Lateralisation of Exemplar and Prototype Controls in Category Learning

Sahil Loomba

May 8, 2016

Abstract

Human cognitive systems rely heavily on how knowledge is represented within these systems. One kind of knowledge, namely the categorical, is ubiquitous in everyday human interaction with the world. Two competing theories of how humans learn and infer from category knowledge have been popular in cognitive psychology: the exemplar theory and the prototype theory. Also, like most cognitive functions, category representations have purported to be lateralised in the brain: with the left hemisphere operating under "prototype" conditions, and the right hemisphere operating under "exemplar" conditions. This paper explores the hypothesis as to whether category knowledge representation is indeed lateralised in the light of these theories, and if yes, then whether a transfer of control between the two gets activated at some epoch of learning.

Keywords: *category learning, brain lateralisation, Bayesian cognition, model selection*

Contents

List of Algorithms

List of Figures

List of Tables

1 Introduction

1.1 Knowledge Representation: Categories

Knowledge constitutes one of the most basic foundations of human cognitive systems, may it be of language, reasoning, or thought. It is a well-structured form of information acquired through "learning", and is often used to exercise reason in the form of "inference", to make sense of the world[1]. A very important aspect of knowledge is how it is represented within our cognitive systems, since this can greatly influence the manner and time of retrieval of information from our knowledge base. As we will soon see, it can also influence the spatial location from where a knowledge fact is accessed, in the brain. For the purpose of this paper, let us consider one important kind of knowledge representation of **categories**.

Categorisation is a cognitive activity which comes so naturally to us, that it often goes unnoticed. We categorise every time we group similar things ("things" could loosely refer to objects or concepts) with each other. This is usually done to enable a more compact mental representation of things that share a significant number of characteristics or attributes, and efficient comparison across categories. An example would be to categorise face expressions into "happy" and "sad", or children learning to tell the difference between cats and dogs at a tender age. Therefore, the notion of **similarity** is important to be described , in the context of categorisation. One might argue that similarity means that certain properties hold invariably across the entire category/class. For instance, one might say "has four legs and is made of wood" to be one such property satisfied by all chairs. However, there can exist three-legged chairs made of plastic. Certainly, one can realise more and more general properties which are likely to be satisfied by most members of a class, say "is used to be sit upon" for the class of chairs, but it's difficult to find a universal invariant. Thus, rather than a hard condition of equality, a softer condition of resemblance in attributes is what makes members of a category similar to one another. However, there will always be some exemplars which are more "typical" of their respective class, than others. Say, how pigeons are more typical of the class birds than penguins.

In reaction time experiments, Smith et al. noted that subjects take lesser time to verify categories of more typical exemplars, something called the **typicality effect**[1]. Despite natural categories not having any strict rules of definition, people are able to acquire such a structure. Posner and Keele proposed that based on the exemplars, people slowly form an idea of the "central tendency" of the category, and then judge their categorical decisions based on this tendency, called a prototype[2]. The "closer" an exemplar is to the category prototype, and the farther it is from prototypes of other classes, the higher is its likelihood of being correctly classified. Two important outcomes of this are:

1. Performance on new samples ("transfer performance") is independent of the frequency of individual exemplar features, but depends on the distance of patterns from prototypes. Two operations could possibly be in play here: matching exemplar-specific information for

old patterns being classified better than new ones, and abstraction.

2. When a long delay is inserted between learning and transfer tests, subjects forget old exemplars more than the prototypes. That is, as retention interval increases, judgments are more likely to be based on prototypes and less likely to be based on exemplar-specific information.

1.2 Exemplar and Prototype Theory

There are two popular and competing models of category learning, often used to explain the observations enlisted above: (a) exemplar and (b) prototype theory[1].

Exemplar Theory suggests that when categorising a new object called "exemplar", our cognitive system compares the object characteristics to those of many exemplars already registered in our cognitive system. The typicality effect can be explained here by the fact that objects similar to many already-stored exemplars would be classified faster and correctly, compared to those which are similar to fewer exemplars.

Prototype Theory, given by Rosch (1973), says that when categorising a new object, our cognitive system compares it to prototypes of various categories, thus assigning it to the one whose prototype is the closest (in similarity) to the new object. Typicality here suggests that the more closely a new object resembles a prototype, the quicker it is assigned to the respective category. Note that here, what the object is being compared to is an *average case* of a preconceived category, whose existence is limited to our mental representation, and may not necessarily exist in the real stimulus set.

Although the fact that prototypical exemplars are classified better might appear to be in favour of the prototype theory, one can explain that using exemplar theory as well. Simply because a prototypical stimulus, in itself an exemplar, is bound to be more similar to exemplars of its category, and less similar to those of another, by its very construction. In fact, any naively intuitive model can easily explain why prototypical examples are more likely to be correctly classified[2]. A longer retention interval for old patterns can also be explained (using an exemplar argument) by the fact that with time delays, it becomes less likely for a probe stimulus to access its own mental representation. Thus, it is very important to design experiments to discriminate models of category learning with caution.

For testing both of these theories, cognitive scientists have designed experiments in support of either of the two, coming up with contrasting results. While some claim that exemplar approach is better in explaining atypical and abstract categories, others claim prototype approach better explains large-sized categories[1]. Some researchers have concluded that people use both models of learning, with prototypical model being used in early stages of learning, and exemplar model in later stages to handle class exceptions (since realising exceptions early can be damaging to learning). However, one could just as easily hypothesise the otherwise: when fewer instances

are seen, exemplar model allows a quick retrieval of information, whereas when instances of categories become too large in later stages of learning, prototyping aids efficiency of retrieval. Testing these contradictory hypotheses, and reconciling them is an important objective of this paper.

1.3 Lateralisation in Category Learning and Inference

Brain lateralisation has been a widely researched topic, and it has been well established that many cognitive functions are lateralised, that is, one of the two hemispheres is more specialised in the control of that function[3]. Anatomical correlates of hemispheric specialisation have been well studied, both at a macro level of fissure shapes, and at a micro level of cellular organisation. Split brain studies, wherein the corpus callosum is severed, have revealed important functional specialisations. The left hemisphere shows dominance in language and speech production. It extracts local features of stimulus, while the right hemisphere extracts the global ones. While the left side is described as being analytical and sequential the right one is holistic and parallel. Most arguments in favour of laterialisation suggest some kind of a parsimony that it achieves, from an evolutionary point of view.

