Do additional features help or hurt category learning? The curse of dimensionality in human learners

Wai Keen Vong
Department of Mathematics and Computer Science
Rutgers University—Newark

Andrew T. Hendrickson
Department of Cognitive Science & Artificial Intelligence
Tilburg University

Danielle J. Navarro
School of Psychology
University of New South Wales

Amy Perfors
School of Psychological Sciences
University of Melbourne

Abstract

The curse of dimensionality, which has been widely studied in statistics and machine learning, occurs when additional features causes the size of the feature space to grow so quickly that learning classification rules becomes increasingly difficult. How do people overcome the curse of dimensionality when acquiring real-world categories that have many different features? Here we investigate the possibility that the structure of categories can help. We show that when categories follow a family resemblance structure, people are unaffected by the presence of additional features in learning. However, when categories are based on a single feature, they fall prey to the curse and having additional irrelevant features hurts performance. We compare and contrast these results to three different computational models to show that a model with limited computational capacity best captures human performance across almost all of the conditions in both experiments.

Introduction

Despite the fact that category learning is logically difficult in many ways, people easily and naturally learn real-world categories. Quine (1960) identified one well-known problem, originating from the fact that the category referent for any particular word is under-determined and could be any from an infinite set of possibilities. This is an example of the problem of induction (e.g., Goodman, 1983), which concerns the difficulty in identifying

how to generalize when there are a potentially infinite set of possible bases to do so. The problem of induction, in its different forms, has been widely studied within cognitive science. Proposed solutions often center around the existence of inductive biases, although the exact nature of these biases remains a debated issue (see, e.g., Markman, 1989; Landauer & Dumais, 1997; Griffiths, Kemp, & Tenenbaum, 2008; Chater, Clark, Goldsmith, & Perfors, 2015; Minda, 2015). In this paper, we focus on a less studied but related problem known as the curse of dimensionality. It is similar in that it is a fundamental problem of learnability, but different in that it relates to the specific problem of learning in high-dimensional spaces or with a large number of features. We show why the acquisition of real-world categories *should* be difficult due to the curse of dimensionality, but propose that the structure of real-world categories may alleviate the curse for humans, at least in many situations.

The curse of dimensionality has been well-studied within statistics (e.g., Bellman, 1961; Donoho, 2000) and machine learning (e.g., Verleysen & François, 2005; Keogh & Mueen, 2011), and has a number of interesting and widespread effects. Within computer science, the curse of dimensionality means that if the amount of data on which to train a model (e.g., a classifier) is fixed, then increasing dimensionality can lead to overfitting. This is because as the space grows larger, the examples themselves grow ever sparser; the only way to avoid the issue is to bring in exponentially more data for each additional dimension. In statistics and mathematics, the curse means that it is not possible to numerically optimize functions of many variables by exhaustively searching a discretized search space, thanks to the combinatorial explosion of parameters to explore.

A similar problem arises in the domain of category learning: as we consider categories with more and more features, the size of the possible feature space and number of examples required to fully learn a category grows extraordinarily quickly. For objects with N independent binary features, there are 2^N possible examples and 2^{2^N} possible ways of grouping these objects into two distinct categories. The number of possible category structures grows at a double-exponential rate, as a function of the number of independent features used to describe stimuli (Searcy & Shafto, 2016). As a result, even for moderate values for N, learning categories should be extremely difficult. For instance, items with 16 possible features of two possible values each yields 65536 possible exemplars.

Most real-world categories have a large number of available features for categorization (Rosch, 1973), which suggests that – in theory at least – the curse of dimensionality means that acquiring natural categories should be a difficult learning problem. Yet people, including children, can learn real-world categories with relative ease, often based on only a few exemplars. How do people accomplish this feat?

We know surprisingly little about the answer to this question. Most experimental work in category learning has not run into the problem of the curse of dimensionality, either because studies have used categories that people have already learned or because they tested categories using stimuli with only a few, highly salient features (e.g, Shepard, Hovland, & Jenkins, 1961; Medin & Schaffer, 1978; Nosofsky, 1986). Although this body of work has substantially contributed to our understanding of category learning, it remains an open question how learning is affected when there are a large number of features.

While some studies in the category learning literature have used stimuli with a large number of features (e.g, McLaren, Leevers, & Mackintosh, 1994; Wills & McLaren, 1997, 1998; Jones, Wills, & McLaren, 1998), the focus of these particular studies was not directly

related to how varying the number of features impacts learning. Furthermore, the limited set of studies that have investigated category learning with varying numbers of features have yielded conflicting results, with some studies finding that additional features impair learning (Edgell et al., 1996), others finding that they facilitate learning (Hoffman & Murphy, 2006; Hoffman, Harris, & Murphy, 2008), and others finding that they have no effect on learning at all (Minda & Smith, 2001), or that they do both (Bourne & Restle, 1959).

How can we resolve this apparent discrepancy? One possibility is that each of these studies differ in the kinds of category structures being learned. After all, the curse of dimensionality stems from having so many possible stimuli configurations in a high-dimensional space that it is difficult to learn which set of features people should use for classification. This should lead to the greatest inefficiency when most of the possible features are not predictive of category membership and only one or a few matter, as in Walker and Bourne (1961) and Edgell et al. (1996). By contrast, if all features are predictive to some degree – especially if they are not perfectly correlated with each other – then additional features should be beneficial, or at least not harmful (Hoffman & Murphy, 2006; Hoffman et al., 2008; Minda & Smith, 2001). This possibility is especially interesting given the fact that most real-world categories have precisely this sort of family resemblance structure (Rosch & Mervis, 1975; Murphy, 2002).

This hypothesis – that a family resemblance category structure may mitigate the impact of the curse of dimensionality, but that other kinds of category structures may not – appears plausible on its face, but to date no studies have tested it. The goal of the current paper is to examine this hypothesis by manipulating category structure and the number of features while holding other factors constant. Our results do indeed suggest that people do not succumb to the curse if the categories follow a family resemblance structure. However, if only a single feature is relevant among a set of features for categorization, the curse of dimensionality affects humans. We argue that the pattern of performance reflects capacity limitations that prevent people from learning from and using more than a few features at a time. As a result, learning is impaired where only one or a few features are predictive, but it is not affected when many different features are all predictive (as any given feature will be similarly useful). We support this interpretation by a comparison of several computational models with different assumptions about representation and capacity limitations, and we find that human performance is best fit by a model with limited capacity for learning.

Experiment 1

We addressed two questions in this experiment. First, how does learning performance change as the number of features increase? Second, to what extent does the structure of the category influence this learning? We therefore systematically manipulated the number of features and the manner in which categories were structured within the context of a standard supervised learning task.

