Obtaining psychological embeddings through joint kernel and metric learning

Brett D. Roads Michael C. Mozer

January 25, 2019

B. D. Roads
Department of Computer Science
University of Colorado Boulder
Boulder, CO 80309-0430
brett.roads@colorado.edu
Present address:
Department Experimental Psychology
University College London
26 Bedford Way
London
WC1H 0AP
United Kingdom
b.roads@ucl.ac.uk
tel: +1 720-205-2971

M. C. Mozer
Department of Computer Science
University of Colorado Boulder
Boulder, CO 80309-0430
mozer@colorado.edu
Present address:
Google Brain
1600 Amphitheater Parkway
Mountain View, CA 94304
mcmozer@google.com

Abstract

Psychological embeddings provide a powerful formalism for characterizing human-perceived similarity among a set of stimuli. Obtaining high quality embeddings can be costly due to algorithm design, software deployment, and participant compensation. This work aims to advance state-of-the-art embedding techniques and provide a comprehensive software package that makes obtaining a high quality psychological embeddings both easy and relatively efficient. Contributions are made on four fronts. First, the embedding procedure allows multiple trial configurations to be used for collecting similarity judgments from participants. For example, trials can be configured to collect triplet comparisons or to sort items into groups. Second, a likelihood model is provided for three classes of similarity kernels allowing users to easily infer the parameters of their preferred model using gradient descent. Third, an active selection algorithm is provided that makes data collection more efficient by selecting comparisons that provide the strongest constraints on the embedding. Fourth, the likelihood model allows the specification of group-specific attention weight parameters. A series of experiments are included to highlight each of these contributions and their impact on converging to a high-quality embedding. Collectively, these incremental improvements provide a powerful and complete set of tools for inferring psychological embeddings. The relevant tools are available as the Python package *PsiZ*, which can be cloned from GitHub (https://github.com/roads/psiz).

1 Introduction

In many interactive software systems, it is essential to predict how individuals will respond to a given visual task. Decision support applications anticipate and adjust for novice perception in order to help novice users arrive at an expert-like categorization decision (Fang & Geman, 2005; Ferecatu & Geman, 2009; Roads & Mozer, 2017). Human-in-the-loop computer vision algorithms utilize a model of human similarity to improve categorization performance (Wah et al., 2014). Visual category-training applications use cognitive models to predict learning outcomes (e.g., Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1986).

At the core of these applications is the notion of psychological similarity. The rich psychological literature on human and animal generalization (Shepard, 1987; Tenenbaum, 1999) explores the conditions under which responses associated with one stimulus transfer to another, or properties associated with one stimulus are ascribed to another. The more similar stimuli are, the more likely generalization is to occur. Similarity is based not on external properties of the stimuli, but rather on an individual's internal representation. We refer to this internal representation as a *psychological embedding*. The primary objective of this work is to jointly infer both a multi-dimensional feature representation along with a corresponding similarity function. This work aims to provide a principled integration of state of the art methods with existing psychological theory.

The work presented in this paper has strong ties to existing embedding techniques. Popular metric multidimensional scaling (MDS) approaches assume simple and restrictive similarity functions (e.g., Gower, 1966; Torgerson, 1958). Non-metric MDS algorithms are nearly atheoretical, assuming only a monotonically decaying function (e.g., Kruskal, 1968a, 1968b). The approach presented in this work is situated in between these two extremes, using more complex similarity functions that are well-motivated by psychological theory. While existing approaches use more complex similarity functions (e.g., van der Maaten & Weinberger, 2012), these approaches typically assume a fixed form for the similarity function during inference. In contrast, this work provides a method for jointly inferring the parameters of similarity functions as well as the corresponding multi-dimensional feature representation.

The primary purpose of this work is to provide a unified set of state-of-the-art tools for individuals interested in inferring psychological embeddings. This work has four facets that can be adjusted to suit the needs of the user. First, observations used for inference can be collected using a variety of different trial configurations. Second, two different classes of models can be used for performing inference, one derived from psychological research and the other widely used in machine learning. Additional classes can easily be implemented by the user. Third, the embedding algorithm can be used to infer group-specific attention weights in the same spirit as INDSCAL (Carroll & Chang, 1970). Lastly, high-quality embeddings can be constructed with less data by using an active selection algorithm that intelligently determines which observations should be collected next. Each of these facets is described in turn, followed by experiments highlighting potential benefits available to the researcher. The relevant tools are available as the Python package *PsiZ*, which can be cloned from https://github.com/roads/psiz. The relevant code to run all the experiments can be cloned from https://github.com/roads/psiz-app.

2 Trial configurations

Inference is performed by collecting similarity judgments where subjects compare a *query* stimulus to at least two *reference* stimuli. The ith trial consists of a query stimulus and a set of reference stimuli. On any trial, the set of reference images $\mathcal{R}_i$ can contain 2-8 images (i.e., $|\mathcal{R}_i| \in [2, 8]$). Images in a set are referenced using a *stimulus index*. Given a trial, subjects select a predefined number of reference images that they consider most similar to the query. The set of selected reference images $\mathcal{S}_i$, may be 1 to $|\mathcal{R}_i| - 1$ images depending on the trial configuration. Again, the set $\mathcal{S}_i$ contains the stimuli indices of the selected reference stimuli. In the simplest case, where $|\mathcal{R}_i| = 2$ and $|\mathcal{S}_i| = 1$, subjects provide triplet similarity judgments (Figure 1a). In a more complicated scenario, subjects can partition the set of references into a similar and dissimilar group (Figure 1b). Alternatively, subjects may be asked to rank their selections (Figure 1c).

