with classification accuracy higher in the uncorrelated condition ($M = 0.63$, $SD = 0.11$) compared to the correlated condition ($M = 0.59$, $SD = 0.08$). This difference was statistically significant (0.04, 95% CI [0.02 to 0.06], $t(259) = 3.5$, $p = 0.001$) and is shown in Figure 7. This indicates that people were actually learning better in the UNCORRELATED rather than the CORRELATED condition.

Figure 7. Mean human classification accuracy during training for both the CORRELATED and UNCORRELATED conditions, across all levels of dimensionality. Contrary to expectations, people performed significantly better in the UNCORRELATED condition.[6]

3.3 Hypothesis 2: Was relative category learning improvement greater with greater dimensionality?

[5] An independent samples t-test was not used because the assumption of homogeneity of variance was violated (Levene's test for equality of variances, $F(2, 298) = 6.13$, $p = 0.014$). For all statistical tests reported in this thesis, normality assumptions were tested and met so non-parametric tests were unnecessary.

[6] As SEM values were so low, error bars were not visible (CORRELATED SEM: 0.006, UNCORRELATED SEM: 0.009).

The second hypothesis stated that the difference in category learning between the CORRELATED and UNCORRELATED conditions would be greater for human learners when there were more dimensions, because the difficulty of finding the useful feature would also be greater. Our results did not support this hypothesis. Firstly, a one-way ANOVA was conducted to determine if categorisation accuracy changed as a function of dimensionality in both the correlated and uncorrelated conditions[7]. Accuracy scores were statistically significantly different between the different levels of dimensionality in the correlated condition, $F(2, 160) = 22.49$, $p < 0.0005$), $n^2 = 0.22$. However, post hoc analysis revealed that relative category learning performance actually diminished significantly as a function of dimensionality between the 4 and 8 dimension levels, 0.06, 95% CI[0.03, 0.10], ($p < 0.0005$). However, this decrease in accuracy scores was insignificant between the 8 and 12 dimension levels (0.02, 95% CI[0.006, 0.06], $p = 0.12$). This trend continued in the uncorrelated condition. Accuracy scores were statistically significantly different between the 3 levels of dimensionality. $F(2, 134) = 13.91$, $p < 0.0005$), $n^2 = 0.17$. However, post hoc analysis revealed that there was a significant decrease in categorisation accuracy scores between the 4 ($M = 0.69$, $SD = 0.10$) and 8 ($M = 0.62$, $SD = 0.11$) dimension levels, a mean decrease of 0.07, 95% CI[0.02 to 0.12], ($p = 0.003$). However, there was no statistically significant difference between the 8 and 12 dimension conditions, 0.04, 95% CI[-0.01 to 0.08], $P = 0.17$.

[7] There were no outliers in the data, both conditions were normally distributed for the 8 and 12 dimension levels, however deviated from normality in the 4 dimension level in both conditions (CORRELATED - Shapiro-Wilk's test , $p = 0.006$, UNCORRELATED - Shapiro Wilk's test, $p = 0.03$). However, a one-way ANOVA was still used as they are considered robust to non-normality (Maxwell & Delaney, 2004). There was homogeneity of variances, as assessed by Levene's test for equality of variances. CORRELATED: ($F(2, 160) = 22.49$, $p = .703$). UNCORRELATED: ($F(2, 134) = 13.91$, $p = 0.103$.

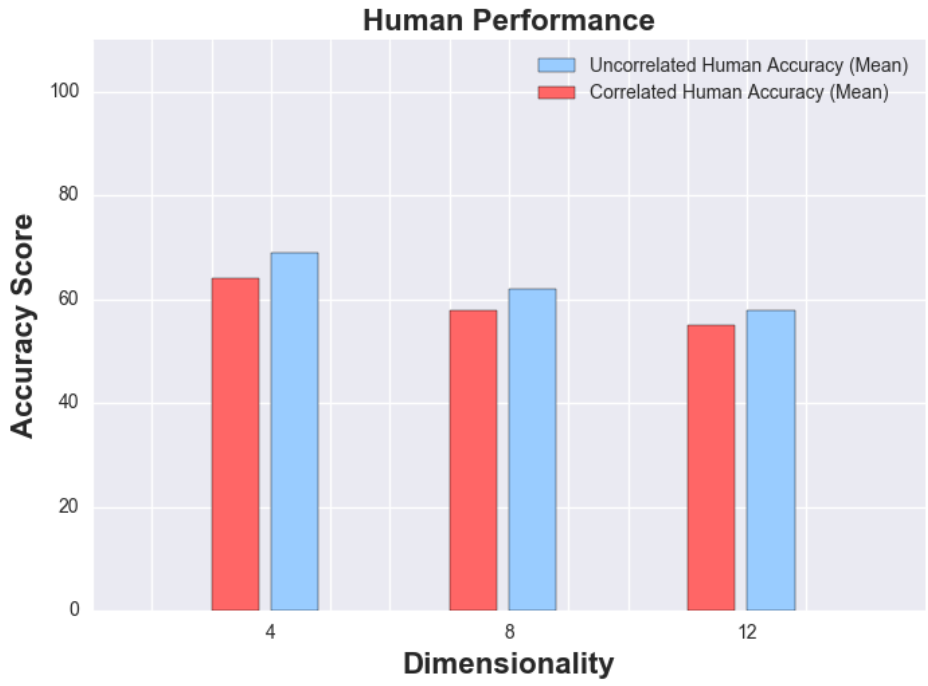

Figure 8. Mean human classification accuracy during training as a function of dimensionality. Overall accuracy decreased in both conditions with dimensionality, and the relative difference between conditions also decreased when there were more dimensions in the stimuli.[8]

Furthermore, accuracy scores were higher in the UNCORRELATED condition for the 4-FEATURE level (N = 45, M = 69, SD = 0.10), compared to the CORRELATED condition (N = 53, M = 0.64, SD = 0.07). This difference was statistically significant, 0.04, 95% CI [0.009 to 0.08], $t(77.36) = 2.49$, $p = 0.03$. Accuracy scores were also higher in the UNCORRELATED condition for the 8-FEATURE condition (N = 45, M = 62, SD = 0.11) compared to the CORRELATED condition (N = 57, M = 0.58, SD = 0.08). This difference was statistically significant, 0.04, 95% CI [0.003 to 0.08], $t(77.12) = 2.18$, $p = 0.03$.

Finally, in the 12-FEATURE condition, accuracy scores were also higher in the UNCORRELATED condition (N = 47, M = 58 SD = 0.08) compared to the CORRELATED condition (M = 55 SD = 0.08). This difference was also statistically significant, 0.03, 95% CI [0.002 to 0.065], $t(98) = 2.1$, $p = 0.04$ (see Figure 8 above).