If one looks at category knowledge learning, it is of key importance to look at two aspects: how the information is represented, or **encoded**, and how information is retained, or **memorised**. While the parietal lobes are critical for abstracting relations, the temporal lobe is the locus of memories of visual representations of shape[3]. The format of representation appears to be different across the hemispheres: while the left side retains abstract, categorical or prototypical information, the right side retains specific characteristics of the exemplars. That is, a prototypical model of learning is favourable in the left hemisphere, while an exemplar model is engaged in the right.

Marsolek (1995) tried to clarify this distinction by designing a unique experiment[3]. The stimuli set consisted of (a) eight prototypes, and an exemplar (b) training and (c) testing stimuli set made by distorting class prototypes. Using only the exemplar training stimuli set b, subjects were trained in stimuli classification. Followed by a speeded test classification task, using all three stimuli set (a, b and c). The stimuli was flashed in either the left or the right visual field. When classifying the training stimuli set b, which they had already studied, the subjects were faster (shorter reaction time) when the stimuli were presented in the left visual field (or to the right hemisphere), and vice-versa for the prototype stimuli set a. While this experiment does seem to suggest a **lateralisation of category learning**, in that the prototypical representations are stored in the left hemisphere, and exemplar ones in the right, the truth about which model is used for classification under usual stimuli circumstances remains occluded. That is, since very few unseen stimuli would be prototypical, for most general unseen stimuli (set c), whether one model is at all favourable over another, and if **category inference is lateralised** in general, remains to be seen. Looking for answers in this direction is a key motivation for this paper.

2 Prior Art

2.1 Context Theory of Classification Learning

Given by Medin et al.[2] the Context Model says that classification judgments are based on retrieval of stored exemplar information. Specifically, they assume that a probe stimulus functions as a retrieval cue to access information stored with stimuli similar to the probe. It attempts to represent the effect of hypotheses in terms of ease of storage and retrieval of information associated with the stimulus dimensions/attributes. They further assume that the likelihood of an exemplar e belonging to a category C is directly proportional to its similarity to exemplars of that category, e_C^+ and inversely proportional to exemplars of other categories, e_C^-.

For their data, they design 16 stimulus cards, by varying 4 binary valued attributes of geometric figues. Through a four-set experiment, the investigators conclude that the context model seems to fit their observations better than the independent cue (prototype) model. While they have themselves not assumed independence of cues (attributes) by using a multiplicative principle to incorporate the total effect of all cues together, they claim that prototype methods wrongly assume an independence of attributes, say that colour is independent of shape in perception of different geometric shapes. Thus, for the purpose of our own experiments, we need to design methods of establishing similarities between object instances, without assuming any independence of attributes.

2.2 Limitations of Exemplar-based Models and Category Abstraction

In this paper by Homa et al.[4] the authors discuss classification for ill-defined categories. Such categories, often termed as most "natural", are those wherein it isn't obvious as to what dimensions/attributes characterise a category, and when the variety of members in the categories can be potentially infinite. (Unlike the previous paper by Medin et al., which uses clearly constructed categories for classification.) Such categories can include musical pieces, hand-written characters, etc. Designing experiments around ill-defined classes allows us to vary the dimensions to great extents, thus coming up with a much bigger and richer stimuli set.

The authors also suggest that the predictions of a prototype model depend on how the prototype is itself characterised, what really accounts for "central tendency" and if information other than that is stored. They identify three possible sources of information: (a) the abstracted prototype, (b) specific information about the exemplars that defined the category during learning and (c) the boundary of the category. We will use these ideas to design our stimuli set for ill-defined categories. They further say, after conducting experiments on transfer delay, that the exemplar model has two major limitations: it overestimates the "category size effect"[1] for prototypes, and underestimates it for old patterns. Therefore in a mixed model setting,

[1]This effect refers to the ease of verification of a categorical fact for categories of smaller size. For instance, verifying "a poodle is a dog" is easier to verify than "a poodle is an animal".

they hypothesise that the following happens. Initially in learning, b kind of information about individual instances represents the category. As the degree of exemplar experience is increased however, a and c kinds of information become more useful. However, they do not provide a quantification of this hunch, as to exactly when in the learning process, this switch of controls happens.

2.3 Decoding Categorisation from Neural Implementation

In this study by Mack et al.[5] the authors apply a novel technique of neuroimaging analysis to add a neuroscientific perspective to the debate of how categorisation occurs in the brain. Without looking into details of the fMRI study, we look at how they fit the behavioural data to the exemplar and prototype models. They use the exact stimuli data as used in Medin et al. More importantly, they fit the observed data using standard model parametrisation.

The exemplar model posits that categories are represented by individual exemplars in memory. Let the exemplars themselves be represented in a multidimensional space, that is, $e \in \mathbb{R}^d$. (In this study, $d = 4$.) They represent the psychological distance between exemplars i and j by:

$$d_{ij} = \sum_{m=1}^{d} w_m |x_{im} - x_{jm}|$$

Thus, the similarity can now be represented by the exponential decay function as:

$$s_{ij} = e^{-cd_{ij}}$$

where c is a model parameter. Now, the probability that exemplar i is categorised into category A is given by:

$$P(A/i) = \frac{\left(\sum_{a \in A} s_{ia}\right)^{\gamma}}{\left(\sum_{a \in A} s_{ia}\right)^{\gamma} + \left(\sum_{b \in B} s_{ib}\right)^{\gamma}}$$

where γ is another model parameter, which dictates the "tightness" of similarity. Analogously, the prototype model posits that categories are represented by category prototypes, which come from the same multidimensional space as the exemplars, that is $p \in \mathbb{R}^d$. We can define similarity of an exemplar to a prototype just as above. Now, the probability that exemplar i is categorised into category A is given by:

$$P(A/i) = \frac{s_{i\hat{a}}^{\gamma}}{s_{i\hat{a}}^{\gamma} + s_{i\hat{b}}^{\gamma}}$$

where $\hat{a}$ and $\hat{b}$ are prototypes of categories A and B respectively. Using (in)correct categories assigned by the subject to latter half of the testing stimuli set, the two models' parameters can be found using standard maximum likelihood techniques, and the fits can be compared for each subject using a statistical test, such as the chi-squared test.