Each judged trial constitutes one observation. Information for the ith judged trial $\mathcal{D}_i$, is a vector formatted such that $\mathcal{D}_i = (q, \mathbf{s}, \mathbf{u})$, where q indicates the index of the query stimulus, $\mathbf{s}$ is a row vector comprised of the selected reference stimulus indices, and $\mathbf{u}$ is a row vector comprised of the unselected reference stimulus indices. Depending on the trial configuration, the length of $\mathcal{D}_i$ will vary. For example, if $|\mathcal{R}| = 2$ and $|\mathcal{S}| = 1$, $\mathcal{D}_i = (q, a, b)$, where a indicates the index of the selected reference stimulus and b indicates the index of the unselected reference stimulus. If $|\mathcal{R}| = 8$, $|\mathcal{S}| = 2$ and subjects make ranked selections, then

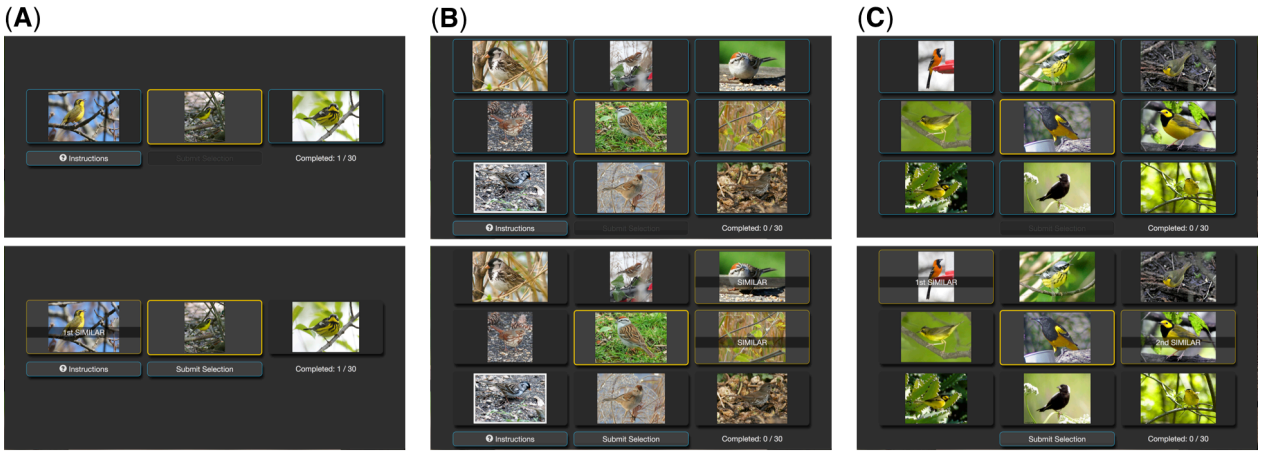

Figure 1: Sample displays shown to subjects. The center image is the query stimulus while the surrounding images are the reference stimuli. (**A**) Given two reference images, subjects select the one reference image that is most similar to the query. (**B**) Given eight reference images, subjects select the two reference images that are most similar to the query. (**C**) Given eight reference images, subjects select two reference images in the order of their similarity.

$\mathcal{D}_i = (q,\ a,\ b,\ c,\ d,\ e,\ f,\ g,\ h)$. Now a indicates the index of the most similar reference and b indicates the index of the second most similar reference. Indices $c - h$ are the remaining unselected references.

Each trial configuration specifies a certain number of implied triplet constraints. Holding all other factors constant, providing more triplet constraints improves the quality of the inferred solution. In the simplest case ($|\mathcal{R}| = 2$, $|\mathcal{S}| = 1$), each trial provides one triplet constraint of the form $q : a > b$, where a is the reference stimulus that was selected as more similar to the query q. More generally, for unordered selections, each display yields $|\mathcal{S}|\,(|\mathcal{R}| - |\mathcal{S}|)$ triplet constraints. For ordered selections, each trial yields $|\mathcal{S}|\,(|\mathcal{R}| - |\mathcal{S}|) + \binom{|\mathcal{S}|}{2}$ triplet constraints. The best trial configuration will depend in part upon the time needed to complete a single trial and the level of noise in subject responses (e.g., Wilber, Kwak, & Belongie, 2014).

3 Likelihood model

The primary objective is to jointly infer a psychological embedding, which constitutes a multi-dimensional feature representation and a similarity function. To improve conceptual clarity, the free parameters representing and the feature representation ($\mathbf{Z}$) and the parameters controlling the similarity function ($\boldsymbol{\theta}$) are written separately. Given a set of observations $\mathcal{D}$, the likelihood of the of the data given the model parameters is:

$$\mathcal{L} = \prod_i p\left(\mathcal{D}_i | \mathbf{Z}, \boldsymbol{\theta}\right) \tag{1}$$

It should be noted that the likelihood easily permits different display configurations to be used together.

In order to infer a psychological embedding from our observations, we make two key assumptions. First, we assume that the psychological embedding is composed of points in a D-dimensional space. Second, we assume that the similarity function belongs to one of two classes. Regardless of the similarity function used, we must also formalize the selection model that specifies how perceived similarity is converted into behavior (i.e., selecting reference images). After describing the selection model, we turn to the two different classes of similarity functions.

3.1 Selection model

Given a similarity function, the likelihood of subject selections are modeled in the same spirit as Luce's ratio of strengths formulation (Luce, 1959). Given two embedding points $\mathbf{z}_1$ and $\mathbf{z}_2$, the similarity function $s\left(\mathbf{z}_1, \mathbf{z}_2\right)$ returns a value, where 0 indicates that the two points are maximally dissimilar. The exact form of the similarity function is discussed shortly. The probability of selecting a given reference is proportional to

the similarity between the query and that reference. Depending on the trial configuration, the exact form of the likelihood will vary. For example, when subjects make only one selection ($|\mathcal{R}_i| \in [2, 8]$, $|\mathcal{S}_i| = 1$),

$$p\left(\mathcal{D}_i|\mathbf{Z}, \boldsymbol{\theta}\right) = \frac{s\left(\mathbf{z}_q, \mathbf{z}_a\right)}{\sum_{r \in \mathcal{R}_i} s\left(\mathbf{z}_q, \mathbf{z}_r\right)}. \tag{2}$$