Method

Participants. 886 participants (496 male, 388 female, 2 other) were recruited via Amazon Mechanical Turk. This is a relatively high number of participants because we ran two experiments with slightly different methodologies (described below) but pooled the

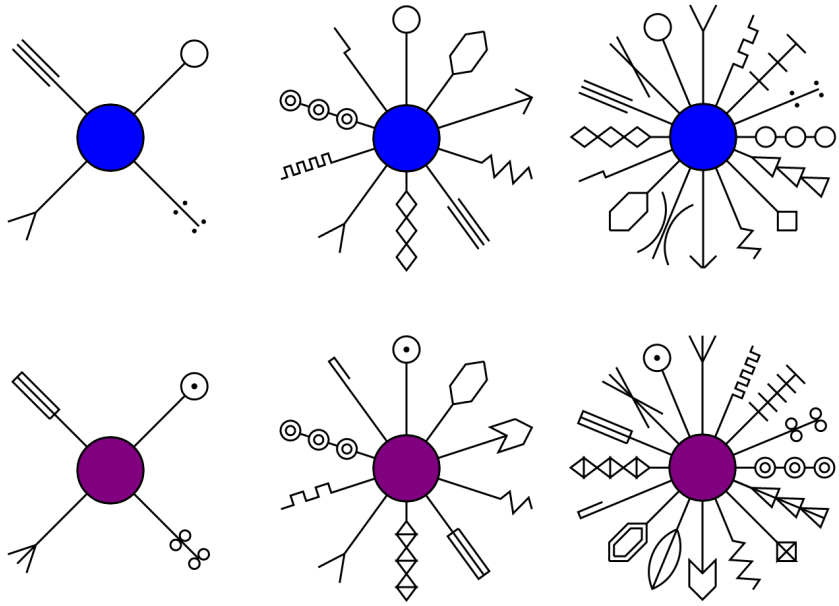


Figure 1. Example stimuli, displaying two instances from each of the three possible *Dimensionality* conditions (4, 10, and 16, from left to right). Features were binary and correspond to the legs of the amoebas. Together, the two 16-FEATURE examples show all possible feature values.

results since they were qualitatively identical. Participants ranged in age 18 to 76 (mean 34.2). They were paid US$2.00 for completion of the experiment, which took roughly 12 minutes. Data from an additional 42 participants were excluded from analysis, either from failure to complete the task (37 participants) or participating in a pilot version of the study (5 participants).

Design. The experiment presented people with a supervised category learning problem, in which they were asked to classify an amoeba stimulus as either a bivimia or lorifen. Each amoeba consisted of a circular base with a set of binary features (legs). The full set of 16 unique pairs of features are shown on the two stimuli in the right column of Figure 1.

Nine experimental conditions were created by manipulating the *Dimensionality* of the stimuli and the *Structure* of the category in a 3 × 3 between-participants design; people were randomly assigned to each condition. The three levels of *Dimensionality* reflect the number of binary features present on the stimuli: 4-FEATURE (N = 302), 10-FEATURE (N = 277), or 16-FEATURE (N = 307). For the lower-dimensionality conditions, the set of displayed features were a randomly selected subset of the features used in the 16-FEATURE condition. The position of features on the amoeba were randomized differently for each participant.

The three category *Structures* were designed in the following way. In every condition there was one feature (chosen randomly) that was 90% predictive of the category label, such that 90% accuracy could be achieved by using that feature alone. However, the predictiveness of the other features differed as a function of *Structure* condition. In the SINGLE condition (N = 294), all other features were completely non-predictive (i.e., the value of that feature predicted a given label 50% of the time). As such, the best performance in the

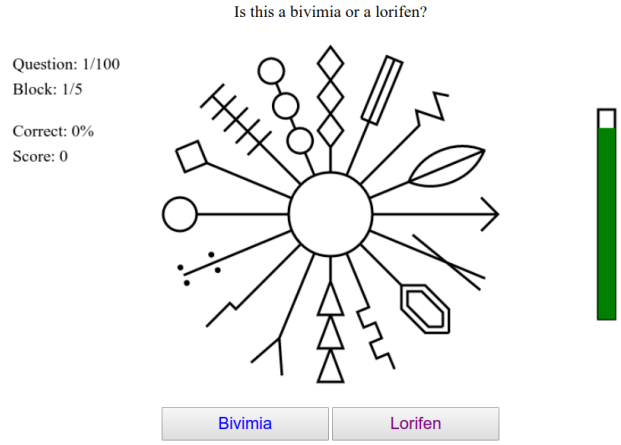


Figure 2. An example of a trial from the 16-FEATURE condition. Participants were asked to classify the amoeba as either a bivimia or a lorifen. In one version of the experiment, a green timer on the right was displayed to incentivize participants to respond faster for additional points, which they were given for correct answers only. Another version of the experiment was also run that did not include the timer.

SINGLE condition would be achieved by identifying the single predictive feature and making categorization decisions using only it. By contrast, in the ALL condition ($N = 301$), all of the features were 90% predictive, matching a family resemblance structure. As a consequence the best possible performance is achievable by aggregating the information provided by all features. Finally, in the INTERMEDIATE condition ($N = 291$), the other features were 70% predictive. Thus, one feature was most diagnostic but it would be theoretically possible to achieve better performance by using all of the features in concert.

Procedure. The experiment consisted of five blocks of 20 learning trials each, resulting in a total of 100 trials. On each trial people were presented with an amoeba as shown in Figure 2 and were instructed to classify it as either a bivimia or a lorifen.[1] People received points for correct answers but did not lose points for incorrect ones. In one version of the experiment ($N = 439$) people were given as much time as they wanted to respond; in the other ($N = 447$), they were still given as much time as they liked but they saw a timer (the green bar on the right of Figure 2) that slowly decreased, and they received more points for faster answers. There were no differences in performance between these two versions so the data was pooled and results reported are from the combined dataset[2].

Participants were given feedback which was displayed for three seconds. It consisted

[1]In all conditions, the stimuli were generated probabilistically and independently of one another, rather than pre-generating 100 specific stimuli and showing the same ones to everybody. To perform this, one of the two categories was randomly selected on each trial, and then the features of the stimulus were generated according to the conditional probabilities based on the category structure for that condition.

[2]For the main analyses reported in this paper, we performed a model comparison between a model with TIMER as an additional discrete predictor and model without. Results from this model comparison for both experiments produced Bayes factors that favoured the models which did not include an effect for TIMER. This suggested that the presence or absence of the timer had little effect on participant's accuracy in the task, so it was sensible to pool the data into a single dataset.