This study primarily suggests that concrete experiences during "learning" are stored in similarity-based cortical representations and that these representations are activated during later

categorisation or "inference". And although exemplar model appears to match the neural representation only slightly better, prototype predictions were also not entirely inconsistent with brain response.

3 Hypotheses

Taking from all the debates and past literature surrounding both category learning and lateralisation, this paper proposes to test and establish the following hypotheses:

1. There exists a **lateralisation** of the cognitive function of **category representation, learning, as well as inference** in the brain. While the left hemisphere stores and operates under the prototypical model, the right hemisphere operates under the exemplar model, giving rise to a mixed model of category knowledge representation.

2. The exemplar model is favoured over the prototype model for smaller category sizes. Given that Hypothesis 1 is true, and as learning of and inference from the two categories progresses, there exists an epoch when **cognitive control shifts from the prototype model to the exemplary model**, that is, the left hemisphere begins to dominate the right hemisphere in category knowledge representation.

4 Methods of Experimentation

4.1 Stimuli Data Design

We work with data on patterned shapes, akin more to the dot-line pattern dataset used by Homa et al. While in the dataset used by Medin at al. they purport that dimensions of the data are the very four dimensions defined by them (namely *colour*, *form*, *size* and *position*), and that every stimulus can be represented as a binary vector amounting to a total of 16 possible stimuli, we intuit that a cognitively-sound vector representation of every stimulus can and must be richer than just this. We must also uphold the ill-defined nature of most natural categories, and be able to generate a stimuli set of any size that we fancy. We thus divide the data generation process into three steps:

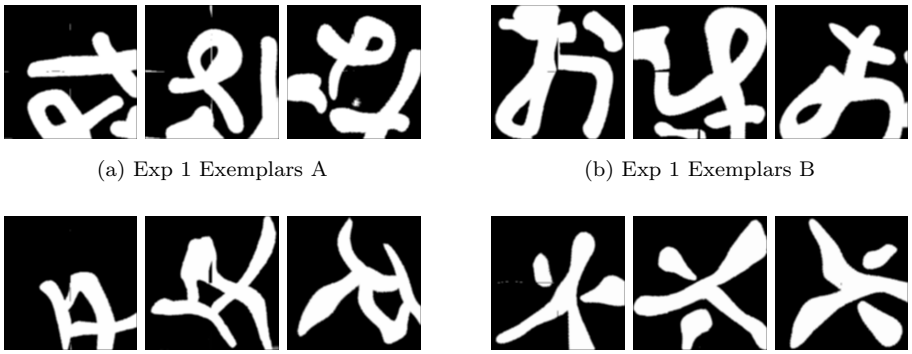
1. **Selecting Prototype Images:** So as to not engage multiple perceptive functions at the same time, and to ensure simplicity of the dataset, the m sized stimuli set would consist of two-dimensional geometric figures, differing only in their shape. To generate this dataset, two markedly different shapes are manually chosen as prototypes for the two categories. In their black-and-white $n \times n$ sized image form, they can be easily represented in the space of Boolean vectors $\mathbb{B}^{n^2}$. We use kanji script characters for both parts of our study, since none of the subjects are literate in Japanese, thus making the characters work well

as abstract patterns, ensuring that subjects treat them strictly as visual stimuli. While choosing the two prototypes, we need to enforce that they are at a "sweet-spot" of being neither too similar, nor too distinct. The images we use are given in Figure 1.

(a) Exp 1 Prototype A (b) Exp 1 Prototype B (c) Exp 2 Prototype A (d) Exp 2 Prototype B

Figure 1: Sample Kanji Characters used in Experiments 1 and 2 of this study

2. **Generating Exemplar Images:** To generate exemplars of a category, we pick its prototype image and apply a random-warp algorithm on it (see Algorithm 1 for details). We generate 40 exemplars for every category. Therefore in total, for an experiment of two categories, we have 82 images forming the image set I_m. Few of the exemplar images for both experiments are shown in Figure 2.



(a) Exp 1 Exemplars A (b) Exp 1 Exemplars B

(c) Exp 2 Exemplars A (d) Exp 2 Exemplars B

Figure 2: Sample Exemplars (image stimuli) used in Experiments 1 and 2 of this study

3. **Creating Representation Set:** Once we have I_m, we realise that this is a dataset of very high dimensionality. (Say if every image is of size 256×256, then our image space is of 65536 dimensions.) Moreover, these dimensions are most likely dependent on each other, which could make estimating distances between them difficult (since modelling how these dimensions depend on each other would be an insurmountable challenge.) Therefore, we use a technique of linear dimensionality reduction, called Principal Component Analysis, or PCA. PCA essentially projects the data onto dimensions which are independent to one another, by finding out dimensions of maximum data variance. Also, we can pick the top-d PCA dimensions which cover the significant variance in our data, thus reducing the

Algorithm 1 Random Warp Algorithm

Input: Prototype Image $I \in \mathbb{B}^{n \times n}$, number of anchor points k, randomness window w, number of exemplars m
for $i = 1$ to m **do**
    $A \leftarrow$ Pick k out of n^2 points, at random, as anchor points
    $A' \leftarrow$ Pick 1 new location for every anchor, randomly in a window of $\pm w$ around it
    $f \leftarrow splineFunctionMap(A, A')$ ▷ Use the spline transform to map points in original image to the warped image
    $I'_i \leftarrow f(I)$
    $I'_i \leftarrow randomRotate(I'_i)$ ▷ Randomly rotate the warped image
return I'

dimensionality of data from 65536 to d (we keep a small $d = 8$). We now obtain the Representation Image Set $R_m \in \mathbb{R}^8$, for the purpose of estimating psychological distance. Every representation vector is essentially a tuple of the weights of different eigenvectors corresponding to that particular image stimulus. Using PCA here is a good idea, since it extracts the most significant, and mathematically orthogonal dimensions of the data, thus eliminating the common fallacy of treating attributes independently.

4.2 Experiment Design

For testing lateralisation, since we are constrained by simple psychological experimentation techniques, we employ the Divided Visual Field (DVF) paradigm. This method works by projecting visual stimulus briefly (in the order of a few hundred milliseconds) onto the left and right visual fields :LVF correspond to the right hemisphere, or RH, and RVF corresponds to LH. (See Figure 3). Depending on the response times for the two kinds of stimuli, one can then either thwart or establish asymmetry in the cognitive function being tested. To conduct these computer-based DVF tests, we use an online programmable toolkit for conducting cognitive psychology experiments called PsyToolkit. However, just using response times could be a very non-robust method of establishing this asymmetry. Therefore, we use the metric of correctness of answers and a probabilistic model (described below) to do so. Experiments were written in the scripting language provided by PsyToolkit, and can be accessed online at this link.