In a more complicated trial configuration, such as $|\mathcal{R}_i| \in [3, 8]$ and $|\mathcal{S}_i| = 2$ with unranked selections,

$$p\left(\mathcal{D}_i|\mathbf{Z}, \boldsymbol{\theta}\right) = \frac{s\left(\mathbf{z}_q, \mathbf{z}_a\right)}{\sum_{r \in \mathcal{R}_i} s\left(\mathbf{z}_q, \mathbf{z}_r\right)} \frac{s\left(\mathbf{z}_q, \mathbf{z}_b\right)}{\sum_{r \in \mathcal{R}_{i \neg a}} s\left(\mathbf{z}_q, \mathbf{z}_r\right)} + \frac{s\left(\mathbf{z}_q, \mathbf{z}_b\right)}{\sum_{r \in \mathcal{R}_i} s\left(\mathbf{z}_q, \mathbf{z}_r\right)} \frac{s\left(\mathbf{z}_q, \mathbf{z}_a\right)}{\sum_{r \in \mathcal{R}_{i \neg b}} s\left(\mathbf{z}_q, \mathbf{z}_r\right)}, \tag{3}$$

and with ranked selections,

$$p\left(\mathcal{D}_i|\mathbf{Z}, \boldsymbol{\theta}\right) = \frac{s\left(\mathbf{z}_q, \mathbf{z}_a\right)}{\sum_{r \in \mathcal{R}_i} s\left(\mathbf{z}_q, \mathbf{z}_r\right)} \frac{s\left(\mathbf{z}_q, \mathbf{z}_b\right)}{\sum_{r \in \mathcal{R}_{i \neg a}} s\left(\mathbf{z}_q, \mathbf{z}_r\right)}. \tag{4}$$

3.2 Exponential-family kernel

Integrating various psychological models (e.g., Jones, Love, & Maddox, 2006; Jones, Maddox, & Love, 2006; Nosofsky, 1986; Shepard, 1987) into their most general form, we obtain:

$$s\left(\mathbf{x}, \mathbf{y}\right) = \exp\left(-\beta \left\|\mathbf{x} - \mathbf{y}\right\|_{\rho, \mathbf{w}}^{\tau}\right) + \gamma, \tag{5}$$

where β, τ, and γ control the gradient of generalization. The norm $\left\|\mathbf{x} - \mathbf{y}\right\|_{\rho, \mathbf{w}}$ denotes the weighted Minkowski distance:

$$\left\|\mathbf{x} - \mathbf{y}\right\|_{\rho, \mathbf{w}} = \left(\sum_{j=1}^{D} w_j \left|x_j - y_j\right|^{\rho}\right)^{\frac{1}{\rho}}, \tag{6}$$

where $w_j \geq 0$ and $\sum_{j=1}^{D} w_j = D$. The parameter ρ controls the type of distance (e.g., $\rho = 2$ results in Euclidean distance). The weights correspond to attention weights and allow the similarity function to capture differences in how individuals or groups attend to different dimensions in the psychological embedding. In the most general case, these weights are allowed to vary by individual or group. In most restrictive case, we assume a single population-level model and set all the weights to one. Note that the weights sum to D so that when all the weights are equal, i.e., $w_j = 1$, we recover the standard (unweighted) Minkowski distance. Since the most common parameter settings result in a Laplacian kernel ($\tau = 1$, $\rho = 2$, $\gamma = 0$) and Gaussian kernel ($\tau = 3$, $\rho = 2$, $\gamma = 0$), we refer to this class of similarity functions as the *exponential-family* kernel.

3.3 Student-t kernel

Although substantial psychological evidence supports the idea that individuals use an exponential similarity function, other similarity functions have been used with success. In machine learning, a popular similarity function is the Student-t kernel (van der Maaten & Weinberger, 2012):

$$s\left(\mathbf{x}, \mathbf{y}\right) = \left(1 + \frac{\left\|\mathbf{x} - \mathbf{y}\right\|_{2, \mathbf{w}}^{2}}{\alpha}\right)^{-\frac{\alpha+1}{2}}. \tag{7}$$

A primary advantage of the Student-t kernel is that it has a heavy tail. The heavy tail is advantageous during inference because it provides a signal to the inference algorithm to continue pushing similar points together and dissimilar points apart.

3.4 Heavy-tailed kernel

By itself, the Student-t kernel lacks the flexibility of the exponential kernel. By generalizing the Student-t kernel with additional free parameters, you obtain a heavy-tailed kernel with comparable flexibility to the exponential kernel:

$$s\left(\mathbf{x}, \mathbf{y}\right) = \left(\kappa + \left\|\mathbf{x} - \mathbf{y}\right\|_{\rho, \mathbf{w}}^{\tau}\right)^{-\alpha}. \tag{8}$$

4 Inference procedure

Tools for performing inference are provided by the PsiZ package. Given a set of observations $\mathcal{D}$, one selects a particular similarity kernel and inference is performed using gradient decent on the log-likelihood of the data given the model parameters:

$$\max_{\mathbf{Z},\boldsymbol{\theta}} \sum_i \log\left(p\left(\mathcal{D}_i|\mathbf{Z},\boldsymbol{\theta}\right)\right). \tag{9}$$

Internally, gradients are computed using the TensorFlow Python library (Abadi et al., 2015).

Before executing the inference procedure, one must first decide the dimensionality of the embedding. The PsiZ package includes a procedure that uses five-fold cross validation in order to estimate the appropriate dimensionality of the embedding. By tracking generalization performance on the validation set, the procedure continues to add dimensions until validation performance no longer improves.