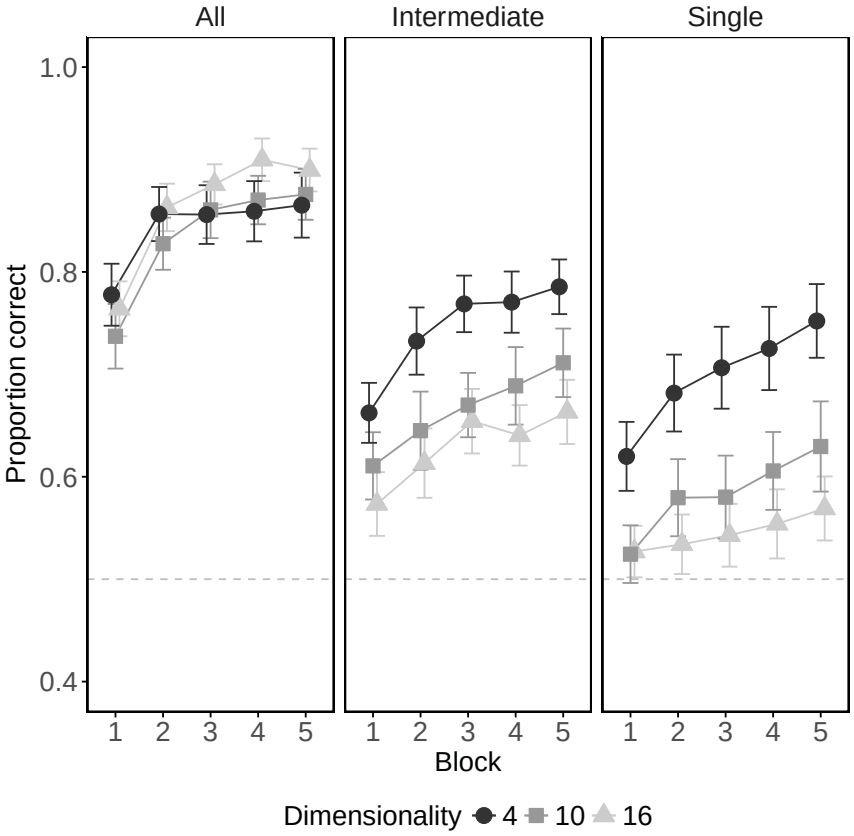


Figure 3. **Accuracy in Experiment 1**. Human learning across the *Dimensionality* and *Structure* conditions. While learning within the ALL condition was unaffected by the number of features, more features led to poorer performance in the SINGLE and INTERMEDIATE categories. Error bars depict 95% confidence intervals, and the dotted line reflects chance performance.

of a correct or incorrect message, the number of points earned, the correct category label, and a change to the color of the circular base of the amoeba to indicate category membership (blue 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 20 trials, people were given a short summary of their performance, showing them their accuracy and points earned in the current block and all previous blocks.

Results

Participants learned well in all conditions, with accuracy increasing across training block (Figure 3). We quantified this effect through the use of Bayesian mixed effects model comparison in which we compared a baseline model that contains only a random intercept for each participant to a model that includes a linear effect of *Block*.[3] The Bayes factor for this comparison ($BF > 10^{77} : 1$) overwhelmingly favors the model that includes an effect of

[3]All mixed effects models in this paper assume a random intercept for each subject. Bayes factors were calculated using the default parameters (Rouder, Morey, Speckman, & Province, 2012; Liang, Paulo, Molina, Clyde, & Berger, 2012) of the BayesFactor package 0.9.12-2 (Morey & Rouder, 2015) in R 3.2.3.

Block. The posterior estimate of block shows a positive slope of 0.025 (95% CI is 0.023 to 0.028) indicating that average accuracy increased by about 2.5% for each block of training.

While it is reassuring that there is a general improvement in accuracy throughout training, one of our main questions was whether accuracy differed as a function of category *Structure*. Figure 3 suggests two things: first, that accuracy in the ALL structure is much higher than the INTERMEDIATE and SINGLE category structures; and second, that the learning rate may be identical across all category structures. To investigate the first issue, we evaluated whether there was an effect of *Structure* on accuracy. Indeed, a model with two predictors (*Structure*, coded as a three-level categorical variable, and *Block*) is strongly preferred over a model containing only *Block* ($BF > 10^{90} : 1$). Posterior estimates reveal that accuracy in the ALL condition is 0.16 higher (95% CI is 0.12 to 0.20) than the INTERMEDIATE structure, which is slightly higher than the SINGLE (0.08 more, 95% CI from 0.04 to 0.12). In order to investigate the second issue, we compared the *Structure* and *Block* model to a more complex model that also included an interaction between *Structure* and *Block*. The model without an interaction is strongly preferred ($BF = 40 : 1$), suggesting that the rate of learning across blocks is not different in the three *Structure* conditions.

Our second question was whether there is evidence for an effect of stimulus *Dimensionality* on performance. We found that there was: a model containing *Dimensionality* (coded as a three-level categorical variable) and *Block* was strongly preferred over a model containing only *Block* ($BF > 10^{7} : 1$). The posterior estimates of the effect of number of dimensions show that the only reliable difference was between the 4-FEATURE and 16-FEATURE conditions, with the 4-FEATURE one being on average 0.08 more accurate (95% CI is 0.03 to 0.13); all other 95% confidence intervals of the difference span zero (4-FEATURE vs. 10-FEATURE and 10-FEATURE vs. 16-FEATURE).

Of course, we are less interested in whether dimensionality or category structure *alone* has an effect on learning, and most interested in whether there is an interaction: do more stimulus dimensions, as hypothesized, hurt learning in the SINGLE category structures but not in the ALL category structures? To evaluate this, we compared a Bayesian mixed effects model containing *Block*, *Structure*, and *Dimensionality* alone to a model with these three variables plus an interaction term between *Structure* and *Dimensionality*. This shows strong evidence in favor of the model containing the interaction ($BF > 10^{8} : 1$), indicating that additional features have different effects on learning in different category structures.

What differences drive this interaction? Figure 3 suggests that accuracy in the INTERMEDIATE and SINGLE category structure conditions decreases much more strongly as the number of stimulus features increases. In order to investigate this quantitatively, we conducted a post-hoc analysis of the effect of *Dimensionality* on accuracy within each category structure. The results, shown in Table 1, indicate the Bayes factor in favor of a mixed effects model containing *Dimensionality* and *Block* relative to a model containing only *Block*. The results show that for the INTERMEDIATE and SINGLE structures, the model with the higher Bayes factor includes the *Dimensionality predictor*, suggesting that the number of features affects learning for these structures. However, for the ALL structure the preferred model based on its Bayes factor is one without *Dimensionality* as a predictor. This is consistent with our hypothesis that additional features should hurt learning much more strongly when categories do not follow a family resemblance structure.

	All	Intermediate	Single
Bayes factor	0.3:1	$\mathbf{10^8 : 1}$	$\mathbf{10^{11} : 1}$
4 vs. 10	0.06 (-0.05 to 0.06)	**0.08 (0.01 to 0.14)**	**0.10 (0.03 to 0.19)**
10 vs. 16	-0.02 (-0.08 to 0.03)	0.04 (-0.03 to 0.10)	0.03 (-0.04 to 0.11)
4 vs. 16	-0.02 (-0.08 to 0.04)	**0.11 (0.05 to 0.18)**	**0.15 (0.07 to 0.22)**

Table 1

Bayes factors and parameter estimates for the post-hoc analyses of the effect of stimulus Dimensionality for each category Structure in Experiment 1. The first row indicates the Bayes factor in favour of a model with Dimensionality and Block as predictors relative to a model with only Block; all other rows show the posterior estimates of the differences between Dimensionality conditions within that category structure. Results indicate that the effect of stimulus dimensionality was larger in the SINGLE *and* INTERMEDIATE *than the* ALL *category structure. The 95% confidence interval estimates are shown inside the brackets.*

Summary

Experiment 1 suggests that increasing the number of features has a differential impact depending on the underlying category structure. In the two conditions that contain a single highly predictive feature and other features that are less predictive (SINGLE and INTERMEDIATE), learning is clearly improved when there are fewer features overall. This is most evident in the final two columns of Table 1, which show 10-14% increases in overall accuracy for learning from four rather than 16 features in the INTERMEDIATE and SINGLE conditions. The same advantage does not occur in the ALL category structure .