Experiments will be carried out on close to 70 subjects, who were preferably male and right-handed, so as to rule out other factors affecting lateralised cognitive functioning. There was a pre-experiment survey to collect useful subject information, such as handedness, gender and vision correctness. For Experiment (and hypothesis) 1, subjects went through two phases. Phase 1, called the "learning phase", involved learning the categories from the 30 exemplars, selected and presented randomly on either the left or right side, with feedback on the actual category name turned on. Phase 2, called the "testing phase", involved a test of learning on the remaining

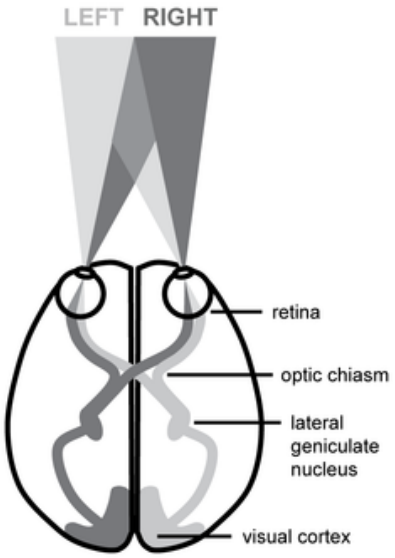


Figure 3: The DVF Paradigm

Source: *The Lateralizer* [6]

50 unseen exemplars, presented randomly, without any feedback on correctness. For Experiment (and hypothesis) 2, subjects went through a "simultaneous learning and testing phase" on an entirely new set of two categories. All 80 stimuli were (randomly) presented, with feedback, and the subjects were tested while they were learning the two categories.

5 Methods of Data Analysis

We describe the analysis for hypothesis 1. Before beginning the data analysis, it is important to prune the dataset based on certain metrics, so that only the most significant and good quality data points can get through. One metric to rule out is the confidence of the subject in his/her answers. However, subjects might feel confident for the wrong reasons, without knowing what the experiment is essentially trying to test. Thus, to rule out that, and to rule out random answering, we set an accuracy threshold on every subject. We expect a subject to be better than chance, and so we set 0.5 as the threshold. Roughly 50 subjects survive this pruning. We use two different approaches to model this data.

5.1 Bayesian Model Selection Approach

Let us define a random variable $X_s^i \in \{0, 1\}$ which represents the correctness of an answer on presentation of image i to a subject s. Let us also define 4 possible models/hypotheses in the set H, which the subject could be operating under: h_1 for LVF-Exemplar-RVF-Exemplar model, h_2 for LVF-Exemplar-RVF-Prototype model, h_3 for LVF-Prototype-RVF-Exemplar model and h_4

for LVF-Prototype-RVF-Prototype model. Say we know the probability of finding the data X, given the hypothesis h, that is, we know $P(X|h)$. (We drop the indices in notation for clarity.) Then by using Bayes' Rule, we can find out the probability of hypotheses as [7]:

$$P(h|X) = \frac{P(X|h) \cdot P(h)}{\sum_{h \in H} P(X|h) \cdot P(h)} \tag{1}$$

Let us assume that every hypothesis is equally likely. Then, we can reduce equation 1 to:

$$P(h|X) = \frac{P(X|h)}{\sum_{h \in H} P(X|h)} \tag{2}$$

Now, how do we estimate these probabilities? For this, we reintroduce the notion of similarity s_{ij} of images i and j, and the probability of classification, as described in Section 2.3. For an image i belonging to Category A, its probability of correct classification is given by:

$$P_e(A/i) = \frac{\left(\sum_{a \in A} s_{ia}\right)^\gamma}{\left(\sum_{a \in A} s_{ia}\right)^\gamma + \left(\sum_{b \in B} s_{ib}\right)^\gamma} \qquad \text{(for the exemplar model)} \tag{3}$$

$$P_p(A/i) = \frac{s_{i\hat{a}}^\gamma}{s_{i\hat{a}}^\gamma + s_{i\hat{b}}^\gamma} \qquad \text{(for the prototype model)} \tag{4}$$

We can similarly define the probability of correct classification if $i \in B$. Now, depending on the hypothesis under study, whose probability we are trying to find out given the data stream, and depending on which side the image stimulus was presented, we use either equation 3 or 4. For example, say the hypothesis under consideration is h_2 (LVF-Exemplar-RVF-Prototype). Now an image k is presented on LVF as a stimulus, and $k \in B$, then according to h_2, we use equation 3 to obtain:

$$P(X_s^i = 1|h) = P_e(B/i) \qquad \text{(used if subject answers correctly)}$$

$$P(X_s^i = 0|h) = 1 - P_e(B/i) \qquad \text{(used if subject answers incorrectly)}$$

Now, since we present a sequence of images $X = \{X_1, X_2, \cdots X_t\}$ in the testing phase, assuming that the images are independent of one another, we can write:

$$P(X_s|h) = \prod_{i=1}^{t} P(X_s^i|h) \tag{5}$$

We are now ready to use equation 2 on our data stream, to find out $P(h|X)$ for all h (see Algorithm 2 for details). Then, the hypothesis with the highest conditional probability is selected as the operating model for that subject. However, a very critical consideration in our analysis is what the similarity metric s_{ij} should be. We consider the following two metrics separately, and enclose results on both in Section 6.