5 Active selection

The final aspect of this work focuses on intelligently selecting trials to show participants. A simple strategy is to randomly select images for each trial, with the sole constraint that all images be unique. Although a reasonable approach, random displays have a drawback. Imagine for the moment that judged displays have been collected using all but one stimulus. Ideally, the next trial shown to a participant would include that unused stimulus. More generally, the embedding procedure will be less confident about the position of some images in an embedding. An active selection approach constructs trials that have the best chance of minimizing uncertainty associated with the embedding. Active selection proceeds via multiple iterations of generating trials and collecting the corresponding observations. This component is heavily inspired by previous active selection research (Tamuz, Liu, Belongie, Shamir, & Kalai, 2011), but has been generalized to handle arbitrary trial configuration and uses a different set of heuristics.

5.1 Formalizing uncertainty

The uncertainty of an embedding point's location is formally characterized using a posterior distribution,

$$p\left(\mathbf{z}_k|\mathcal{D},\mathbf{Z}_{\neg k},\boldsymbol{\theta}\right) \propto p\left(\mathcal{D}|\mathbf{z}_k,\mathbf{Z}_{\neg k},\boldsymbol{\theta}\right) p\left(\mathbf{z}_k|\mathbf{Z}_{\neg k},\boldsymbol{\theta}\right). \tag{10}$$

For simplicity, we assume that the prior distribution of the embedding points is characterized using a Gaussian distribution

$$p\left(\mathbf{z}_k|\mathbf{Z}_{\neg k}\boldsymbol{\theta}\right) \sim \mathcal{N}\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right). \tag{11}$$

Assuming a Gaussian prior mirrors the assumptions of the embedding algorithm. The likelihood is the same as previously described and predicts how participants select references given a particular query,

$$p\left(\mathcal{D}|\mathbf{z}_k,\mathbf{Z}_{\neg k},\boldsymbol{\theta}\right). \tag{12}$$

The posterior distribution is approximated by sampling from the joint distribution using Gibbs sampling. Since the posterior distribution has a Gaussian prior, elliptical slice sampling (Murray, Adams, & MacKay, 2010) can be used to sample points in a relatively efficient manner. In effect, the sampling procedure produces a set of points for each stimulus. If the distribution of points is tightly clustered, then there is relatively low uncertainty about the position of that stimulus. If the distribution is wide, then there is relatively high uncertainty about the location of the stimulus in the embedding. For clarity, the posterior samples are denoted using a three-dimensional tensor $\boldsymbol{\zeta}$ such that $\boldsymbol{\zeta}_k^{(s)}$ indicates the sth sample of the kth stimulus. Since each point exists in a d-dimensional space, the third dimension of $\boldsymbol{\zeta}$ corresponds to the dimensionality of the embedding. The matrix $\boldsymbol{\zeta}^{(s)}$ can be thought of as a sampling snapshot of the entire embedding where $\boldsymbol{\zeta}_k^{(s)}$ is the embedding point of the kth stimulus in that sampling snapshot.

5.2 Maximizing information gain

To maximize information gain, we need to compute the expected information gain for a candidate trial. Given a candidate trial $\mathbf{c}$, the expected information gain is equal to the mutual information,

$$I\left(\mathbf{Z};Y|\boldsymbol{\mathcal{D}},\mathbf{c}\right) = H\left(\mathbf{Z}|\boldsymbol{\mathcal{D}}\right) - H\left(\mathbf{Z}|\boldsymbol{\mathcal{D}},Y,\mathbf{c}\right), \tag{13}$$

where Y is a discrete random variable indicating all possible outcomes when the candidate trial is shown to a participant. For example, if the candidate trial displays two references and participants must select one reference, then there are two possible outcomes. The first term indicates the entropy associated with the current embedding. The second term indicates the expected entropy of the embedding if we collect an observation for the candidate trial. Since we would like to minimize entropy (i.e., uncertainty) associated with the embedding we are looking for a candidate trial such that $H\left(\mathbf{Z}|\boldsymbol{\mathcal{D}}\right) > H\left(\mathbf{Z}|\boldsymbol{\mathcal{D}}, Y, \mathbf{c}\right)$ and information gain is positive.

Since $\mathbf{Z}$ is a continuous variable, computing information gain appears non-trivial. Fortunately, the computation can be simplified by exploiting the identity of mutual information (i.e., $H(A) - H(A|B) = H(B) - H(B|A)$) and using our previously obtained samples taken from the posterior distribution in order to approximate the integrals. After all modifications and approximations, the equation for information gain becomes,

$$I\left(\mathbf{Z}; Y|\mathcal{D}, \mathbf{c}\right) = -\sum_{i=1}^{M} P\left(y_i|\mathcal{D}, \mathbf{c}\right) \log P\left(y_i|\mathcal{D}, \mathbf{c}\right) + \frac{1}{N}\sum_{s}^{N}\sum_{i}^{M} \log\left(p\left(y_i|\boldsymbol{\zeta}^{(s)}, \mathcal{D}\right)\right) p\left(y_i|\boldsymbol{\zeta}^{(s)}, \mathcal{D}\right), \quad (14)$$

where M indicates the number of possible outcomes associated with the candidate trial and N is the number of samples being used to approximate the integral.

5.3 Heuristic search procedure

For simple scenarios is is possible to evaluate all candidate trials in order find the trial that maximizes expected information gain. Unfortunately, for most scenarios, particularly those involving larger stimulus sets, exhaustive search becomes prohibitively expensive. As an alternative, we employ a two-stage heuristic search strategy. In the first stage a query stimulus is stochastically selected based on it's relative uncertainty. In the second stage, a candidate set of references is stochastically selected based on their similarity to the query stimulus. This process is repeated until the desired number of trials have been created. Ideally the embedding would updated to take into account observations for the new trial. In practice, multiple trials can be generated at once by limiting the number of times a particular stimulus can be used as a query.

In the first stage, relative uncertainty is determined by summing the Kullback–Leibler divergence between the stimulus of interest and all other stimuli. Intuitively, this prioritizes stimuli that exhibit high uncertainty in the embedding. However, no all uncertainty is equal. Highly uncertain stimuli that are far away from other stimuli are less crucial to pin down than highly uncertain stimuli that have close neighbors. The asymmetric nature of Kullback–Leibler divergence is also exploited in this heuristic. Given two stimuli, one that has high uncertainty and one that has low uncertainty, only the stimulus with high uncertainty should be given higher priority.