[8]As SEM values were so low, error bars were not visible (CORRELATED 4,8 & 12 SEM: 0.009, 0.01, 0.01, UNCORRELATED 4,8, & 12 SEM: 0.01, 0.02, 0.01).

What we are seeing from these results, is that there was no greater relative category learning improvement as dimensionality increased between the CORRELATED and UNCORRELATED conditions. What we actually saw was diminishing relative category learning performance between the two conditions as a function of dimensionality (see Figure 9). Lack of support for this hypothesis was not surprising, as support for this hypothesis was contingent on support for hypothesis 1.

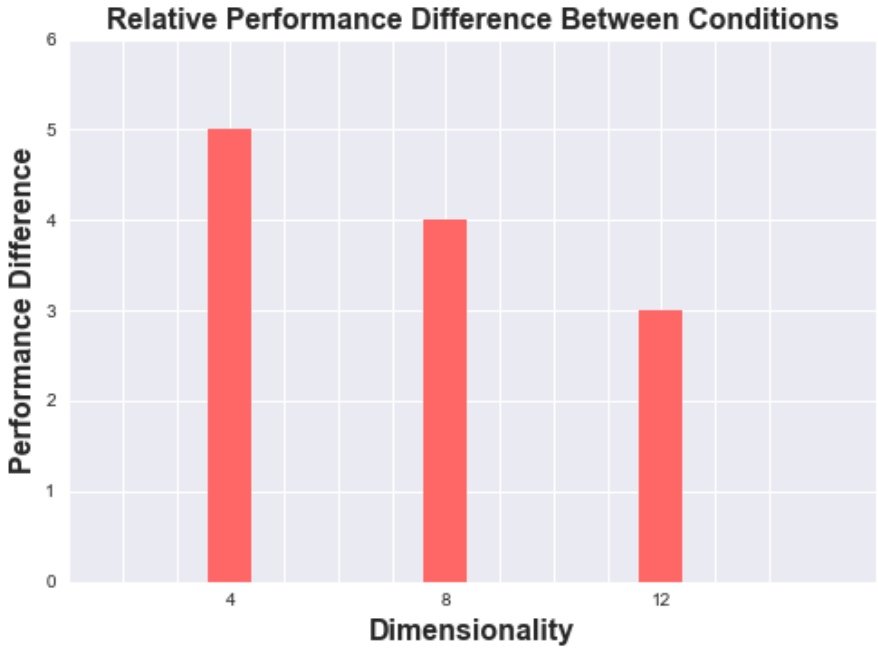

Figure 9. Relative performance differences between the CORRELATED and UNCORRELATED conditions across all levels of dimensionality. Contrary to expectations, relative category learning improvement was not greater with higher levels of dimensionality.

3.4 Hypothesis 3: Did people learn to identify important features better if there were correlated?

To test our third hypothesis, we analysed the data from the Isolated and Combined tests. In the Isolated test, participants were shown a stimulus with only one binary feature present and asked to classify it. The purpose of this test was to identify whether participants had learnt the relationship between each feature and the corresponding category label. This is an important first step before testing whether people could learn the within-category feature

correlations: if they could not identify the features in the first place, such correlations would be difficult. Given that our participants did not appear to be using the correlations, we might ask whether the problem came in learning the correlations or in learning the features themselves.

Recall that the four predictive features could predict category membership 75% of the time, whilst the irrelevant features could only predict category membership at chance (i.e., 50% of the time). We would, therefore, expect that there would be an accuracy ceiling of 75%, with any significant performance above 50% being indicative of some learning occurring.

To assess this intuition, a one-sample t-test was used.[9] In the CORRELATED condition, performance on the Isolated test was significantly higher by a mean of .19, 95% CI [.15 to .22], than an accuracy score that was no greater than that of chance ($t(162) = 10$, $p < 0.0005$, $d = .78$. In the UNCORRELATED condition, performance on the Isolated test was also significantly higher than an accuracy score that was no greater than that of chance by a mean of .17, 95% CI[.13 to .21], ($t(136) = 8.6$, $p < 0.0005$, $d = .73$. These findings suggest that people could learn which features were predictive of which category across both conditions, at least to some extent (see Figure 10).

[9] There were no outliers in the data as assessed by inspection of a boxplot for values greater than 1.5 box-lengths from the edge of the box, however, inspection of a Q-Q Plot revealed that the distribution of scores had a slight negative skew in both conditions. One-sample t-tests were still used due to their robust nature regarding type 1 error rates (Wilcox, 2012) and the large sample size used ($N = 300$). This also applies to the following combined test analysis.

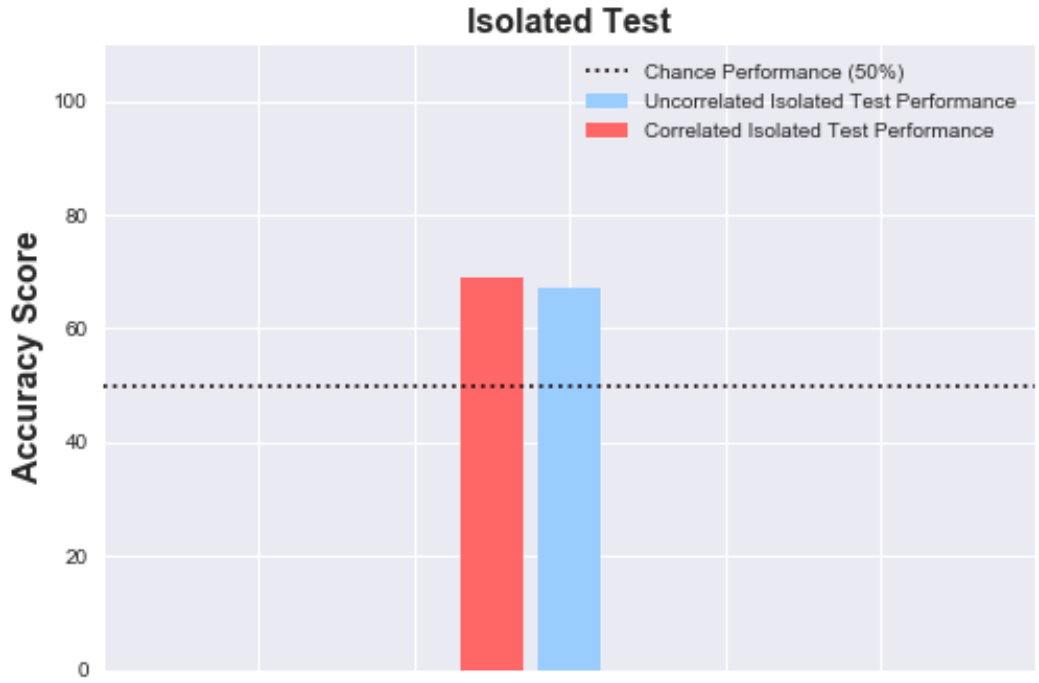

Figure 10. Isolated test performance for the CORRELATED and UNCORRELATED conditions. Accuracy scores were significantly higher than chance in both conditions, suggesting that participants could both detect and utilise the category predictive features.[10]

In addition to the Isolated test described above, participants also took part in a Combined test intended to assess how well they had detected the presence of within-category feature correlations. This involved giving them stimuli with a single missing feature and then asking them to identify which feature was missing. In the UNCORRELATED condition, it was expected that participants would perform no better than 75% (which is the class-conditional probability of any feature given correct identification of the category). In the CORRELATED condition, it was expected that performance would be higher than the UNCORRELATED condition, since participants also had access to the within-category correlational information.