The fact that learning was not impaired in the ALL category structure may not be particularly surprising, given that all features were equally useful and there were no features that were less predictive. In that sense it is the lack of *advantage* for more features that is perhaps more surprising, especially since other studies have shown a learning advantage when there are additional features (e.g., Hoffman & Murphy, 2006; Hoffman et al., 2008). One possibility here is that performance in the ALL condition reflects a ceiling effect. Since all of the features were 90% predictive, it could be that the task was quite easy no matter how many features there were. We test this directly in Experiment 2 by investigating only family resemblance structures, but manipulating the degree to which the features are predictive of the category label.

Experiment 2

This experiment explicitly tests whether additional features have an effect on category learning within family resemblance categories when the features are less predictive than in the previous experiment. If there is no effect of the number of features on learning, we can be more certain that the differences due to category structure found in Experiment 1 are actually due to category structure rather than to the informativeness of the features. We test this by systematically manipulating the predictiveness of the features. Does this affect the degree to which additional features affect learning?

Method

Participants. 888 people (459 male, 425 female, 4 other) were recruited via Amazon Mechanical Turk. As before, the high number of participants reflects the fact that we ran two experiments with slightly different methodologies and pooled the results since they were qualitatively identical ($N = 436$ for the version without the timer, $N = 452$ for the version with it). Participants ranged in age from 19 to 74 (mean 34.6). They were paid US$2.00 for completion of the task, which took 12 minutes. Data from an additional 37 participants were excluded from analysis, either from failure to complete the task (32 participants) or participating in an earlier version of this study (5 participants).

Design. The task and stimuli were identical to Experiment 1, with participants randomly allocated in a 3×3 between-participants design. As before, we manipulated the *Dimensionality* by altering the number of features present in the stimuli to make three conditions: 4-FEATURE (N=286), 10-FEATURE (N=327), and 16-FEATURE (N=275). Unlike before, all the category structures were family resemblance structures, with all features being equally predictive of the category. This time we manipulated the degree of *Predictiveness* to make three conditions: 70% predictive (N=310), 80% predictive (N=258), and 90% predictive (N=320). The 90% condition was a replication of the ALL structure in Experiment 1.

Procedure. The procedure was identical to Experiment 1. Similar to the previous experiment, in one version of the experiment ($N = 436$), there was no time limit for providing a response on each trial. In the other version of the experiment ($N = 452$), there was still no time limit, but they saw a timer that slowly decreased (see Figure 2), and they received more points for faster responses.

Results

How was learning affected by *Dimensionality* and *Predictiveness*? We evaluated this question by comparing Bayesian mixed effects models that included some combination of *Block* as a continuous variable and *Dimensionality* and *Predictiveness* as discrete variables. Reassuringly, we found that people did indeed learn over the course of training: a model including *Block* was strongly preferred over a model that only contained a random effect for each participant ($BF > 10^{74}$:1). As before, posterior estimates suggest that average accuracy increased by about 2.3% for each block of training (95% CI is 0.021 to 0.026).

How did the *Predictiveness* of features affect learning? As Figure 4 shows, and as one would expect, overall learning was lower in categories with lower predictiveness. This is borne out in a Bayesian model comparison between a model with *Predictiveness* (coded as a three-level categorical variable) and *Block* as compared to a model with only *Block* as a predictor. The two-predictor model was strongly preferred ($BF > 10^{116} : 1$), and the posterior estimates indicate that accuracy was 11% higher for both the 90% to the 80% condition (CI is 0.07 to 0.14) as well as the 80% to the 70% condition (CI is 0.08 to 0.15).

Our main question, of course, was whether additional number of features had an impact on categorization accuracy. Figure 4 suggests that *Dimensionality* does not have an effect on learning, and a Bayesian mixed effects model comparison confirms this: a model containing only *Block* was preferred ($BF > 17 : 1$) over a model containing both *Dimensionality* and *Block*. This is further supported by post-hoc analyses that find strong preference

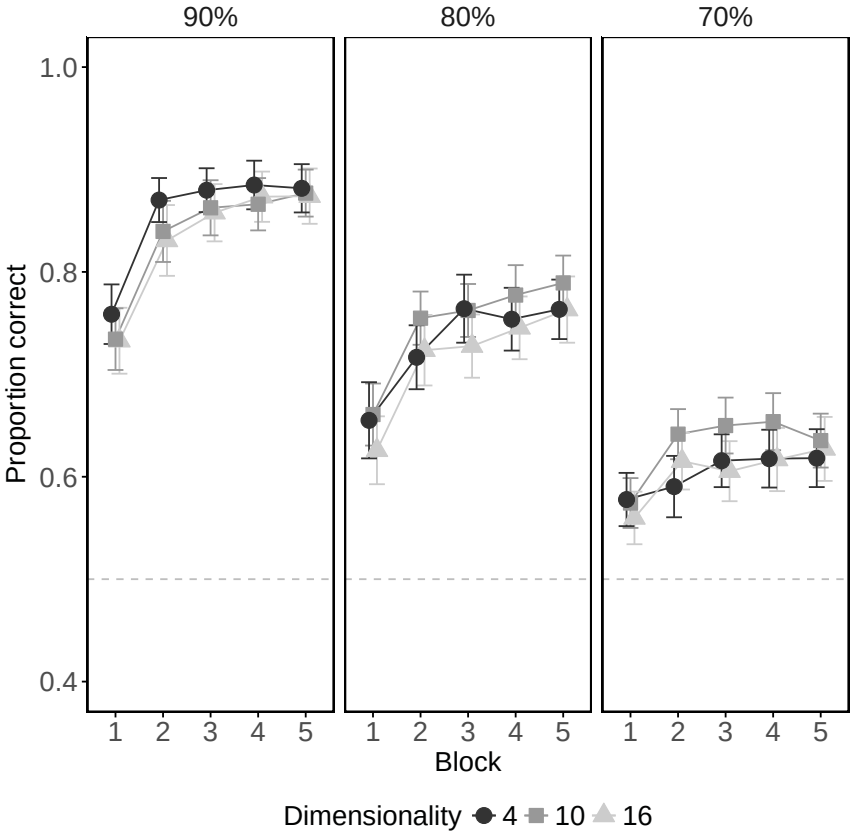


Figure 4. **Results from Experiment 2**. Mean accuracy across the three *Predictiveness* and *Dimensionality* conditions. While the mean performance decreased as the level of *Predictiveness* was reduced, within each *Predictiveness* condition there was no change based on the number of features. Error bars show 95% confidence intervals, and the dotted line reflects chance performance.

for a model containing *Block* and *Predictiveness* predictors over all models containing *Dimensionality*.