In the second stage, a set of candidate references are selected based on similarity using the current best estimate of the similarity function. In effect, this heuristic biases the reference set to include stimuli that are close neighbors of the query stimulus. If all the reference stimuli are excessively dissimilar from the query stimulus, the corresponding similarity judgments will not provide much information. When few observations have been collected, this heuristic behaves as if it were randomly selecting reference stimuli.

6 Experiments

6.1 Experiment 1: Kernel comparison

The following experiment compares the ability of three different similarity kernels to predict human similarity judgments. The exponential-family kernel is motivated by psychological theory, white the Student-t kernel and heavy-tailed kernel are largely motivated by common practice in machine learning.

6.1.1 Methods

Participants A population of 342 participants were recruited from Amazon Mechanical Turk and paid at a rate of approximately $6.00 per hour.

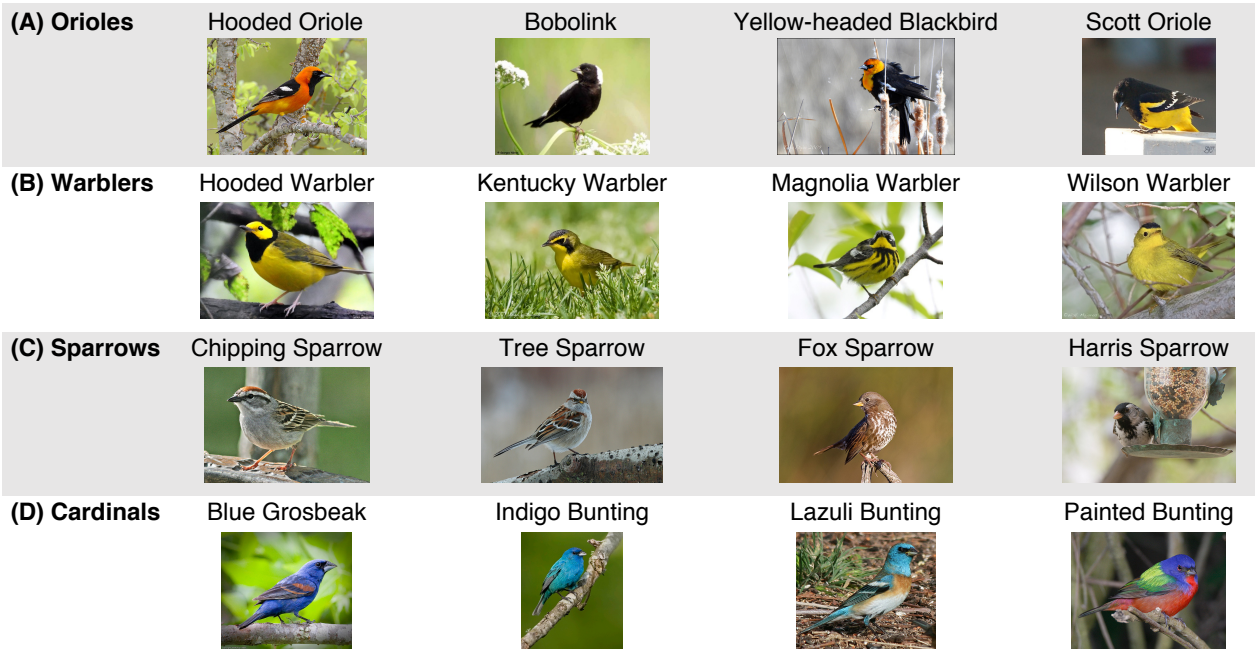

Figure 2: Example stimuli of the different bird species used in this work. Each row contains four similar bird species, each of which belongs to the same or similar taxonomic family. The images shown in this figure were drawn from the set of 208 images used in the experiments.

Materials A small dataset of 16 species of birds was selected from the CUB 200 image dataset (Wah, Branson, Welinder, Perona, & Belongie, 2011). Species were chosen such that there were four groups of birds composed of four similar looking species, roughly corresponding to four taxonomic families (Figure 2A-D). For each species, we selected 13 images, yielding a total of 208 unique images. Images were selected and cropped such that each image displayed a single bird, the bird was clearly visible, the image was of a good resolution, and no text was present.

Procedure Similarity judgments were collected during short 10 minute sessions via a web-based application. Each 10 minute session used one of two possible trial configurations. Participants either saw trials with two references and selected the most similar reference (2-choose-1) or saw eight references and selected two references in a ranked order (8-choose-2). During a 2-choose-1 session, participants saw between 60-120 trials. The number of displays varied in order to calibrate each session to be approximately 10 minutes. During a 8-choose-2 session, participants saw 30 trials. Participants were allowed to complete more than one 10 minute session. Collectively, participants judged 7,520 2-choose-1 trials and 8,772 8-choose-2 trials. All judged trials were combined to create a single dataset of observations ($\mathcal{D}$).

The collected similarity judgments were used in a ten-fold cross validation procedure in order to evaluate the capabilities of an exponential-family kernel, a heavy-tailed kernel, and a Students t kernel. Similarity judgments were partitioned into 10 roughly equal folds such that each fold had a the same proportion of 2-choose-1 and 8-choose-1 trials. For each fold, the dimensionality was estimated and then an embedding was inferred using nine folds as training data. The remaining fold was used as a validation set. For each fold, the validation loss (i.e., -log-likelihood) was recorded. Since individuals may perceive similarity differently and individuals themselves may not be consistent, we do not expect to infer embeddings with zero loss.