To test this hypothesis, an independent samples t-test was conducted comparing performance on the Combined test across structure conditions. As Figure 11 shows, our

[10] As SEM values were so low, error bars were not visible (CORRELATED SEM: 0.02, UNCORRELATED SEM: 0.02).

hypothesis was not supported, although the trend was in the predicted direction. Performance tended to be better in the CORRELATED condition (M = .69, SD = .23) compared to the UNCORRELATED condition (M = .64, SD = .22), although these findings did not attain significance (95% CI [.00 to .10], $t(298) = 1.9$, $p = 0.054$, $d = 0.22$).

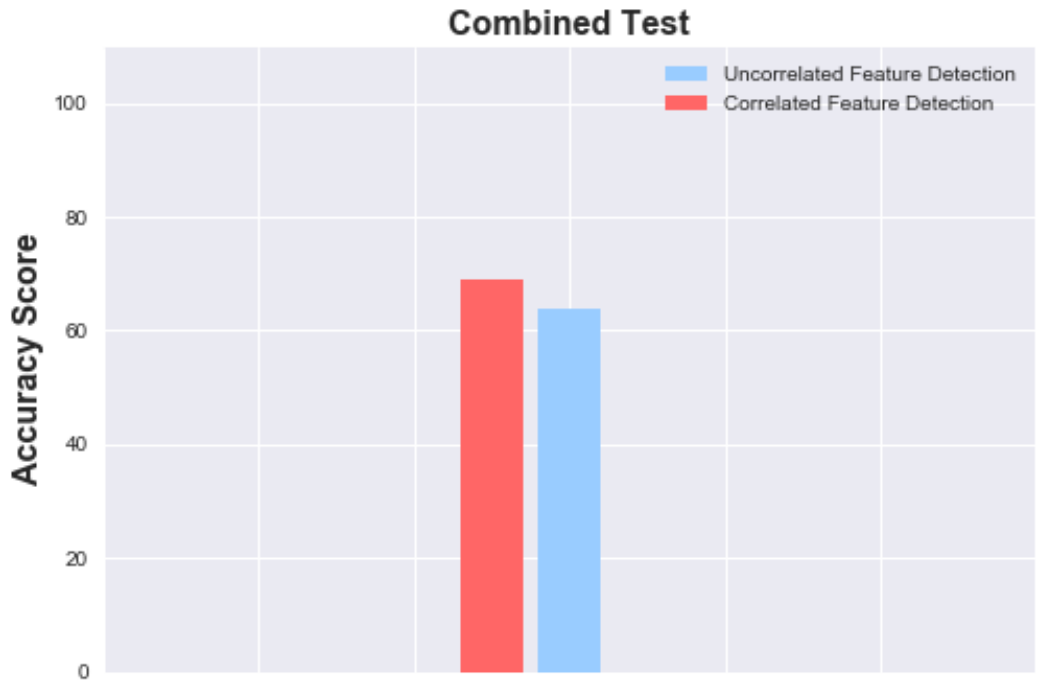

Figure 11. Feature detection scores between the CORRELATED (red) and UNCORRELATED conditions (blue). Feature detection accuracy was higher in the CORRELATED condition; however, these findings did not attain significance.[11]

3.5 Hypothesis 4: Could behaviour in this task be accounted for by a model that assumes class-conditional feature independence?

Our final hypothesis stated that behaviour in this task could not be accounted for by a model that assumed class-conditional feature independence. To test this hypothesis, human performance from Part I was compared to the classification performance of a Naïve Bayes

[11] As SEM values were so low, error bars were not visible (CORRELATED SEM: 0.02, UNCORRELATED SEM:0.02).

model. Human learners performed significantly better in the UNCORRELATED rather than the CORRELATED condition across all dimensionalities, although overall accuracy decreased in both conditions when there were more features. Does a Naïve Bayes classifier reproduce these same qualitative effects?

Just as people did, the Naïve Bayes classifier performed better in the UNCORRELATED condition (Figure 12), although Native Bayes was considerably better overall (with a mean accuracy of 85%, compared to 63%). In the CORRELATED condition, accuracy scores for both human learners and the model decreased as a function of dimensionality; however, in the UNCORRELATED condition, classification accuracy increased as a function of dimensionality in the model, whereas it decreased in the human model.[12]

Overall, support was mixed for this hypothesis. A model assuming class-conditional feature independence, like human learners, did perform better in the uncorrelated rather than correlated conditions; however, human performance decreased as a function of dimensionality whereas model performance remained constant in the uncorrelated condition.

[12] As the model was deterministic, there were no error bars. Furthermore, multiple runs were not conducted over multiple data sets, thus, statistics were inappropriate.