Algorithm 2 Bayesian Model Selection Algorithm

Input: Training Image Set I_s^{train}, Data stream of subject on the testing set $X_s = \{X_s^1, \cdots, X_s^m\}$, Current Hypothesis h, Best-fit Tightness γ

$\{A, B\} \leftarrow findExemplars(I_s^{train})$

$\{\hat{a}, \hat{b}\} \leftarrow findPrototypes(I_s^{train})$

$P_s \leftarrow 0$

for $i = 1$ to m **do**

  $P_s^i \leftarrow 0$

  if $X_s^i.leftOrRight == left$ **then** ▷ Check if image was projected onto LVF or RVF

    if $h.LVF == exemplar$ **then** ▷ Check for the model on this VF for the given h

      $P_s^i \leftarrow \frac{\left(\sum_{a \in A} s_{ia}\right)^\gamma}{\left(\sum_{a \in A} s_{ia}\right)^\gamma + \left(\sum_{b \in B} s_{ib}\right)^\gamma}$ ▷ Probability of being correctly categorised in A

    else

      $P_s^i \leftarrow \frac{s_{i\hat{a}}^\gamma}{s_{i\hat{a}}^\gamma + s_{i\hat{b}}^\gamma}$

  else

    if $h.RVF == exemplar$ **then**

      $P_s^i \leftarrow \frac{\left(\sum_{a \in A} s_{ia}\right)^\gamma}{\left(\sum_{a \in A} s_{ia}\right)^\gamma + \left(\sum_{b \in B} s_{ib}\right)^\gamma}$

    else

      $P_s^i \leftarrow \frac{s_{i\hat{a}}^\gamma}{s_{i\hat{a}}^\gamma + s_{i\hat{b}}^\gamma}$

  if $i \in B$ **then** ▷ If the image belonged to B, we correct for it

    $P_s^i \leftarrow 1 - P_s^i$

  if $X_s^i.correct == false$ **then** ▷ If the image was wrongly categorised, we correct for it

    $P_s^i \leftarrow 1 - P_s^i$

  $P_s \leftarrow P_s + log(P_s^i)$ ▷ To prevent a loss of precision, sum in log-form instead of product

$P_s \leftarrow e^{P_s}$ ▷ Restore to $P(X_s|h)$

return P_s

5.1.1 Exponential Euclidean Distance Based Similarity

Akin to usual methods in literature, a popular way of measuring psychological distance is a weighted Manhattan distance, as described in section 2.3. However, to eliminate the need to define "weights", we use PCA to reduce data to just 8 orthogonal dimensions. We can now take the distance to be simple Euclidean distance in this vector space. Thus for (representative) images u and v, we get the distance as

$$d_{uv} = \sqrt{\sum_{i=1}^{8}(u_i - v_i)^2}$$

Now, we convert distance to similarity by taking a negatively raised exponent of it as

$$s_{uv} = e^{-cd_{uv}}$$

where c is a model parameter. Note that small ds give similarities lose to 1 and large ds give similarities close to 0. Also, the value of c is very sensitive, since it decides the weight of similarity of the image with a single exemplar. Thus, making a large value of c biased towards an exemplar model. Thus, how do we select for c? We assume it to be a parameter specific to the kind of model being used. Thus, to find the best c_i for our hypothesis h_i, we simply estimate the following:

$$\hat{c_i} = argmax_{c_i \in dom(c)} P(X|h = h_i, c = c_i)$$

That is, we find the parameter which best fits our hypothesis to the data. We can do this by treating parameter estimation as an unconstrained optimisation problem, with the objective function of maximising the conditional $P(X|h = h_i, c = c_i)$. For our experiments, we use the Nelder-Mead Algorithm to do this: seed c at 0.001, ensure the constraint that $c > 0$, and the solution does converge to a (local) optima.

Finally, using that $\hat{c}$, we find out the probabilities.

$$P(h = h_i|X, \hat{c_i}) = \frac{P(X|h = h_1, \hat{c_1})}{\sum_{i=1,4} P(X|h = h_i, \hat{c_i})}$$

5.1.2 Cosine Similarity

In the similarity metric above, there was a need to introduce a model parameter, which can be the key in deciding which model is eventually selected. Therefore, one way to circumvent this issue is to choose a parameterless metric, such as cosine similarity. It is a metric for inner product spaces, which is essentially a measure of the angle between two vectors. Thus, length of the vector is immaterial here. It is measured as

$$s_{uv} = \frac{u \cdot v}{|u||v|}$$

Thus for vectors with a very small angle, similarity is close to 1, and for those almost orthogonal, similarity is close to 0. Since this metric is parameterless, we would like to prefer using this over distance-based similarity, to reduce the complexity of our model.

5.1.3 Note on Few Engineering Concerns

Some key points to note while applying the Bayesian Cognition model:

- For a subject s whose training representation set I_s^{train} has 30 images, prototypes of both categories are estimated as the mean of the training exemplars of each category. That is,

$$\hat{a} = \frac{\sum_{i \in I_s^{train}} i\{cat(i) == A\}}{\sum_{i \in I_s^{train}} \{cat(i) == A\}}$$

  where $\{cat(i) == A\}$ is an indicator function which is 1 if $i \in A$, else 0.

- In both equations 3 and 4, sets A and B are a split of only the training set I_s^{train}, on the grounds of whichever category a training image belongs to. This is because a subject has learnt categories A and B based on feedback on the training set, and would thus compare new stimuli to these training exemplars only.

- There is a parameter γ which is inherent to our model. Similar to how we use an unconstrained optimisation in Section 5.1.2, we do the same to find a γ which best fits our data for the hypothesis under consideration. Hence, when using the exponential distance-based metric, we run a two-variable optimisation over the space of $\{c, \gamma\} \in \mathbb{R}^2$, while for the cosine similarity metric, we run a one-variable optimisation over the space of $\gamma \in \mathbb{R}$.

- Every image representation vector is normalised such that the entire imageset has zero mean and unit variance. This is a key preprocessing step for any data modelling problem.

5.2 Average Probability Distribution Approach

Let us define a random variable $Y \in [0, 1]$, which corresponds to the probability of correctness, over a space of subjects. Thus, we try to model the probability of the probability of correctness. Using equations 3 and 4, we can find the average/mean probability of correctness for a subject s given by

$$Y_s^h = \frac{\sum_{i \in I_s^{test}} P(X_s^i | h)}{|I_s^{test}|} \qquad \text{(where } I_s^{test} \text{ is the testing image set)} \tag{6}$$

Once we obtain this for all subjects, we can essentially form a histogram of all Y_s^h, which is nothing but the probability distribution of probability of correctness $P_h(Y)$ as per hypothesis h. Besides these 4 distributions, we also form the ground truth distribution, that is a histogram of actual average correctness on the test image set: call this probability distribution $P_0(Y)$. Now, it is intuitive to realise that among the four distributions $P_i(Y)$, whichever is the "closest" to the ground truth distribution must be the "best" distribution.