Summary

Experiment 2 provides strong evidence that the number of features does not affect learning when the category structure follows a family resemblance pattern, and that this cannot be attributed to a ceiling effect. Interestingly, learning in the 70% predictive family resemblance category structure is only slightly above chance ($M = 0.63$ in the final block). Despite the fact that there was evident room for improvement, there was no benefit of increasing the number of features, suggesting that these results do not reflect a floor effect either. Taken in conjunction with Experiment 1, these findings suggest that the curse of dimensionality affects people more with category structures where correct classification relies on a single feature. Indeed, in family resemblance categories, there appears to be no detrimental effect of additional features at all (but neither is there much benefit).

Category learning models

These empirical results deserve some theoretical explanation. Our design was motivated by the intuition that the curse of dimensionality *should* have a differential impact in different situations, in a fashion not dissimilar to what we actually observed. Specifically, we predicted that if people learn by *searching* for predictive features, then the curse of dimensionality should hurt performance in situations where only a few features are predictive, but should have little impact when many features are predictive. Crucially, this prediction assumes that people have capacity limitations that prevent them from learning or using all of the features at once. In the family resemblance conditions, for instance, adding more features means that each stimulus contains more independent information about the category label, and one might intuitively expect performance to *improve* when more information is made available. No such effect is evident in our data, suggesting that capacity limitations play a critical role in governing category learning.

Armed with these insights, in the rest of this paper we evaluate three category learning models that incorporate different capacity limitations on the learner. Which best explains the empirical data from Experiments 1 and 2? The three models are designed to vary systematically in the strength of the capacity limitation they impose. At one end of the spectrum, we consider a statistical learning model that can attend to all stimulus information available in the task, learns by Bayesian belief updating, equivalent to a probabilistic prototype model. At the other end, we consider a hypothesis testing model that can only attend to a single feature at a time and learns by applying simple belief updating rules. In between these extremes we consider a statistical learning model in which learning and decision-making are limited.

Notation

We briefly describe the notation used to describe the each of these models. The input for each trial is a D-dimensional stimuli vector $\mathbf{x} = (x_1, x_2, \ldots, x_D)$, where D is the dimensionality of the stimulus and each x_i is a binary feature, i.e. $x_i \in \{0, 1\}$. The predicted category response $\hat{y} \in \{0, 1\}$ for trial N is defined by the feature information from trial N along with the representation learned by the model based on the previous $N - 1$ trials.

An ideal observer model

First, we describe the details of an ideal statistical learner which we call OPTIMAL. In our experiments, the stimuli were generated by following the principle of class-conditional independence (e.g., Anderson, 1990; Jarecki, Meder, & Nelson, 2013). As long as one knows the true category label y, then the probability of any particular feature value x_i is completely independent of any other feature. As a consequence, every category can be represented in terms of a single feature vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)$ where $\theta_i = p(x_i|y)$ describes the probability that feature i will have value x_i. Although class-conditional independence is not always satisfied in real life where feature correlations are possible (Malt & Smith, 1984), it is a reasonable simplification in many situations (Jarecki et al., 2013), and one that is appropriate to our experimental design. Moreover, because the category can be represented

in terms of a single idealised vector $\boldsymbol{\theta}$ that describes the central tendency of the category, it is broadly similar to standard prototype models (Posner & Keele, 1968).[4]

Formally, we implement this statistical learning model using a naive Bayes classifier which makes the same assumption of class-conditional independence. In it, the posterior probability that novel object $\boldsymbol{x}$ belongs to the category y is given by:

$$p(y|\mathbf{x}) \propto \prod_{i=1}^{D} p(x_i|y)p(y) \tag{1}$$

where the marginal probability $p(x_i|y)$ is given by the posterior expected value of θ_i given the previously observed category members. Specifically, if the learner has observed n_y previous exemplars that belong to category y, of which n_{yi} were observed to have the feature x_i, then the model estimates the following probability:[5]

$$p(x_i|y) = E[\theta_i|n_{yi}, n_y] = \frac{n_{yi}+1}{n_y+2} \tag{2}$$

Applying a similar logic, the model learns the base rate of the category labels over time, and so the prior probability $p(y)$ of category y is computed by applying a (smoothed) estimate of the observed base rate so far:

$$p(y) = \frac{n_y+1}{n+2} \tag{3}$$

Finally, as an ideal observer model, the OPTIMAL model is assumed to always choose the category label with highest posterior probability, and thus the response $\hat{y}$ is selected deterministically by applying the rule:

$$\hat{y} = \arg\max_y p(y|\mathbf{x}) \tag{4}$$

The OPTIMAL model is appealing for two reasons. Firstly, it serves as an ideal observer model for this experiment, insofar as it is a statistical learning model whose structure precisely captures the structure of the task (i.e., conditional independence) and learns the specific categories by applying Bayes rule. As such it can reasonably be claimed that the performance of this model represents the upper bound on what might be achievable in the learning task. Secondly, because of its connection to prototype models, it may be taken as a representative of a broad class of "family resemblance models" that have dominated the theory of category learning since the 1970s. This model is not intended to be a fully general model of human categorization, but rather to act as a gold standard to compare to human performance in these tasks across different levels of dimensionality, category structure, and feature predictiveness.

[4] Although we do not explicitly evaluate any exemplar models (Nosofsky, 1986) or mixture models (Sanborn, Griffiths, & Navarro, 2010; Love, Medin, & Gureckis, 2004), we expect that their behavior would be very similar to the prototype model on these category structures. Exemplar models are perfectly capable of learning prototype-like category structures (Nosofsky, 1988), and as such we would not expect this experimental design to be predictive as regards to the prototype vs. exemplar distinction. Rather, we expect that the lessons implied for the optimal model would turn out to be similar for exemplar models and any other sufficiently rich statistical learning model.

[5] Formally, this expression arises if the learner places a uniform prior over an unknown Bernoulli probability θ_i and updates those beliefs via Bayes' rule. It is equivalent to the Laplace smoothing technique.

A hypothesis testing model

The second model we describe is a hypothesis testing model which we call RULE, that classifies objects into categories by considering hypotheses based only on a single feature. On each trial, the model considers a single hypothesis h_{ia} for the rule that defines how it makes categorization decisions. All of the rules in the hypothesis space take the following form: `if` $x_i = a$ `then` $\hat{y} = 0$`, otherwise` $\hat{y} = 1$, such that each rule learns to use one feature (a) to predict the category outcome ($\hat{y}$). The space of hypotheses is defined by the set of features. As an example, a particular hypothesis the model might use is: `If the third feature takes value 0` (i.e. $x_3 = 0$)`, then respond bivimia` ($\hat{y} = 0$)`, otherwise respond lorifen` ($\hat{y} = 1$)`.` Relative to other rule-based models in the literature (e.g., Nosofsky, Palmeri, & McKinley, 1994; Goodman, Tenenbaum, Feldman, & Griffiths, 2008) this is fairly simplistic because it can never adapt and use more than a single feature to make a category judgment. This was deliberate because we wanted to consider a model at the other end of the spectrum, capturing the important intuition that the learner attends to and uses only one feature at a time.