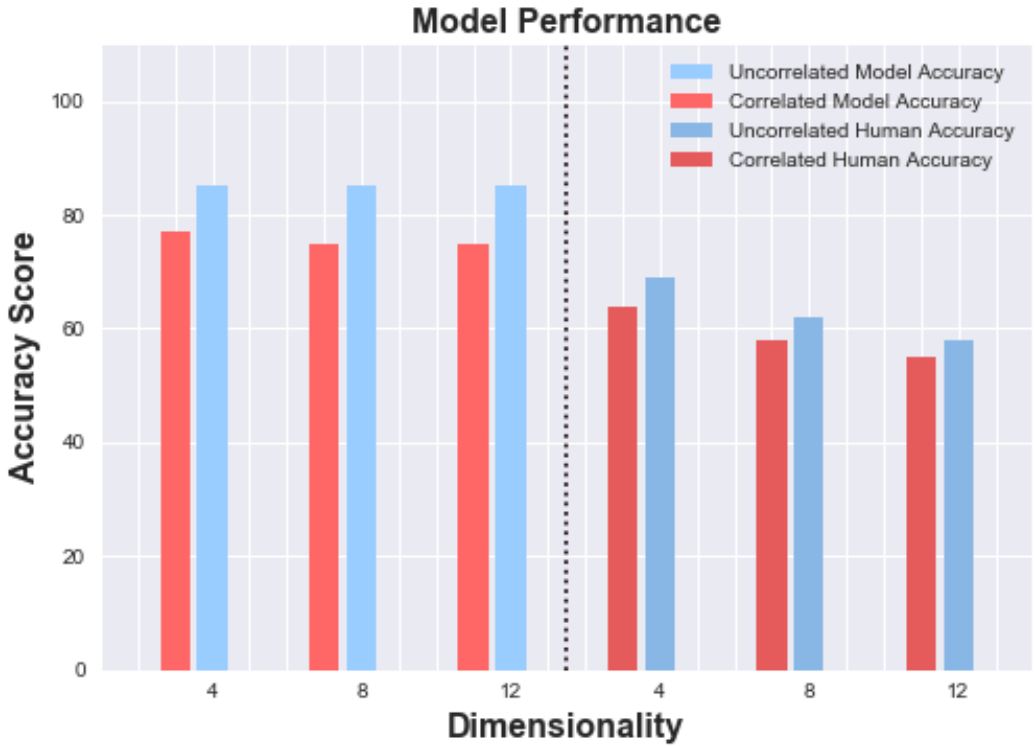

Figure 12. Classification accuracy scores for the Naïve Bayes model (left of line) and the human model (right of line) for both conditions across all levels of dimensionality (4, 8, and12). Contrary to our fourth hypothesis, participants performed significantly better in the UNCORRELATED condition across all levels of dimensionality (like the Naïve Bayes model). Human performance plots in the right panel are the same as those in Figure 8. However, they have been re-plotted so direct comparison to model performance can be made.

4. Discussion

4.1 Overview

The current study examined whether human learners utilise within-category feature correlations as a heuristic to overcome the curse of dimensionality in environments with increasing dimensionality. Participants took part in a category learning experiment on the Amazon Mechanical Turk platform where they were asked to classify Amoeba like stimuli as either Bivimias or Lorifens. Both CORRELATED and UNCORRELATED conditions were presented to participants and classification accuracy was recorded. Results from the present study indicated that participants did not use within-category feature correlations as a heuristic to identify category predictive features, and subsequently, overcome the curse of dimensionality. Results from the Isolated Test revealed that participants did appear to learn some correlational structure, however, this finding was insignificant and did not affect category learning performance in a way that was useful. As a result of these findings, a Naïve Bayes model mostly accounted for human performance in the main category learning task, with both the Naïve Bayes and human model performing better in the UNCORRELATED condition.

4.2 Hypothesis 1: Did human category learning improve in the CORRELATED condition?

Hypothesis one stated that human category learning will be improved when within-category feature correlations are present in the experimental stimuli, resulting in a higher number of correct classifications in the correlated condition. Support for this hypothesis was not attained, with overall classification accuracy significantly lower in the CORRELATED condition, indicating that people were actually learning better in the UNCORRELATED rather than the CORRELATED condition. It is important to acknowledge that it appears as if participants did figure out some of the correlational structure, as can be seen in Figure 11, but this was not enough to make a difference in overall classification accuracy scores (and these results were just shy of significance). There are two main explanations for these contradictory findings which will be discussed below.

4.2.1 Feature Attribute Variations

Firstly, the category learning task was much harder than was anticipated by the study authors. This was not realised until the application for the category learning task was complete, and given the time constraints faced in an honours year, design and coding of the task could not be re-done. To elaborate, upon the authors completion of the categorisation task, the within-category correlational structure appeared to be highly difficult to perceive. Given this observation, we may ask what was it about this correlational structure that was not analogous to perceived world structure? The following observations may help to elucidate an answer to this question.

In the category learning task used in this experiment, the features themselves remained spatially constant across the stimuli. However, the attributes of the features (the lines, dots, shading of the features etc) changed between each trial. This could be problematic, as

participants may not have had enough time to actually learn not just the correlations between features, but the subsequent correlations or lack thereof between the feature attributes and features themselves. For example, remember that in the CORRELATED condition, there were 4 category predictive features that could predict each other's presence 100% of the time. Now the number of unique attributes on each feature varied, but for the sake of argument, lets stipulate that there were 4 feature attributes on average. Given what we know about the curse of dimensionality, even with 4 category predictive features containing just 4 unique attributes, we will end up with 4^4 unique feature/attribute combinations, bringing us to a total of 256 unique combinations that need to be rapidly learnt between each trial, a highly demanding computational task.

If we consider how such a learning problem may be encountered in the natural world, it seems unlikely that this category structure would be common. For example, let's consider a young child trying to learn a novel category. Even though this child will constantly encounter highly dimensional environments filled with irrelevant, non-category predictive features, they will also have highly specified and consistent exposure to features of a category where both the attributes and the features themselves will remain relatively constant. Let's consider a child learning the category *cat*. In this category, the structure of the *cat's* nose, eyes and ears will remain the same, unless some sort of accident occurs. The *cat's* height will remain constant once fully grown and the colour of its fur will not change until old age. In this sense, making accurate categorisation decisions in highly dimensional environments may not just be contingent upon the capacity to detect within-category feature correlations, but also consistent exposure to stable feature attributes over a period of time. Something that the stimulus in this category learning task did not provide.

4.2.2 Stimulus Isolation

Another methodological shortcoming was that participants saw each stimulus in isolation, and never had the opportunity to compare each stimulus to one another. Reconsidering perceived world structure, we do not experience stimulus in isolation in the natural world, we tend to experience category members in either clusters, or in a way that allows direct comparison to other category and non-category members. For example, imagine going through your kitchen. In your draw, you may encounter the *cutlery* category. Members of this category will be clustered together and will share common features and attributes, e.g., shiny, mostly small, mostly made of silver or steel, used for eating etc. In this moment, all category members are experienced unanimously and the opportunity to directly contrast them to out of category members is salient. One may also observe that the plates in the next draw do not belong in the *cutlery* category, and neither do the cups. In this sense, the natural world provides ample opportunity for category members to be experienced unanimously and in contrast to non-category members. Therefore, the opportunity for within-category feature correlations to become salient, relative to category and non-category members is consistently provided, an element of category structure that was not provided in this experiment.

Future studies could easily address this short coming by presenting the stimulus used in this experiment side by side or spread out across the visual field. Participants could then progress through the task and classify each stimulus whilst comparing them to other category and non-category members. Thus, potentially allowing for easier identification of within-category feature correlations.