6.1.2 Results

The primary goal of the model comparison is to determine which model is best able to predict unseen human similarity judgments. The results are presented in Figure 3. Significance tests use a Bonferroni corrected alpha value of .05 (.017 corrected). Focusing on validation loss, pair-wise t-tests of the ten-fold cross-validation validation procedure reveal that the differences between the exponential-family kernel ($M = 2.93$, $SD = 0.09$), the heavy-tailed kernel ($M = 2.98$, $SD = 0.07$), and the Student-t kernel ($M = 2.92$, $SD = 0.06$)

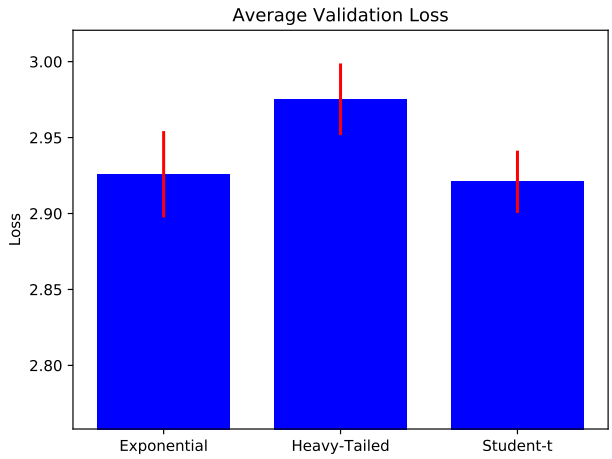

Figure 3: Model fitting results using a ten-fold cross validation procedure. The bars show the mean validation loss (i.e., -log-likelihood) for three different similarity kernels: exponential, heavy-tailed and Student-t. Error bars indicate standard error of the mean across the ten folds.

are all non-significant.

6.1.3 Discussion

The three kernels appear equally capable of predicting human similarity judgments. For domains that are similar to the set of birds used here, it is likely reasonable to use either kernel. One advantage of using the exponential-family kernel is that many computational models of human category learning also use an exponential-family kernel. By assuming an exponential-family kernel, the resulting psychological embedding could be integrated with a category learning model (e.g., Roads, Xu, Robinson, & Tanaka, 2018).

6.2 Experiment 2: Data collection strategies

Having compared different kernels, we turn to the issue of comparing different strategies for collecting observations. The goal is to determine the strategy that results in the highest-quality embedding at the lowest cost. Since the primary cost of collecting similarity judgments is paying participants for their time, we evaluate different collection strategies based on how many worker hours are required to reach a given quality level.

Data collection strategies are evaluated along two dimensions. First, the trial itself can take on many different configurations. Second, trials can be generated randomly or via active selection. The different collection strategies are evaluated using simulations of human similarity judgments.

6.2.1 Methods

Participants No new participants were recruited for this experiment. All similarity judgments collected for the previous experiment were re-used to infer a ground-truth model of human behavior.

Materials The experiment used the same set of 208 bird images as the previous experiment.

Procedure Three different collection strategies were evaluated using simulated human responses. The first collection strategy presented trials containing two references, where simulated participants selected one reference. The content of the trials was chosen randomly, subject to the constrain that a single image couldn't appear more than once on a trial (Random 2-choose-1). The second collection strategy presented trials containing eight references and required participants to select two references, in ranked order. The particular images for each trial were chosen randomly (Random 8-choose-2). The last strategy used an 8-choose-2 trial configuration, but selected the trial content using active selection (Active 8-choose-2).

Simulated responses were generated by treating a fitted psychological embedding as a generative model of human behavior (i.e., a virtual subject). Once a psychological embedding predicts the probability of all

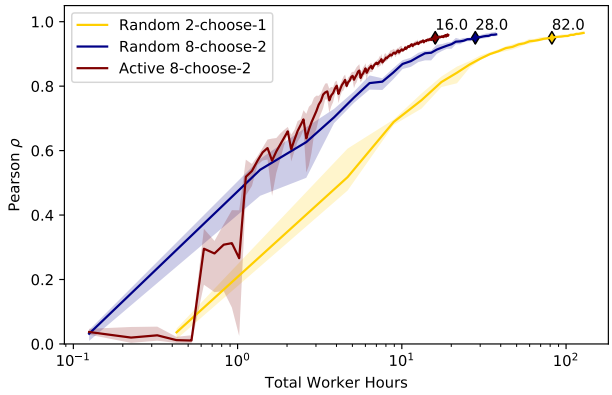

Figure 4: Simulated results of Experiment 2. Each line indicates a different collection strategy. Since cost is determined by total number of worker hours needed, the quality of the inferred embeddings is plotted with respect to worker hours. Each line indicates the mean between five independent simulation runs. The shaded regions indicate the maximum and minimum envelope across runs.

possible response outcomes for a particular trial (see Section 3.1), a specific response is generated by stochastically sampling from the possible outcomes. To ensure that the simulated responses mirror human behavior, an exponential-family psychological embedding was fitted to all human similarity judgments described in the previous experiment (7,520 2-choose-1 trials and 4,733 8-choose-2 trials). The fitted model served as a virtual subject and the ground-truth psychological embedding by which other models were evaluated.

Each collection strategy is used to generate trials, collect observations, and infer a strategy-specific psychological embedding. The quality of a strategy-specific embedding is determined by comparing its predictions to those of the ground-truth embedding. The critical predictions of a psychological embeddings can be summarized by generating a corresponding pair-wise similarity matrix S. The element s_{ij} indicates the similarity between the ith and jth stimulus. The predictions of a strategy-specific and ground-truth psychological embedding can be compared by computing the Pearson correlation coefficient between the respective similarity matrices. When computing the Pearson correlation, we only use the upper diagonal portion of the matrix less the diagonal elements, since the matrix is symmetric and the diagonal elements indicate self-similarity. If the strategy-specific embedding has successfully modeled the ground-truth embedding, the Pearson correlation will be high.