We use here the simplest notion of comparison between two probability distributions, that of Kullback-Leibler (KL) Divergence, popular in the field of information theory. For two probability distributions $P(i)$ and $Q(i)$ over some random variable i, KL divergence is given by:

$$D_{KL}(P||Q) = \sum_i P(i) log\left(\frac{P(i)}{Q(i)}\right) \tag{7}$$

Larger is the value of D_{KL}, lesser is the similarity between the two distributions. Note that D_{KL} is always non-negative. The hypothesis with lowest KL-Divergence is selected as the most "popular" operating model for subjects under consideration. Note that to find D_{KL}, we need to discretise the probability distributions at an appropriate binning size, and the divergence values can be sensitive to the way binning is done: too small a bin size can make the distribution sparse and too large a bin size can distort the characteristic shape of the distribution. Eventually, we settle for distributions discretised to 10 bins.

6 Results and Discussion

We have two choices of similarity metrics, and two methods of model selection. Therefore we enclose results for all four combinations in Tables 1, 2, 3 and 4. Note that for the Bayesian method, we apply an extra threshold called the hypothesis probability threshold, since we would only like to count those hypotheses which are significantly fit more than other hypotheses. We keep this $HypoThresh$ at 0.5 (twice that of a chance significance of 0.25). We then plot probability of hypotheses versus subject accuracy, thus making points on the upper side of these plots more significant (see Figure 4). For the average distribution method, we plot these distributions to see which one fits the actual distribution better (see Figure 5).

Distance-based Similarity: When we look at the Bayesian approach, this metric leads us to believe that most subjects populate h_3 = LVF-Prototype-RVF-Exemplar, a form of lateralisation which is an exact inversion of the expected result that our brains follow h_2. However, when we use the average distribution approach, h_1 = LVF-Exemplar-RVF-Exemplar comes across as the best fit model. Which of the two approaches should be believed, then? The former method is more principled, and could be trusted over the latter. However, more importantly, this uncovers a key issue with a distance-based similarity metric, one that is used often in literature[5]. The notion of distance puts too much importance on the "weights" given to every dimension, which overcomplicates our analysis. (In fact, we have also had to define a separate parameter c, which adds another layer of complexity.) This brings us to the results of the second metric.

Cosine Similarity: This metric looks at similarity by estimating angles, and not distances, between vectors. If this notion of similarity holds for our purpose (and we assume for it to work here, in our inner product space), then it can be used as a metric. Since

angles don't depend on the "weight" of a dimension, this simplifies our analysis as well. Additionally, this metric is parameterless. Looking at results for cosine similarity is more encouraging. We obtain a consistent ranking of our hypotheses from both Bayesian and average distribution approaches: h_1 = LVF-Exemplar-RVF-Exemplar is a highly preferred model, followed by h_2 = LVF-Exemplar-RVF-Prototype, which is the hypothesised model of category learning and inference, as given in Hypothesis 1.

Hypothesis	**Frequency**	**Rank**
LVF-Exemplar-RVF-Exemplar	4	2
LVF-Exemplar-RVF-Prototype	4	2
LVF-Prototype-RVF-Exemplar	10	1
LVF-Prototype-RVF-Prototype	2	4

Table 1: Frequency of the most probable hypothesis, after applying $SubjectThresh = 0.5$ and $HypoThresh = 0.5$, using the distance-based similarity metric

(Only 20 subjects survive $HypoThresh$ and show a "significant" preference for their model)

Hypothesis	**KL Divergence**	**Rank**
LVF-Exemplar-RVF-Exemplar	0.644	1
LVF-Exemplar-RVF-Prototype	0.944	4
LVF-Prototype-RVF-Exemplar	0.822	3
LVF-Prototype-RVF-Prototype	0.782	2

Table 2: KL Divergence of hypothesis probability distributions to actual probability distributions, after applying $SubjectThresh = 0.5$, using the distance-based similarity metric

Hypothesis	**Frequency**	**Rank**
LVF-Exemplar-RVF-Exemplar	7	1
LVF-Exemplar-RVF-Prototype	3	2
LVF-Prototype-RVF-Exemplar	0	3
LVF-Prototype-RVF-Prototype	0	3

Table 3: Frequency of the most probable hypothesis, after applying $SubjectThresh = 0.5$ and $HypoThresh = 0.5$, using the cosine similarity metric

(Only 10 subjects survive $HypoThresh$ and show a "significant" preference for their model)

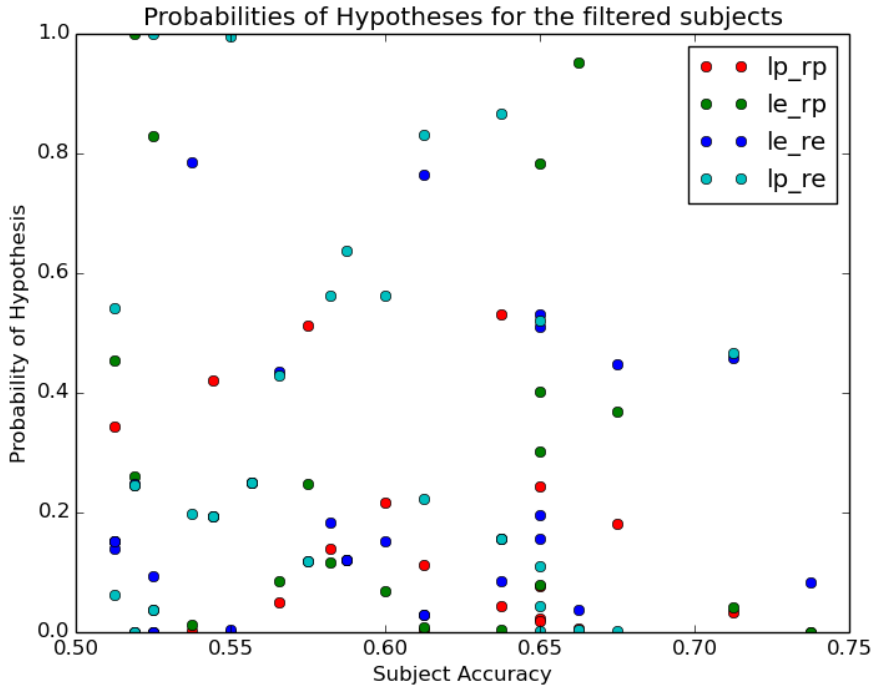
This interesting observation leads us to believe that at least for this experimental setting, category learning and inference is **not lateralised**, and the exemplar model is used when the image is presented on either side of the screen. Either this complete disproves the accepted