4.3 Hypothesis 2: Was relative category learning improvement greater with greater dimensionality?

Hypothesis two stated that relative improvement in category learning would be greater when there were more potential features, because the difficulty of finding the useful feature

would also be greater. Our results did not support this hypothesis, relative category learning accuracy actually decreased as a function of dimensionality in both the correlated and uncorrelated conditions. Lack of support for this hypothesis was not surprising, as support for this hypothesis was contingent upon support for hypothesis one. However, it is important to point out that people were responding in a reasonable way given the methodological limitations of this study. Participants were not just randomly responding or failing to learn anything, they were actually responding in a way that was consistent with the findings of previous research, showing that people find it harder to make accurate classification decisions as dimensionality increases (Searcy & Shafto, 2016; Vong, Hendrickson, Navarro & Perfors, 2016).

4.4 Hypothesis 3: Did people learn to identify important features better if there were correlated?

The third hypothesis stated that people would learn to identify important features better if they were correlated. This hypothesis was tested in the Combined Test and a small effect was obtained, however, this effect was just shy of significance. Therefore, this hypothesis was not supported.

There are two main explanations to consider in light of these results. Firstly, the fact that this hypothesis was just shy of significance may come down to a lack of statistical power. As the attained effect was so small (d = 0.22), a larger sample may be needed in future studies to significantly detect an effect of this size. This study may therefore be of use to inform effect size estimates in the *a priori* calculation of sample size in future studies. However, even if this lack of statistical significance was to be resolved in future studies, learning such a small amount of correlational structure did not appear to have an effect on performance in the main category learning task (see Figure 7), at least in a way that was useful to category learning.

We must also consider whether the lack of support for hypothesis one really did come from the absence of feature correlation learning, or from an inability to learn the features themselves. In the Isolated Test, one sample t-tests revealed that participants could identify important category predictive features significantly better than chance in both conditions. Suggesting that the problem in our category learning task did not come down to learning the features, but the correlations between them, thus, supporting our previous supposition. These findings further point to the need for future studies to assess power, hold feature attributes constant across features, and consider the use of stimulus clustering or the opportunity for direct stimulus comparison within each trial.

4.5 Hypothesis 4: Could behaviour in this task be accounted for by a model that assumes class-conditional feature independence?

The final hypothesis stated that behaviour in this task could not be accounted for by a model that assumed class-conditional feature independence. This hypothesis was not supported. Accuracy scores actually decreased as a function of dimensionality in the CORRELATED condition for both human learners and the Naïve Bayes model, further providing evidence for participants inability to detect feature correlations in the main category learning task.

In the UNCORRELATED condition, model performance actually stayed constant as a function of dimensionality (85%), whereas it decreased in the human model. This finding is not surprising as each feature predicted 75% to the category, but since each feature is independent, each provided an independent source of information. What this means, is that if the model takes all of these features into account it can actually do better than 75% accuracy. An interesting finding in lights of this, is that human learners actually performed worse as a function of dimensionality in the uncorrelated condition. If people were performing like the

model, they should also be able to treat each category predictive feature as independent sources of information and achieve better than 75% accuracy in the category learning task. The fact that they did not, points to the difficulty of finding the category predictive features amidst increasing dimensionality, a capacity constraint not inherent in the model. This observation further reiterates the importance of a heuristic to identify category predictive features. If one did not exist, the ability to detect category predictive features in the natural world would be highly constrained, due to the cognitive capacity limitations and constant sensory overload experienced in the natural world.

4.6 Limitations and suggestions for future research

Finally, there is always the possibility that people do not actually use within-category feature correlations as a heuristic for identifying category predictive features. Maybe there is an alternative explanation for this phenomenon that this thesis has not considered. Previous work has investigated alternative mechanisms for learning category predictive features, such as prior biases (e.g., shape bias / whole word bias) (Choi, McDaniel & Busemeyer, 1993) and visual salience (Liu et al., 2011). However, it is still unclear if these heuristics usefully apply in the general case of early stage category predictive feature detection and for all the kinds of categories that people learn. Alternative explanations such as these may be worthy of further investigation. However, before within-category feature correlations are rejected as an explanation for category predictive feature identification, future studies need to address the methodological limitations of this study. As discussed in the preceding sections, a logical starting point would be to hold feature attributes constant across inter-correlated features and allow for the comparison of stimulus within each trial. These design changes may have the effect of reducing cognitive load and creating a category structure that is more analogous to that of the natural world, thus, enhancing the external validity of future studies.

4.7 Conclusion

In the natural world, people are faced with constant sensory overload. Theoretically, this poses a problem for human category learning due to the curse of dimensionality. As we consider categories with increasing numbers of features, the size of the feature space grows rapidly, and thus, additional features end up increasing processing complexity at an exponential rate. However, people, attain real-world categories with ease, usually based off only a few available exemplars.

Previous research has suggested that people are more susceptible to the curse of dimensionality when categories follow a rule-based structure, yet, tend to overcome this problem when categories follow a family resemblance structure (Vong et al., 2016). However, for this structure to become salient, people need to be able to identify the category predictive features to being with. This means that the original category learning problem that people are posed with should leave them susceptible to the curse of dimensionality. To overcome this, some form of heuristic must be utilised by human learners for the detection of these category predictive features.

Previous research has demonstrated the detection of both between and within-category feature correlations in category learning tasks (Ahn, 1998; Huttenlocher, & Hedges, 2006), however, the role of within-category feature correlations had not yet been explored as a heuristic for predictive feature identification. The present study aimed to address this.

Results from the present study indicated that participants did not use within-category feature correlations as a heuristic to identify category predictive features, and subsequently, overcome the curse of dimensionality. Results from the Isolated Test revealed that participants did appear to learn some correlational structure, however, this finding was insignificant and did not affect category learning performance in a way that was useful. As a result of these

findings, a Naïve Bayes model mostly accounted for human performance in the main category learning task, with both the Naïve Bayes and human model performing better in the UNCORRELATED condition.

A number of methodological limitations may have contributed to these results. As discussed, feature attributes were not held constant across inter-correlated features, and the opportunity to directly compare stimuli to one another across trials was not provided. Two elements of category structure that that were not analogous to the natural world. Future studies could easily address these short comings by holding feature attributes constant across stimuli and allowing for the direct comparison of each stimuli within each trial.

In summary, the current study found no evidence to suggest that within-category feature correlations are used as a heuristic to overcome the curse of dimensionality in environments with increasing dimensionality. The results from the current study open up new directions for future research, with easily applied methodological changes suggested that may yield more promising results in future studies.

References

Ahn, W. K. (1998). Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality. *Cognition*, 69(2), 135-178. https://doi.org/10.1016/S0010-0277(98)00063-8

Ahn, W. K. (1999). Effect of causal structure on category construction. *Memory & Cognition*, 27(6), 1008-1023.

