Finding Feature Representations of Stimuli: Combining Feature Generation and Similarity Judgment Tasks

Matthew D. Zeigenfuse (mzeigenf@uci.edu)
Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Sciences, University of California, Irvine
Irvine, CA, 92697-5100

Abstract

A widely-used assumption cognitive modeling is that stimuli are represented in terms of features. Two experimental approaches to finding appropriate features, and characterizing stimuli in terms of these features, involve feature generation and similarity judgment tasks. In feature generation, people list a set of candidate features, and then decide whether or not each stimulus has each feature. In similarity judgment tasks, people rate the similarity between pairs of stimuli, and models like additive clustering are used to infer features, and their patterns of belonging to stimuli. In this paper, we show how relating feature generation and similarity judgment can provide a powerful method for finding feature representations. We describe a model that constrains a potentially large set of generated features to only those that are needed to explain similarity judgments. Using modern computational Bayesian methods, we apply our model to part of the Leuven natural language database, considering a set of 30 mammals and 764 candidate representational features. We show that the inferred feature representation is interpretable, is able to describe the existing similarities, and provides good generalization performance to withheld similarities.

Keywords: Feature Representation; Feature Generation; Similarity Data; Additive Clustering; Bayesian Methods

Introduction

Model of higher-level cognitive processes must make assumptions about how stimuli are represented. One widely-used assumption is that stimuli are represented in terms of whether or not they have each of a set of binary features. This assumption is widely used in models of memory (e.g., Dennis & Humphreys, 2001; Hintzman, 1984; Raaijmakers & Shiffrin, 1981; Shiffrin & Steyvers, 1997), of category learning (e.g., Medin & Schaffer, 1978; Lee & Navarro, 2002; Love, Medin, & Gureckis, 2004), decision-making (e.g., Gigerenzer & Goldstein, 1996; Payne, Bettman, & Johnson, 1990), and other cognitive processes.

There are at least two well-established experimental classes of tasks that are useful for defining stimuli in terms of a set of features. The first class involves *feature generation* tasks, in which subjects generate a set of candidate representational features, and then judge whether or not each stimulus has each of the features. The second class involves *similarity judgment* tasks, in which subjects assess—via rating scales, association or confusion probabilities, or a range of other possible approaches—the psychological similarity between each pair of stimuli. These similarities can then be used by models like additive clustering (Shepard & Arabie, 1979), or its various extensions (e.g., Navarro & Lee, 2004) to infer a feature set, and the assignments of each feature to each stimulus.

Both classes of experimental task have strengths and weaknesses. Feature generation leads directly to stimulus representations, but requires faith in the introspective accuracy of the people who generate and assign the features. The free-form nature of the generation is especially problematic, because it is not clear how people generate the features, nor how they decide to terminate a list of candidate features. Making inferences from similarity judgment has a better understood theoretical basis, but is much more challenging computationally (see Navarro & Griffiths, in press, for the state of the art). In addition, additive clustering models do not provide any semantic interpretation for the features they derive.

In this paper, we show we show how relating feature generation and similarity judgment can provide a powerful method for addressing the feature representation problem, combining the best aspects of both approaches. We first develop our model, and explain it using a simple toy example. We then apply the model to part of the Leuven natural language database, considering a set of 30 mammals and 764 candidate representational features. We present results showing how the inferred feature representation is interpretable, is able to describe the existing similarities, and provides good generalization performance to withheld similarities.

Relating Feature Generation and Similarity Judgment Tasks

The key insight of our approach is that the observed data in feature generation and similarity judgments tasks can be related to each other, if we assume that both types of data were based on the same underlying feature representations of the stimuli. The form of the relationship between generated features and similarity judgments is shown in Figure 1 as a commutative diagram. In the commutative diagram, $\mathcal{F}$ denotes the 'true' feature-based representations, $\mathcal{F}^+$ denotes the 'augmented' or 'additional' feature representations produced in a generation task, and $\mathcal{S}$ denotes the similarity judgments produced in

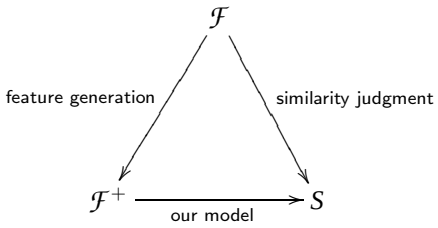

Figure 1: A commutative diagram describing the assumed relationship between the unobserved true feature representation $\mathcal{F}$, the augmented feature representation $\mathcal{F}^+$ observed from a feature generation task, and similarity judgments $\mathcal{S}$ observed from a similarity judgment task.

a similarity task. The important point about Figure 1 is that, if we are willing to make specific assumptions about how the generation and similarity tasks work, we can automatically derive a specific relationship, based on the requirement of commutativity, between the augmented features and the similarities.