The RULE model learns by updating the utility u of every hypothesis in the hypothesis space. All utility values are initialized to 0.5 at the start of the learning process and are bounded between 0 and 1. At the end of each trial, the model updates the utility of the currently-considered hypothesis only, and it does so by assuming that the utility is proportional to the number of correct decisions that the rule has led to on those trials where the learner was considering that rule. Formally, this utility function is denoted:

$$u(h_{ia}) = \frac{1 + (\text{correct predictions with } h_{ia})}{2 + (\text{trials with } h_{ia})} \tag{5}$$

At the end of every trial, the model updates the utility of the current hypothesis. If it makes the correct prediction, the hypothesis is retained for the next trial, otherwise it is discarded and a new hypothesis is selected from the set of hypotheses with probability proportional to the utility, as in Equation 6:

$$p(h_{ia}) = \frac{u(h_{ia})}{\sum_{x,y} u(h_{xy})} \tag{6}$$

A limited capacity statistical learner

The OPTIMAL and RULE model differ in several respects, and if one of them learns in a more human-like fashion than the other we would like to know *why* this is the case. The OPTIMAL statistical model employs a category representation that closely mirrors a probabilistic prototype, whereas the RULE model represents categories using simple decision rules. The OPTIMAL model updates its category representations using all the information available to it, whereas the RULE model only updates its beliefs about the one specific rule it is currently considering. Finally, the OPTIMAL model makes its categorization decisions by always choosing the most likely category, whereas the rule based model – though also deterministic – is entirely capable of following a particular rule to make a correct decision and then immediately discarding that rule.

Given these differences, we developed a LIMITED model variant of the OPTIMAL statistical learning model that retains the prototype-style representation but is limited in both how many features to use when making decisions and the ability to learn from their observations.

As in the OPTIMAL model, the category label selected on a given trial by the LIMITED model is dictated by Equation 4. However, instead of multiplying across all features when making a decision as in Equation 1, a single feature f drives decision making:

$$p(y|\mathbf{x}) \propto p(x_f|y)p(y) \tag{7}$$

Learning is also limited in this model, which we implemented by applying Equation 2 to only a *single* feature on each trial. This limitation captures the same qualitative principle that underpins the single-hypothesis belief updating procedure used by the RULE model. Highlighting this connection, the updating process of the LIMITED model shifts its attention across features using a utility-based rule that is almost identical to Equations 5 and 6 for the RULE model:

$$u(f_i) = \frac{1 + (\text{correct predictions with } f_i)}{2 + (\text{trials with } f_i)} \tag{8}$$

$$p(f_i) = \frac{u(f_i)}{\sum_x u(f_x)} \tag{9}$$

The LIMITED model is thus very restricted: it absorbs information only from a single attended feature, and this is the only feature that contributes to the categorization decision. The differences between this model and the RULE model are fairly modest.

Model results

Each of the three models were simulated 10,000 times in each of the experimental conditions from both experiments, where each simulation mimicked a 100-trial experiment. On each trial, a new stimulus was generated in exactly the same manner as the experiment. The model then made a prediction of the category label of the current stimulus, and then received feedback which it would use to update its category representation.

Figures 5 and 6 shows the correlation between human performance and the predictions from each of the three models across each condition. There are a number of observations we can make based on the different scatterplots shown here. First, the results show that the OPTIMAL model consistently outperforms humans in all conditions. While this provides a theoretical upper bound for how well people could do in the task, the results show that this limit was not attained in any of the conditions.

Second, both figures demonstrate that the predictions from both the LIMITED and RULE models fit reasonably well to the human data. In some of the conditions, they both appear to underpredict human performance, but they nevertheless capture the main qualitative patterns across both experiments. In particular, when categories are not family resemblance structures, both models predict that more features should hurt performance, and when the category structure is family resemblance based, both models predict that additional features should make no difference. Since this behavior precisely matches the

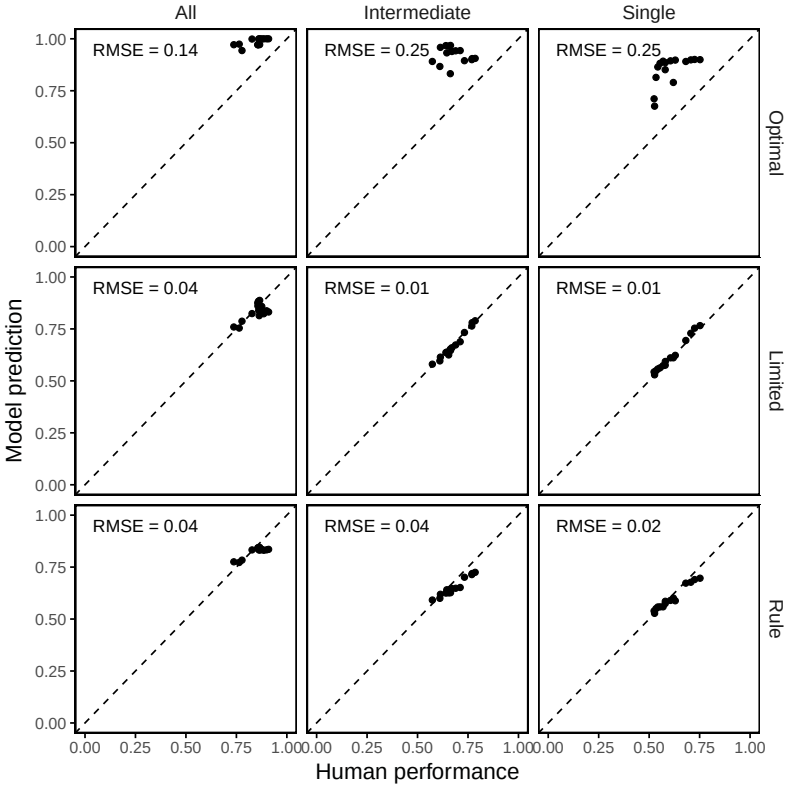


Figure 5. **Scatterplot comparing the predictions of the three computational models to human performance in Experiment One.** Each panel contains a dot for the performance at each block and across all *Dimensionality* conditions. Our results show that the OPTIMAL model consistently overestimates performance, while both the LIMITED and RULE models both closely match human performance across the different conditions in Experiment One.

qualitative pattern shown by people, it is perhaps no surprise that the RULE and LIMITED models provide a strong quantitative fit to human performance.

Perhaps somewhat disconcertingly, neither the quantitative nor the qualitative fits give compelling reason to prefer either the RULE or LIMITED model over each other. That said, it is important to realise that this analysis so far reflects *aggregate* data: how well each model predicts the overall population average amongst our participants. Yet we know that population averages may be highly misleading when the goal is to infer what kinds of individual processes give rise to the behavior in question.