Ahn, W. K., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive psychology*, 41(4), 361-416. https://doi.org/10.1006/cogp.2000.0741

Ahn, W. K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, 16(1), 81-121. doi:10.1207/s15516709cog1601_3

Anderson, J. R. (1990). The adaptive character of thought: *Psychology Press*.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, 98(3), 409. http://dx.doi.org/10.1037/0033-295X.98.3.409

Anderson, J. R., & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(2), 259.

Anderson, J. R., & Matessa, M. (1992). Explorations of an incremental, Bayesian algorithm for categorization. *Machine Learning*, 9(4), 275-308.

Austerweil, J. L., & Griffiths, T. L. (2011). A rational model of the effects of distributional information on feature learning. *Cognitive psychology*, 63(4), 173-209. https://doi.org/10.1016/j.cogpsych.2011.08.002

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological review*, 120(4), 817. http://dx.doi.org/10.1037/a0034194

Bellman, R. (1961). *Adaptive control processes: a guided tour*. Princeton University Press. Princeton, New Jersey, USA.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in cognitive sciences*, 7(1), 19-22. https://doi.org/10.1016/S1364-6613(02)00005-0

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition*, 30(3), 353-362.

Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Incorporating prior biases in network models of conceptual rule learning. *Memory & Cognition*, *21*(4), 413-423.

Chomsky, N. (1959). A review of BF Skinner's Verbal Behavior. *Language*, 35(1), 26-58.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, 82(6), 407. http://dx.doi.org/10.1037/0033-295X.82.6.407

Crawford, L. E., Huttenlocher, J., & Hedges, L. V. (2006). Within-category feature correlations and Bayesian adjustment strategies. *Psychonomic bulletin & review*, 13(2), 245-250.

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

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. AMS Math Challenges Lecture, 1, 32.

Edgell, S. E., Castellan Jr, N. J., Roe, R. M., Barnes, J. M., Ng, P. C., Bright, R. D., & Ford, L. A. (1996). Irrelevant information in probabilistic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6), 1463. http://dx.doi.org/10.1037/0278-7393.22.6.1463

Garner, W. R. (2014). *The processing of information and structure:* Psychology Press.

Gibson, E. J. (1973). Principles of Perceptual Learning and Development. *Leonardo*, *6*(2), 190. doi:10.2307/1572721

Goldmeier, E. (1972). Similarity in visually perceived forms. *Psychological issues*.

Goldstone, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human perception and performance*, 26(1), 86. http://dx.doi.org/10.1037/0096-1523.26.1.86

Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. *Perceptual organization in vision: Behavioral and neural perspectives*, 233-278.

Goldstone, R., Gerganov, A., Landy, D., & Roberts, M. (2008). *Cognitive biology: Evolutionary and developmental perspectives on mind, brain, and behaviour* (10th ed., pp. 163-188). Cambridge, MA: MIT Press.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of experimental psychology: General*, 130(1), 116. http://dx.doi.org/10.1037/0096-3445.130.1.116

Hammond, K. R., McClelland, G. H., & Mumpower, J. (1980). Human judgment and decision making: Theories, methods, and procedures: *Praeger Publishers*.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior*, 16(3), 321-338. https://doi.org/10.1016/S0022-5371(77)80054-6

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

Hoffman, A. B., Harris, H. D., & Murphy, G. L. (2008). Prior knowledge enhances the category dimensionality effect. *Memory & Cognition*, 36(2), 256-270.

Hoffman, A. B., & Murphy, G. L. (2006). Category dimensionality and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology-Learning Memory and Cognition*, 32(2), 301-315.

Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. *Cognition*, 18(1), 65-96. https://doi.org/10.1016/0010-0277(84)90022-2

Keogh, E., & Mueen, A. (2011). Curse of dimensionality. In *Encyclopedia of Machine Learning* (pp. 257-258). Springer US. doi: 10.1007/978-0-387-30164-8_192

Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(2), 317. http://dx.doi.org/10.1037/0278-7393.14.2.317

Lin, E. L., & Murphy, G. L. (1997). Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human perception and performance*, 23(4), 1153. http://dx.doi.org/10.1037/0096-1523.23.4.1153

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., & Shum, H. Y. (2011). Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, *33*(2), 353-367. doi: 10.1109/TPAMI.2010.70

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, 23(2), 250-269. https://doi.org/10.1016/S0022-5371(84)90170-1

Margolis, E., & Laurence, S. (1999). *Concepts: core readings*: Mit Press.

Markman, A. (1998). *Knowledge Representation* (10th ed.). Mahwah NJ: Lawrence Erlbaum.

Maxwell, S., Delaney, H., & Kelley, K. (2004). *Designing experiments and analyzing data* (1st ed.). New York: Psychology Press.

McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive psychology*, 11(1), 1-37. doi: https://doi.org/10.1016/0010-0285(79)90002-1

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(1), 37. http://dx.doi.org/10.1037/0278-7393.8.1.37

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive psychology*, 19(2), 242-279. https://doi.org/10.1016/0010-0285(87)90012-0

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(3), 775. http://dx.doi.org/10.1037/0278-7393.27.3.775

Murphy, G. (2004). Introduction. In G. Murphy, *The Big Book of Concepts* (1st ed., pp. 1-20). MIT Press.

Murphy, G. L., & Wisniewski, E. J. (1989). Feature correlations in conceptual representations. *Advances in cognitive science*, 2, 23-45.

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

Neumann, P. G. (1974). An attribute frequency model for the abstraction of prototypes. *Memory & Cognition*, 2(2), 241-248.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive psychology*, 9(4), 441-474. https://doi.org/10.1016/0010-0285(77)90016-0

Palmer, S. E. (1983). The psychology of perceptual organization: A transformational approach. *Human and machine vision*, 1, 269-339.

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

Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, Duchesnay, (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825-2830.

Pevtzow, R., & Goldstone, R. L. (1994). Categorization and the parsing of objects. In *Proceedings of the sixteenth annual conference of the cognitive science society* (pp. 717-722).

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive psychology*, 3(3), 382-407. https://doi.org/10.1016/0010-0285(72)90014-X

Rock, I. (1973). *Orientation and form:* Academic Press New York.

Rosch, E. (1973). Natural categories. *Cognitive psychology*, 4(3), 328-350. https://doi.org/10.1016/0010-0285(73)90017-0

Rosch, E. (1999). Principles of Categorization. In E. Margolis & S. Laurence, *Concepts: Core Readings* (4th ed., p. 190). MIT Press.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4), 573-605. https://doi.org/10.1016/0010-0285(75)90024-9

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, 8(3), 382-439.

Ross, B.H., & Spalding, T.L. (1994). Concepts and categories. In R. J. Sternberg (Ed.), *Handbook of perception and cognition: Vol. 12. Thinking and problem solving* (pp. 119-148). San Diego: Academic Press.