Hypothesis	**KL Divergence**	**Rank**
LVF-Exemplar-RVF-Exemplar	0.646	1
LVF-Exemplar-RVF-Prototype	1.092	2
LVF-Prototype-RVF-Exemplar	1.371	3
LVF-Prototype-RVF-Prototype	1.309	4

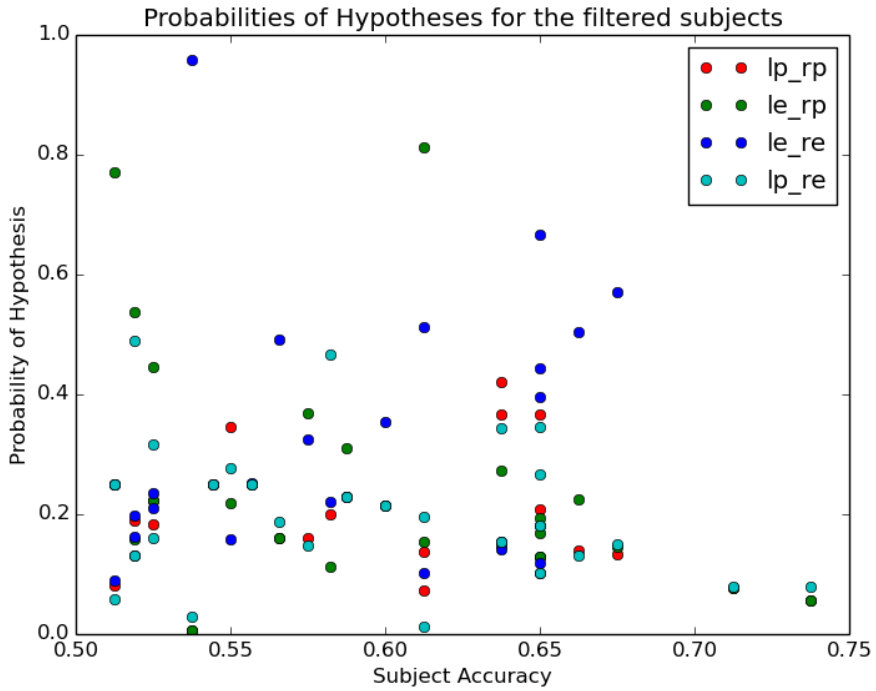
Table 4: KL Divergence of hypothesis probability distributions to actual probability distributions, after applying $SubjectThresh = 0.5$, using the cosine similarity metric

hypothesis, or, this could hint into the nature of lateralisation. Since the experiment was of a short duration, consisting of a small set of training and testing exemplars, perhaps this was not sufficient for a subject to prototype the two categories A and B. Indeed, a few subjects do show lateralisation of category learning, and they operate under the hypothesised "left-side is prototypical while right-side is exemplary" conditions. This could mean that the degree of lateralisation depends on the amount of learning which has taken place. That initially, due to few examples, the exemplar model is followed, and slowly the left hemisphere picks up on prototyping and leaves the right hemisphere alone in an exemplar paradigm. A way to further this analysis is to look at Hypothesis 2. For this, assuming that the selected hypothesis for a subject is correct, we simply plot the reaction times of stimuli presented to either of the hemispheres/VFs. Now, assuming that a preferred learning model accords a benefit in cost by being more efficient, and thus quicker in response, this model would have a low reaction time. Let us look at some of these plots, as shown in Figures 6 and 7. We enclose a few select subjects' plots here, who have cleared the filter thresholds. (These plots have been low-pass filtered for visual acuity. Kindly ignore any periodicities induced owing to this filtering.) Subjects under the same (exemplar) model in both hemispheres show very similar reaction time trends. Whereas we observe reaction times to be generally higher for the prototype conditioned hemisphere. For some, we even observe a crossing-over of the two trends: the hypothesised point of a **"switch"** between exemplar and prototype controls at some epoch of learning. These observations strengthen our claims about both Hypotheses 1 and 2. **That there is indeed a lateralisation of models of category learning and inference, the degree of which depends on the current epoch of learning.**

In conclusion, the primary aim of this project was to provide a mathematically nuanced and principled method of testing hypotheses in cognitive psychology. In that effort, some key contributions of this project were to better define notions of psychological distance, and probabilistic hypothesis testing, for undertaking simple brain lateralisation studies.