Each strategy-specific embedding was inferred using a different number of trials in order to map out how the number of trials affects the quality of the inferred embedding. Starting with an initial set of observations, additional observations were added in an incremental fashion. Each strategy-specific embedding was evaluated based on how many worker hours it took to reach a Pearson correlation of .95. Since there are two sources of stochasticity (trial generation and response simulation), five separate runs were conducted for each strategy. For each run, Random 2-choose-1, Random 8-choose-2, and Active 8-choose-2 were seeded with 500, 50, and 50 trials respectively. For all strategies, the seed trials had their content generated randomly. During active selection, 40 trials (each with a unique query image) were generated per round. In between every round, the posterior distribution of the embedding points was updated, while holding constant the parameters of the similarity function. Every fifth round the parameters of the similarity function were updated. For simplicity, all inference is performed assuming a dimensionality of three–matching the dimensionality of the ground-truth embedding.

6.2.2 Results

From the actual human data, it is clear that a 2-choose-1 display ($M = 4.73$, $SD = 11.18$) and 8-choose-2 ($M = 13.07$, $SD = 23.49$) display require different amounts of time to complete $t(12251) = -26.41$, $p < .001$. Number of trials are converted to worker hours based on the median human response time of the 2-choose-1 ($Mdn = 3.06$) and and 8-choose-2($Mdn = 8.98$) trials. Since the human response times include dramatic outliers, median response times provide an appropriate measure of central tendency.

The simulation results for three different collection strategies are presented in Figure 4. The simulation results show that Random 8-choose-2 (M=28.0) is more efficient than Random 2-choose-1 (M=82.0) in

reaching a Pearson correlation of .95. For the same embedding quality, only about 34% of the worker hours are necessary when using randomly selected 8-choose-2 versus 2-choose-1 trials. The results also reveal that Active 8-choose-2 (M=16.0) is more efficient than either Random strategy. When using an 8-choose-2 trial configuration, active selection requires about 57% of the worker hours compared to random selection.

6.2.3 Discussion

Using 8-choose-2 displays allows more data to be collected in a cost-effective manner, allowing a high-quality embedding to be inferred at nearly a third the cost. If the goal is to obtain a psychological embedding using the most effective trial configuration, the 8-choose-2 trial configuration is a good way to save money. These results replicate the findings of Wilber et al. (2014), except with a proper likelihood model.

Additional savings are achieved when switching to a strategy that uses active selection. The benefit of active selection appears to be greatest when the quality of the inferred embedding is starting to asymptote. Without much data, active selection behaves similarly to random selection. In effect, the active selection procedure is highly uncertain about all embedding points and the chosen displays provide the same amount of information at randomly generated displays. As more data is accumulated, active selection starts to focus on uncertain stimuli, allowing the inferred embedding to reach asymptote more quickly. According to these simulations, active selection provides an efficient and cost-effective way to obtain high-quality embeddings.

6.3 Experiment 3: Group-specific attention weights

In the final experiment, we demonstrate how the embedding procedure is capable of inferring group-specific attention weights in a similar spirit to INDSCAL (Carroll & Chang, 1970). More importantly, the experiment demonstrates how a shared embedding has the potential to reduce the cost of collecting data. The demonstration uses a shared set of fictitious stimuli and two simulated groups. These two groups can be likened to novices and experts. Inspired by novice and expert attention differences with musical notes (Shepard, 1982), we assume a scenario where novices pay attention to one set of feature dimensions, while experts attend to a complementary set of feature dimensions.

6.3.1 Methods

Participants No human participants were used in this experiment. All observations were simulated.

Materials A fictitious set of 100 stimuli was used in this experiment. The stimuli coordinates were drawn from a four-dimensional Gaussian distribution with zero mean and a spherical covariance matrix of 0.03.

Procedure Following the design of Experiment 2, we assume a known ground-truth psychological embedding. In contrast to the previous experiment, this embedding is not based on actual human behavior, but assumes that a set of stimuli are distributed in a four dimensional space. Furthermore, it is assumed that novices focus on the first two dimensions ($\mathbf{w} = [1.8, 1.8, .2, .2]$) while experts focus on the last two dimensions ($\mathbf{w} = [.2, .2, 1.8, 1.8]$) when judging similarity. Novice and expert responses are simulated by using the respective attention weights.

Multiple-group inference is examined using two conditions. In the Independent condition, randomly generated 8-choose-2 displays are presented to novices and experts and a separate psychological embedding is inferred for each group. In the Shared condition, a shared psychological embedding (with group-specific attention weights) is inferred and compared to the ground-truth embedding. The quality of an inferred embedding is evaluated in the same manner as Experiment 2, with a small twist. Since there are two groups, there are group-specific similarity matrices: a novice similarity matrix and an expert similarity matrix. The quality of a inferred embedding is determined by comparing a group-specific similarity matrix to the corresponding ground-truth similarity matrix.

Using five independent runs, we determine how many expert worker hours are necessary to reach psychological embeddings that correctly capture novice and expert behavior with at least a .95 Pearson correlation. Worker hours are estimated using the same conversion values used in Experiment 2.

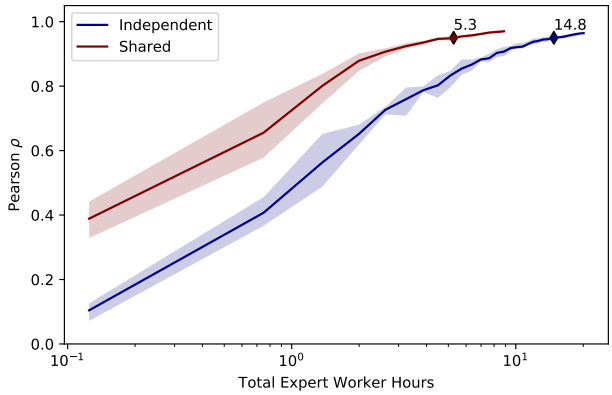

Figure 5: Experiment 3 simulation results. Each line indicates a different simulated scenario evaluating how many expert worker hours are necessary. Since cost is determined by total number of worker hours needed, the quality of the inferred embeddings is plotted with respect to worker hours. The blue line indicates the number of expert worker hours necessary to reach criterion when a independent psychological embedding is inferred for each group. The red line indicates the number of expert worker hours necessary to reach criterion when a shared psychological embedding is inferred. Each line indicates the mean between five independent simulation runs. The shaded regions indicate the maximum and minimum envelope across runs.