Russell, S., Norvig, P., & Canny, J. (2003). *Artificial intelligence* (1st ed., p. 478). Englewood Cliffs, N.J.: Prentice Hall.

Rust, N. C., & Stocker, A. A. (2010). Ambiguity and invariance: two fundamental challenges for visual processing. *Current opinion in neurobiology*, 20(3), 382-388. https://doi.org/10.1016/j.conb.2010.04.013

Schyns, P. G., & Murphy, G. L. (1994). The ontogeny of part representation in object concepts. *The psychology of learning and motivation*, 31, 305-349.

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

Searcy, S. R., & Shafto, P. (2016). *Cooperative* inference: Features, objects, and collections. http://dx.doi.org/10.1037/rev0000032

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological review*, 81(3), 214. http://dx.doi.org/10.1037/h0036351

Tversky, A. (1977). Features of similarity. *Psychological review*, 84(4), 327-352. http://dx.doi.org/10.1037/0033-295X.84.4.327

Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series prediction. *Computational Intelligence and Bioinspired Systems*, 85-125.

Vong, W. K**.,** Hendrickson, A. T., Perfors, A. F. & Navarro, D. J. (2016). Do additional features help or harm during category learning? An exploration of the curse of dimensionality in human learners. In A Papafragou, D Grodner, D Mirman and JC Trueswell (Ed.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2471-2476).

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of *category* structure. *Journal of Experimental Psychology*, 124, 181-181.

Wattenmaker, W. D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(5), 908. http://dx.doi.org/10.1037/0278-7393.17.5.908

Wilcox, R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Waltham, MA: Elsevier.

Appendices

Appendix A: Participants by Country

Table 1.
Number of participants by country

Country	No. of Participants
Canada	1
Columbia	1
India	7
Ireland	1
Turkey	1
United States of America	288
Venezuela	1

Appendix B: Experiment Instructions

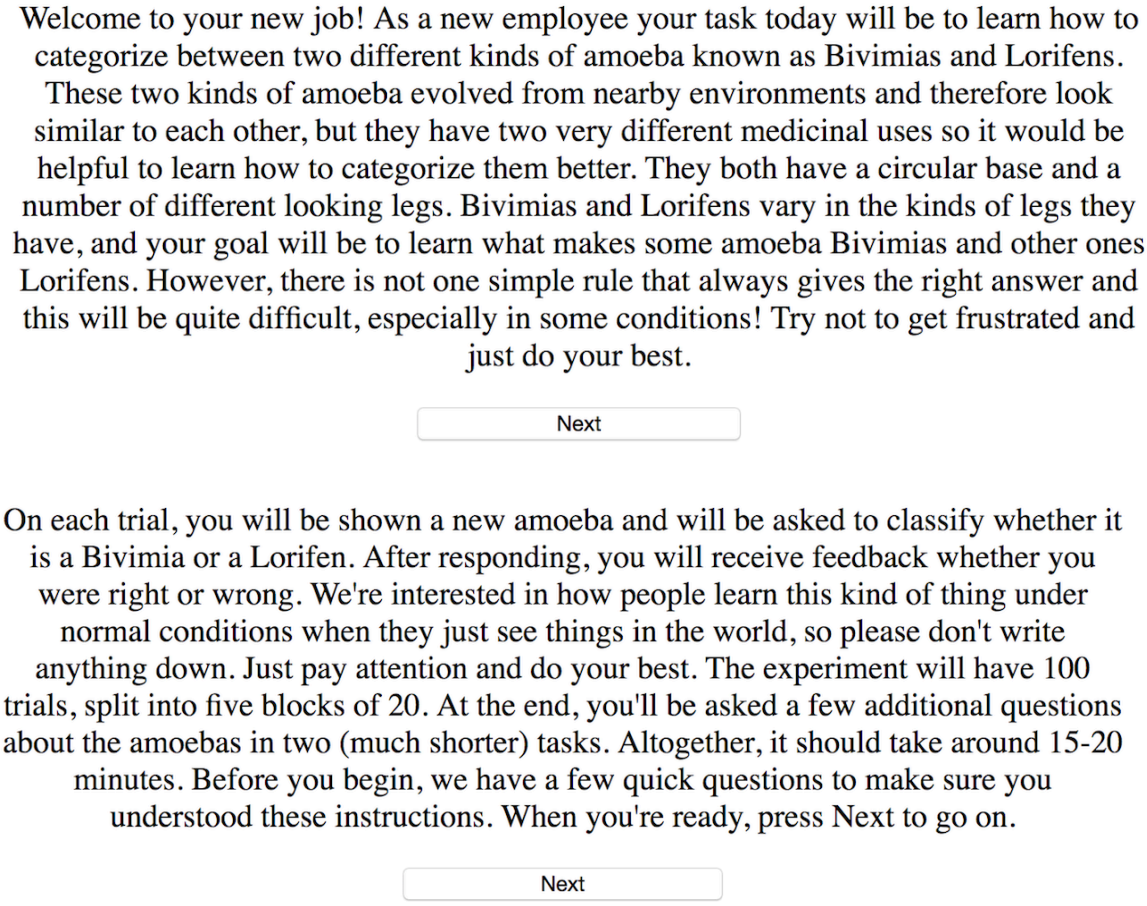

Figure 13. Experiment instructions.