For this reason we also calculated which model best fit each of the individuals in each of the experiments (as reflected in the RMSE between each individual's accuracy and the model predictions from the same experimental condition as the individual), as shown in Figure 7. A random guessing model was also included in the set of models, where the performance for each block was set at chance level (50%). Although there is substantial variation across people, in the majority of conditions most people's performance best matches the LIMITED model. The only exception is that in the most difficult conditions a substantial number are best fit by the RANDOM model, suggesting these participants were

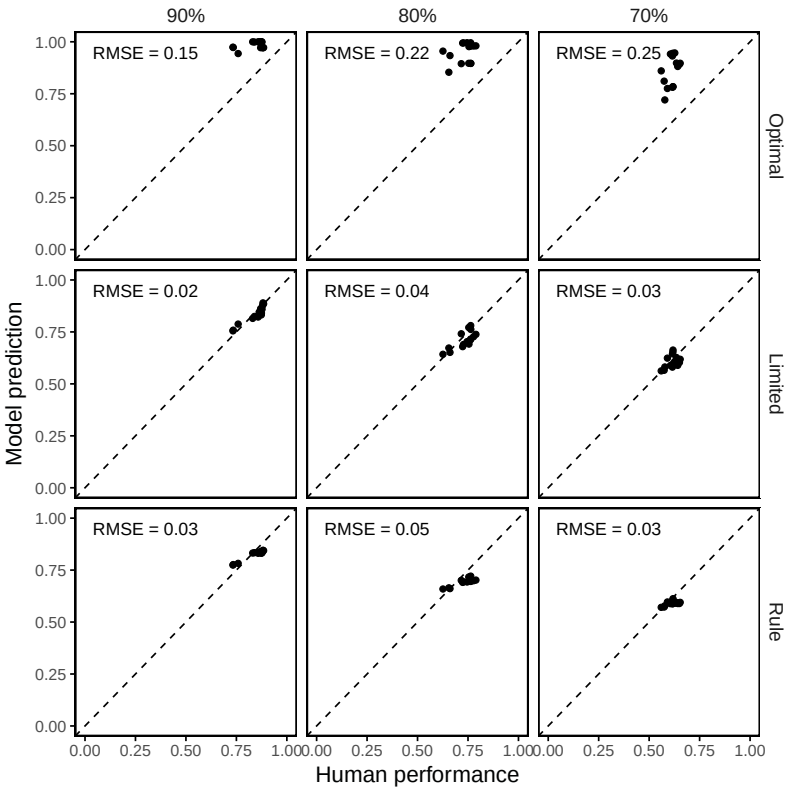


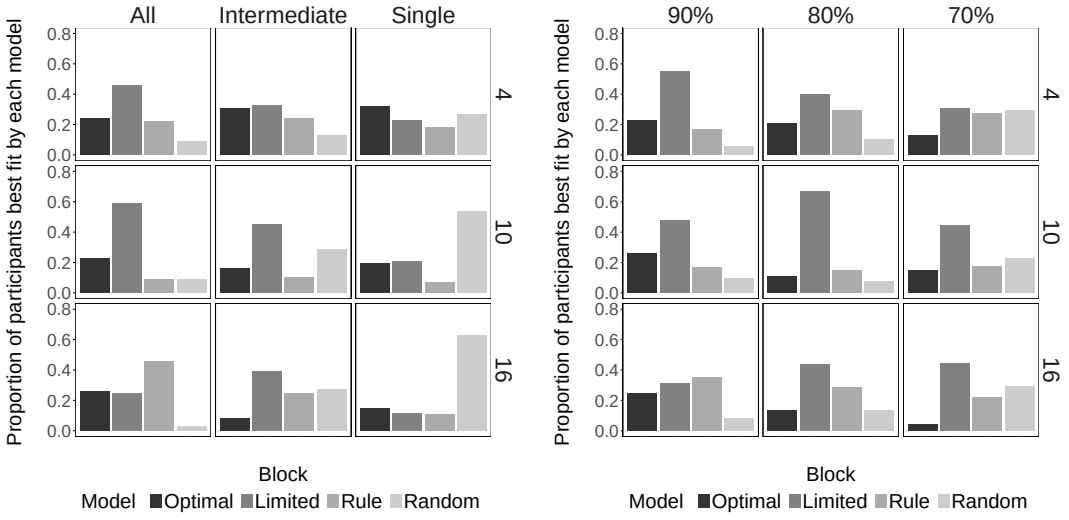
Figure 6. **Scatterplot comparing the predictions of the three computational models to human performance in Experiment Two.** Each panel contains a dot for the performance at each block and across all *Dimensionality* conditions. Similar to the scatterplot in Figure 5, we find that the OPTIMAL model performs far better than humans on this task, while both the LIMITED and RULE models both track human performance closely.

unable to learn the task. The OPTIMAL model describes performance the least well of the three theoretically motivated models.

Discussion

The term "curse of dimensionality" has been applied to a range of problems in machine learning, statistics and engineering, all of which share the common property that the space of possible solutions to an inference problem grows extraordinarily rapidly as the dimensionality increases. The same phenomenon applies to human category learning, and our goal in this paper has been to explore how the curse plays out for human learners.

At an empirical level we observed a clear pattern in which the curse of dimensionality is strongly mediated by the structure of the categories that need to be learned. Categories like those in the SINGLE condition, in which only a small number of features are relevant for predicting category membership, are heavily affected by dimensionality because the search problem (identifying the predictive feature) becomes harder as more irrelevant features are added. In contrast, the number of features does not appear to affect learning for family



(a) Experiment 1 (b) Experiment 2

Figure 7. **Proportion of individuals who are best fit by each of the models in both experiments.** For each individual, we calculated the RMSE between their performance and each of the computational models, and assigned them to the model whose predictions led to the smallest RMSE. Across most of the conditions in both experiments, most people's pattern of performance most closely matched the predictions of the LIMITED model, and not the RULE model, despite the predictions of both models being very similar in the aggregate. The main exception to this pattern is in the more difficult conditions like the SINGLE category structure, in which a model consistent with random guessing did the best.

resemblance categories in which all features are somewhat predictive of category membership. A comparison of several computational models indicates that people's individual performance is best explained by model where learning and decision-making both proceed based on one feature at a time. This pattern was observed across the different category structures we tested, suggesting that people were not using different learning approaches to different category structures. Additionally, this result is consistent with theories of category learning that begin with single features that can be combined later on, such as Combination Theory (Wills, Inkster, & Milton, 2015), rather than beginning from the overall stimulus and divided into distinct attributes.

One of the main conclusions from this work is that it is not very meaningful to discuss the effect of dimensionality without considering what kind of category structure is being learned. In fact, the role of category structure may help explain away apparent differences in the literature. For instance, previous research indicating that additional features hurt performance used features that were not predictive of category membership (Bourne & Restle, 1959; Edgell et al., 1996), consistent with our SINGLE condition in Experiment 1. Other studies that found that adding features did not have any effect used family resemblance categories, consistent with the ALL conditions in both of our experiments (Hoffman and Murphy (2006) Experiments 1 and 2, Minda and Smith (2001) Experiment 4, Hoffman et al. (2008) Experiment 1). There were only two results we were not able to replicate, both reflecting an improvement in learning with more features (Hoffman and Murphy (2006)

Experiment 3 and Minda and Smith (2001) Experiment 1). However, in both of those studies, other aspects of category structure covaried with the number of features, providing an alternate explanation for results that differed from ours. For instance, in Experiment 1 of Minda and Smith (2001), the categories with fewer features were less structured, and when this confound was addressed in Experiment 4 of that paper, the effect went away.