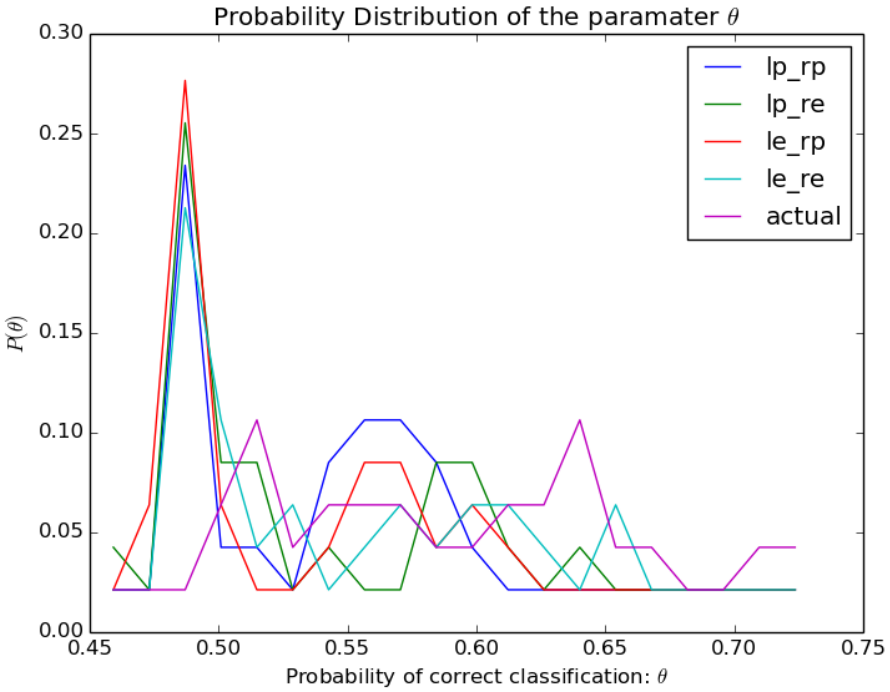


(a) Exponential Euclidean Distance Based Similarity

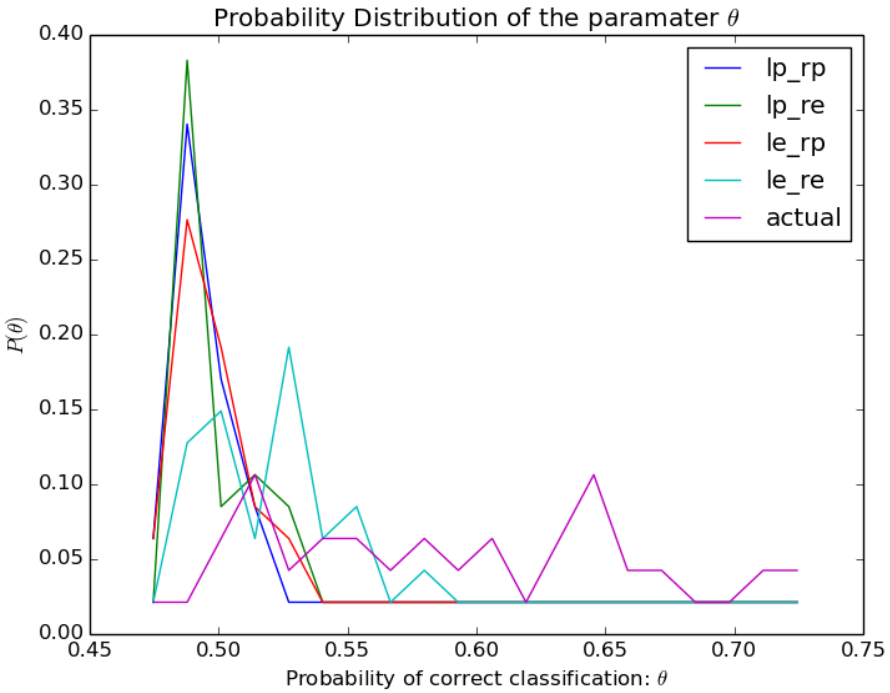


(b) Cosine Similarity

Figure 4: **Bayesian Model Selection:** Probability of various hypotheses against subject accuracy, for the two similarity metrics

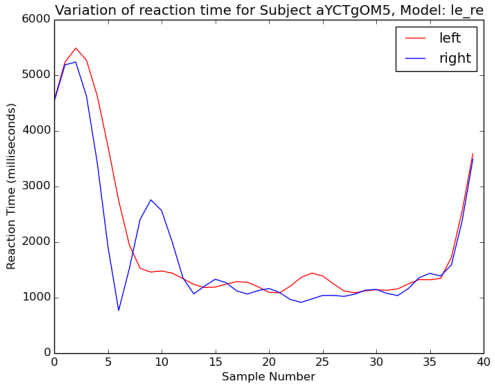


(a) Exponential Euclidean Distance Based Similarity



(b) Cosine Similarity

Figure 5: **Average Probability Distribution:** Probability distribution of probability of correctness

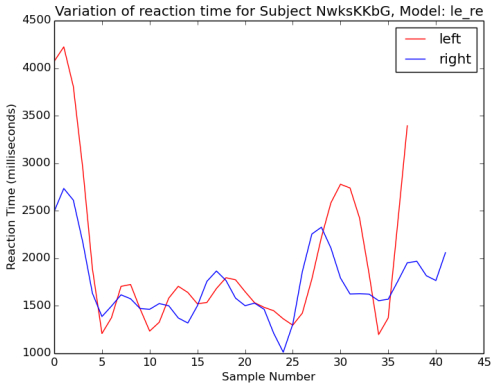


(a) LVF-Exemplar-RVF-Exemplar

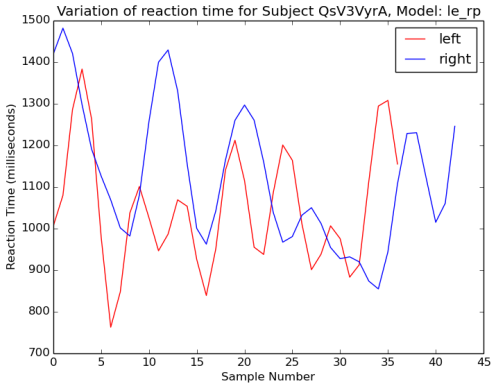
(b) LVF-Exemplar-RVF-Prototype

(c) LVF-Prototype-RVF-Exemplar

Figure 6: **Hypothesis 2:** Reaction times of both visual fields for few select subjects who got through exponential Euclidean distance-based similarity model



(a) LVF-Exemplar-RVF-Exemplar



(b) LVF-Exemplar-RVF-Prototype

Figure 7: **Hypothesis 2:** Reaction times of both visual fields for few select subjects who got through cosine similarity model

Acknowledgements: I would like to thank Dr. Varsha Singh for supervising this project and proposing interesting directions of research (and for letting students of her other course be our subjects), Dharti Tiwari for her massive help in organising the experiments at a large scale, and Anshul Bawa for immediate and helpful advice on way more than one occasion.

References

[1] *Cognitive Psychology and Cognitive Neuroscience/Knowledge Representation and Hemispheric Specialisation*, Wikibooks, linked here

[2] Medin and Schaffer, *Context Theory of Classification Learning*, Psychological Review, 1978, Vol. 85, No. 3, 207-238

[3] Gazzaniga, Ivry and Mangun, *Cognitive Neuroscience: The Biology of the Mind*, W.W. Norton, 2002. 2nd Edition 2000

[4] Homa, Sterling and Trepel, *Limitations of Exemplar-Based Generalization and the Abstraction of Categorical Information*, Journal of Experimental Psychology: Human Learning and Memory, 1981. Vol. 7. No. 6, 418-439

[5] Mack, Preston and Love, *Decoding the Brain's Algorithm for Categorization from Its Neural Implementation*, Current Biology 23, 2023–2027, October 21, 2013, linked here

[6] Motz, James and Busey, *The Lateralizer: a tool for students to explore the divided brain*, Advances in Physiology Education, 36(3), 220-225, 2012

[7] Griffiths, Kemp and Tenenbaum, *Bayesian models of cognition*, 2008