Appendix B Continued: Experiment Instructions

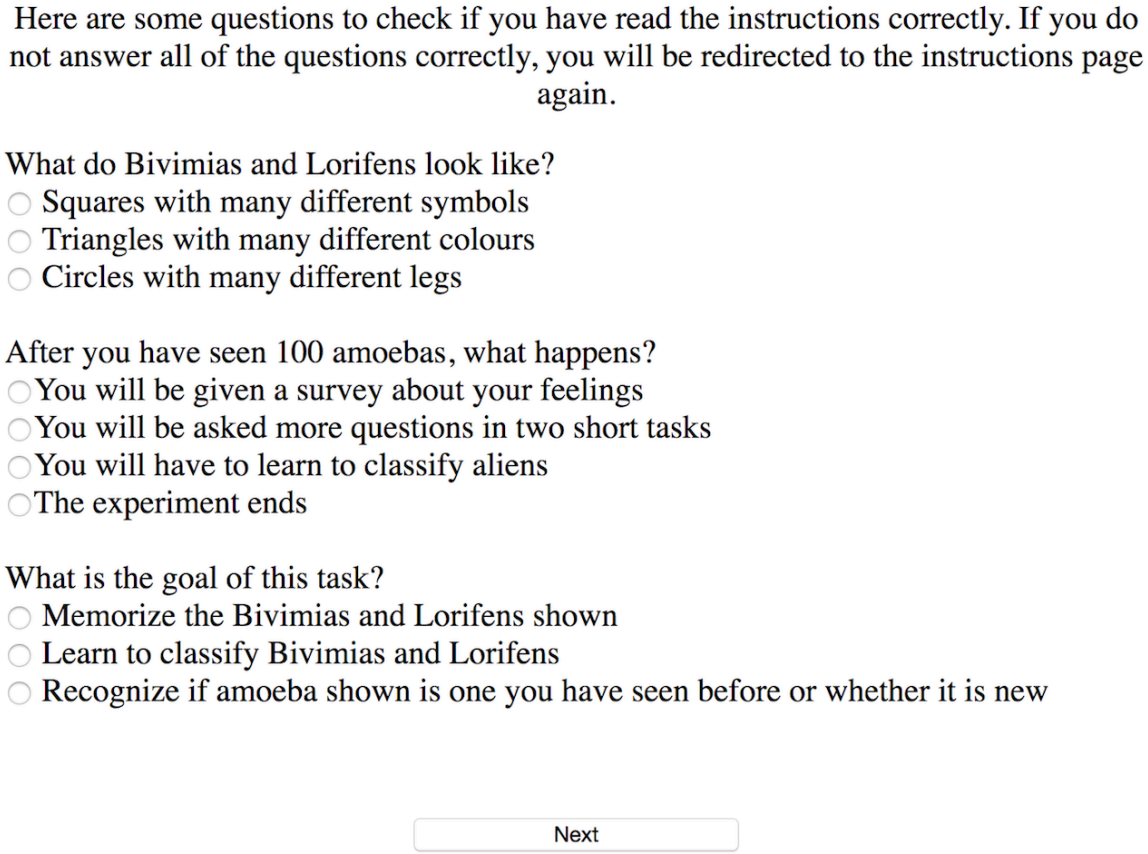

Figure 14. Checks to make sure that the participant has understood the instructions.

Appendix C: Model Data Structure

f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	cat
0	0	1	0	0	0	0	1	0	0	0	0	0
1	1	0	0	1	1	0	0	1	0	1	0	0
1	1	0	0	1	0	0	0	0	0	1	1	0
1	1	0	1	1	0	0	1	0	1	0	0	0
1	1	1	0	1	1	1	0	1	0	1	1	0
0	0	1	1	0	0	1	1	0	0	0	0	0
0	1	1	1	1	1	0	0	0	0	1	1	0
1	1	1	1	1	0	1	0	1	0	0	1	0
1	0	1	1	1	1	0	1	0	1	1	1	0
1	0	1	1	1	0	1	1	0	0	0	1	0
1	1	0	0	0	1	1	1	0	1	1	1	1
0	0	0	0	0	0	1	1	0	1	0	1	1
0	0	0	1	0	1	1	0	1	0	0	0	1
0	0	1	1	0	0	1	0	1	0	1	0	1
1	1	0	0	0	0	0	0	1	1	0	1	1
1	0	0	1	1	1	1	0	1	1	1	0	1
1	0	0	0	0	1	1	1	1	1	1	0	1
0	0	0	0	0	0	0	0	0	1	1	0	1
1	0	0	0	1	0	0	0	1	0	0	1	1
0	0	0	0	1	1	0	1	0	1	0	1	1
0	1	1	1	0	1	1	1	0	1	0	0	0
1	1	0	1	0	0	0	1	1	0	0	0	0
1	0	1	1	0	0	1	1	1	0	1	1	0
1	1	1	0	0	1	1	0	0	1	0	0	0
1	1	1	1	1	0	0	1	1	1	1	0	0
1	0	1	0	1	0	1	1	1	1	1	0	0
1	1	1	1	0	1	1	0	1	0	0	0	0
0	1	1	1	0	1	0	1	1	1	0	0	0
1	0	0	1	1	0	1	0	0	1	0	0	0
1	1	1	1	0	1	1	1	0	1	0	1	0
0	1	0	0	0	0	1	1	1	0	0	0	1
0	0	0	0	0	1	0	0	1	1	0	1	1
0	0	0	1	0	1	1	0	1	1	0	1	1
0	0	0	0	1	0	0	1	0	1	0	1	1
0	0	1	0	0	0	0	1	0	0	0	1	1
1	0	0	0	0	0	1	1	0	0	1	1	1
0	0	0	0	0	0	0	1	0	0	1	1	1
0	1	0	0	0	0	0	1	1	0	0	0	1
0	1	0	1	1	0	0	1	1	0	1	0	1

Figure 15. Randomly generated csv data from R for the 12-dimension uncorrelated condition showing the first forty objects and the corresponding category label. Features one to four are 75% category predictive whereas features five to twelve are 50%cateogry predictive.

Appendix D: Model Code

Correlated Condition - Four Dimensions.

In [2]:
```
import pandas as pd
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score
```

In [3]:
```
correlated_4_features_train = pd.read_csv("correlated_4_features_train.csv", delimiter=',', encoding='utf-8-sig')
correlated_4_features_test = pd.read_csv("correlated_4_features_test.csv", delimiter=',', encoding='utf-8-sig')
#print(correlated_4_features_train)
#print(correlated_4_features_test)
```

In [4]:
```
# format the data and expected values for SKLearn
# y:'Cond': 0 = bivimias, 1 = lorifens

x_train = pd.DataFrame.as_matrix(correlated_4_features_train[['f1', 'f2', 'f3', 'f4']])
y_train = pd.DataFrame.as_matrix(correlated_4_features_train[['Cat']]).ravel()
x_test = pd.DataFrame.as_matrix(correlated_4_features_test[['f1', 'f2', 'f3', 'f4']])
y_test = pd.DataFrame.as_matrix(correlated_4_features_test[['Cat']]).ravel()
```

In [5]:
```
model = GaussianNB() # create a Gaussian Classifier
```

In [8]:
```
model.fit(x_train, y_train) # train the model using the training sets
```

Out[8]:
```
GaussianNB(priors=None)
```

In [9]:
```
# running test set through the trained model
model_prediction = model.predict(x_test)
print(model_prediction)
```

```
[0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0
 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 1 1 0 1 1 1 0 1 1 1 0 1 1 1 0 1 1 1 0 1 1 1
 0 1 1 1 0 1 1 1 0 1 1 1 0 1 1 1 0 1 1 1 0 1 1 1 0 1 1 1 0 1 1 1 0 1]
```

In [10]:
```
model_accuracy = accuracy_score(y_test, model_prediction)*100 # model accuracy score
```

In [12]:
```
print("Model accuracy percent is", (model_accuracy))
```

```
Model accuracy percent is 75.0
```

Figure 16. Model experiment code from Jupyter in the 4-dimension correlated condition. A Gaussian Naïve Bayes model from Scikit learn is being trained and tested on the category learning task.