6.3.2 Results

The simulation expert-specific results for the two different conditions are presented in Figure 5. When inferring an expert-specific psychological embedding for the Independent condition, approximately 14.8 expert worker hours are required to reach criterion. Inferring a novice-specific psychological embedding requires 14.8 novice worker hours to reach criterion (not shown in the figure). When inferring a shared embedding, criterion can be met for both groups using 8.9 novice worker hours and 5.3 expert worker hours. From the perspective of total worker hours, the Independent condition requires 29.6 worker hours while the Shared condition requires 14.2 worker hours.

6.3.3 Discussion

There are two key findings from this experiment. The first interesting finding concerns the fact that fewer total worker hours are required when inferring a single psychological embedding with group-specific attention weights. One interpretation of total time difference is that inference struggles to determine the location of stimuli along the feature dimensions that receive little attention weight. By combining observations from subjects that have complementary attention weights, it becomes easier to determine the location of each stimulus in the psychological embedding.

The second interesting finding is that a shared psychological embedding can reach criterion for both experts and novices by using relative fewer expert hours. Since experts are typically paid more for their time (and expertise), reducing the required number of expert worker hours can substantially reduce the financial burden of collecting data. It is possible more extreme savings can be achieved by shifting more of the inference burden onto novice observations.

The above analysis assumed that novice and experts complete trials in the same amount of time. However, novices and experts may differ on their throughput. One possibility is that experts would be faster given their ability to make quick fine-grained judgments about their domain of expertise (Tanaka & Taylor, 1991).

7 Conclusion

Psychological embeddings are useful in many domains of research. Despite the substantial progress that has been made, a unified and coherent set of tools has been slow to emerge. This work presents the key aspects of a publicly available python package that makes it easy for researchers to infer their own psychological embeddings. In an effort to make the tools as useful as possible, the algorithms have been designed to handle a variety of trial configurations and handle inference of group-specific attention weights. In addition, the

package includes an active selection routine to help researchers get the most out of their budget. While these facets were discussed in the context of visual similarity, the software package can work with similarity judgment based on other modalities.

To accompany the description of the algorithm, three experiments demonstrated the various ways the package can be used. The first experiment demonstrated how different similarity kernels can easily be compared, allowing the researcher to select the one that makes the most sense for their project. The second experiment highlighted how different collection strategies can make data acquisition more cost-effective. In particular, active selection combined with 8-choose-2 trial configurations beat out the other options. Lastly, Experiment 3 illustrated how group-specific attention weights can be inferred using a single model–potentially reducing the cost of collecting data. In isolation, the results presented in this work make incremental contributions on four different fronts. As a whole, a meaningful contribution is made by providing a complete top-to-bottom software package.

Acknowledgements

This research was supported by NSF grants SES-1461535, SBE-0542013, SMA-1041755, and DRL-1631428.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* Retrieved from `https://www.tensorflow.org/` (Software available from tensorflow.org)

Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, *35*(3), 283-319.

Fang, Y., & Geman, D. (2005). Experiments in mental face retrieval. In T. Kanade, A. Jain, & N. K. Ratha (Eds.), *Audio-and video-based biometric person authentication* (pp. 637–646). Springer-Verlag.

Ferecatu, M., & Geman, D. (2009, June). A statistical framework for image category search from a mental picture. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*(6), 1087-1101. doi: 10.1109/TPAMI.2008.259

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, *53*, 325-338.

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *32*, 316–332.

Jones, M., Maddox, W. T., & Love, B. C. (2006). The role of similarity in generalization. In *Proceedings of the 28th annual meeting of the cognitive science society* (pp. 405–410).

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Kruskal, J. B. (1968a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*, 1-27.

Kruskal, J. B. (1968b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, *29*, 115-130.

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

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York, NY: Wiley.

Murray, I., Adams, R. P., & MacKay, D. J. (2010). Elliptical slice sampling. In *Proceedings of the 13th international conference on artificial intelligence and statistics (aistats).*

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

Roads, B. D., & Mozer, M. C. (2017). Improving human-machine cooperative classification via cognitive theories of similarity. *Cognitive Science: An Multidisciplinary Journal*, *41*, 1394-1411. doi: 10.1111/cogs.12400

Roads, B. D., Xu, B., Robinson, J. K., & Tanaka, J. W. (2018). The easy-to-hard training advantage with real-world medical images. *Cognitive Research: Principles and Implications*, *3*(38). doi: 10.1186/s41235-018-0131-6

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review*, *89*(4), 305–333. doi: 10.1037/0033-295X.89.4.305

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

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. T. (2011). Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*(3), 457 - 482. Retrieved from `http://www.sciencedirect.com/science/article/pii/001002859190016H` doi: http://dx.doi.org/10.1016/0010-0285(91)90016-H

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In D. C. M. Kearns S. Solla (Ed.), *Advances in neural information processing systems 11* (pp. 59–65). Cambridge, MA: MIT Press.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

van der Maaten, L., & Weinberger, K. (2012, Sept). Stochastic triplet embedding. In *Machine learning for signal processing (mlsp), 2012 ieee international workshop on* (p. 1-6). doi: 10.1109/MLSP.2012.6349720

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The Caltech-UCSD Birds-200-2011 Dataset* (Tech. Rep. No. CNS-TR-2011-001). California Institute of Technology.

Wah, C., Horn, G. V., Branson, S., Maji, S., Perona, P., & Belongie, S. (2014, June). Similarity comparisons for interactive fine-grained categorization. In *Computer vision and pattern recognition (cvpr)*. Columbus, OH.

Wilber, M. J., Kwak, I. S., & Belongie, S. J. (2014). Cost-effective hits for relative similarity comparisons. In *Second aaai conference on human computation and crowdsourcing*.