Our computational work explains how these apparently qualitatively different effects for different kinds of categories can all emerge parsimoniously from one unified model. A model with limited capacity predicts that additional features should neither hurt nor help in categories with family resemblance structure, since subsequent features provide the same amount of information as existing ones. However, the same limited-capacity model should struggle in more rule-based categories, when learning involves searching over the space of features to identify the (few) predictive ones. The effect only occurs for a learner who is limited enough in capacity that they cannot simultaneously learn over all features at once.

Broader implications for human learning

The fact that human learning deviates systematically from the OPTIMAL model is theoretically interesting and highlights an important difference between real world learning and many category learning experiments. Both our experiments had features with *class-conditional independence*, in which the stimulus features are conditionally independent of one another as long as one knows the category to which the stimulus belongs. This assumption does not hold in general, but in some situations it might provide a reasonable first approximation. Indeed, people do appear to assume class-conditional independence, at least at first, in some category-learning tasks (Jarecki, Meder, & Nelson, in press).

However, from a computational modeling perspective, it is important to recognize the limitations that this assumption imposes: the reason that our ideal observer model is able to perform *better* on family resemblance categories as the number of features increases is that it exploits the fact that every additional feature conveys independent information about the category. When class-conditional independence holds, family resemblance categories become easier to learn as the dimensionality increases. This does *not* match the pattern we observed in our data, in which people's performance on family resemblance categories was the same regardless of the number of features in the stimuli.

Instead, the pattern of learning in these tasks is more accurately predicted by a capacity limited model that only processes a modest amount of information on each trial. However, the reason why people only process a limited amount of information is not clear. One possible interpretation is that memory limitations may restrict how people are able to encode these novel stimuli in memory or reason about the relationship between features and category labels. Depending on the relationship among features and between them and category labels, in real life one might end up making better categorization decisions by using a limited number of predictive features rather than attempting to process all information inherent in the stimuli. In other words, human learners might differ from our optimal statistical learning model not because of the limits of human cognition but because human cognition is shaped by an environment in which class conditional independence is a poor assumption, and that human learners are better described by other kinds of inductive biases.

The question of what other inductive biases are required to explain how humans are affected by the curse of dimensionality in some cases and not others is beyond the scope

of this paper, but we can speculate about possible answers. One possibility is an inductive bias for sparsity (Gershman, Cohen, & Niv, 2010), which assumes that only one (or a limited) number of features is relevant for categorization. Thus, the relevant features for this task could be learned through selective attention, a process where attentional weights for particular features increase or decrease based on their ability to make correct classification decisions. This kind of approach has been successfully employed by a number of existing models of categorization to explain other patterns in human category learning (Nosofsky, 1986; Kruschke, 1992), and is a potential future avenue of exploration for a richer explanation of how people learn categories with many features.

A second, alternative approach towards lifting the curse dimensionality is to reduce the number of features that are represented or encoded in the first place. Such methods focus on reducing the number of dimensions via manifold learning (Tenenbaum, 1997) or using structured representations (Kemp & Tenenbaum, 2009; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Lake, Salakhutdinov, & Tenenbaum, 2015). These kinds of approaches have been pursued with considerable success in semantic representation (Griffiths, Steyvers, & Tenenbaum, 2007; Landauer & Dumais, 1997). It is, of course, possible that human learning is versatile enough to incorporate the fundamental insights from both exploiting limited memory and attentional capacities and reducing the effective dimensionality of incoming stimuli. Pursuing these issues further is a matter for future work.

Acknowledgments

AP was supported by grants from the Australian Research Council (DP110104949 and DP150103280, with salary support from DE12010378). The salary of AH was supported by DP110104949.

References

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

Bellman, R. (1961). *Adaptive control processes: A guided tour.* Princeton, NJ: Princeton University Press.

Bourne, L. E., & Restle, F. (1959). Mathematical theory of concept identification. *Psychological review*, *66*(5), 278.

Chater, N., Clark, A., Goldsmith, J., & Perfors, A. (2015). *Empiricism and language learnability.* Oxford, England: Oxford University Press.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. In *American Mathematical Society Conference on Math Challenges of the 21st Century.*

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), 1463–1481.

Gershman, S. J., Cohen, J. D., & Niv, Y. (2010). Learning to selectively attend. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1270–1275). Austin, TX: Cognitive Science Society.

Goodman, N. (1983). *Fact, fiction, and forecast* (4th ed.). Cambridge, MA: Harvard University Press.

Goodman, N., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

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

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

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.

Jarecki, J., Meder, B., & Nelson, J. (in press). Naive and robust: Class-conditional independence in human classification learning. *Cognitive Science*.

Jarecki, J., Meder, B., & Nelson, J. D. (2013). The assumption of class-conditional independence in category learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2650–2655). Austin, TX: Cognitive Science Society.

Jones, F., Wills, A., & McLaren, I. (1998). Perceptual categorization: Connectionist modelling and decision rules. *The Quarterly Journal of Experimental Psychology Section B*, *51*(1b), 33–58.

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

Keogh, E., & Mueen, A. (2011). The curse of dimensionality. In C. Sammut & G. Webb (Eds.), *Encyclopedia of machine learning.* New York, NY: Springer.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2012). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 250–269.

Markman, E. (1989). *Categorization and naming in children.* Cambridge, MA: MIT Press.

McLaren, I., Leevers, H., & Mackintosh, N. (1994). Recognition, categorization, and perceptual learning (or, how learning to classify things together helps one to tell them apart).

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Minda, J. P. (2015). *The psychology of thinking: Reasoning, decision-making and problem-solving*. New York, NY: Sage.

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), 775–799.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for Common Designs [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.12-2)

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.

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

Searcy, S. R., & Shafto, P. (2016). Cooperative inference: Features, objects, and collections. *Psychological Review*, *123*(5), 510-533.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Tenenbaum, J. B. (1997). Mapping a manifold of perceptual observations. *Advances in Neural Information Processing Systems*, *10*, 682–688.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series. In *Proceedings of the 8th International Workshop on Artificial Neural Networks* (pp. 758–770).

Walker, C. M., & Bourne, L. E. (1961). The identification of concepts as a function of amounts of relevant and irrelevant information. *The American journal of psychology*, 410–417.

Wills, A., Inkster, A. B., & Milton, F. (2015). Combination or differentiation? two theories of processing order in classification. *Cognitive Psychology*, *80*, 1–33.

Wills, A., & McLaren, I. (1997). Generalization in human category learning: A connectionist account of differences in gradient after discriminative and non discriminative training.

The Quarterly Journal of Experimental Psychology: Section A, *50*(3), 607–630.

Wills, A., & McLaren, I. P. (1998). Perceptual learning and free classification. *The Quarterly Journal of Experimental Psychology Section B*, *51*(3b), 235–270.