Generation and Similarity Assumptions

Although many specific assumptions about feature generation and similarity judgment processes are possible, we focus on the following reasonable starting ones. First, we assume that when people generate feature lists in a generation task, they list all of the true features that are important components of their mental representations of the stimuli, but augment the feature list with additional peripheral features. Secondly, we assume that when people make similarity judgments, they follow the 'common features' approach formalized by additive clustering models (Shepard & Arabie, 1979). That is, people assign a weight or salience to each feature and, when asked the similarity between two stimuli, sum the weights of the features both stimuli have in common.

We denote a feature representation of n stimuli using m features by a matrix $F = [f_{ik}]$, where $f_{ik} = 1$ if the ith stimulus has the kth feature, and $f_{ik} = 0$ if it does not. We denote the pairwise similarities between n features by a matrix $S = [s_{ij}]$, where s_{ij} is the similarity between the ith and jth stimuli. Finally, we denote the weight of the kth feature as w_k.

Using this notation, our account of feature generation is that a $n \times (m + m^+)$ matrix $\mathcal{F}^+$ is generated, containing the m true features, and an additional m^+ features. Our account of similarity judgment is that the matrix $\mathcal{S}$ is generated with each pairwise similarity given by

$$s_{ij} = \sum_{k=1}^{m} w_k f_{ik} f_{jk} + c, \tag{1}$$

where c is a constant corresponding to the minimum level of similarity between stimuli, and can be conceived in the additive clustering model as the weight of a 'universal feature' that all stimuli share.

Our Model

With these assumptions in place, it is straightforward to specify the model that relates the observed feature representation from the generation task to the similarity judgment data. We introduce a set of latent indicator variables, one for each feature, whose role it is to indicate whether each feature is a true feature or an additional feature.

Formally, we denote the latent indicator for the kth feature by z_k, with $z_k = 1$ if the kth feature is part of the underlying representation, but $z_k = 0$ if the kth feature is an additional feature produced during the generation task. This means that the model estimates the similarity between the ith and jth stimuli as

$$s_{ij} = \sum_{k=1}^{(m+m^+)} z_k w_k f_{ik} f_{jk} + c. \tag{2}$$

This model in Equation 2 is easy to interpret. It assumes that only those features belonging to the true underlying representation are evident in the similarity judgments. A feature only influences the similarity between stimuli if it is assigned to be a true feature, with $z_k = 1$.

This means that using the model to make inference about the z_k indicator variables corresponds to 'pruning', 'paring back', or 'regularizing' the augmented feature representation provided by a generation task to a smaller set of true features, based on the information latent in the measures of stimulus similarity. Another way of understanding the model is that the commutative diagram in Figure 1 says that the similarities, s_{ij} in Equations 1 and 2 must be the same, which determines which features are true and which are additional, as formalized by the z_k indicator variables.

Bayesian Inference for the Model

To make inferences using our model, we implemented it as a probabilistic graphical model (see Jordan, 2004; Lee, 2008, for statistical and psychological introductions, respectively). This allows us to undertake fully Bayesian inference on the model, using modern computational methods based on posterior sampling.

To make the model probabilistic, we made the standard assumption (e.g., Tenenbaum, 1996; Lee, 2002a) that observed similarity data are noisy, according to a Gaussian error model with common variance, so that

$$\hat{s}_{ij} \sim \text{Gaussian}\big(s_{ij}, \sigma^2\big). \tag{3}$$

and placed a point prior on the variance, so that

$$\sigma^2 \sim \text{Delta}\big(0.05^2\big), \tag{4}$$

relying on the guidelines develop by (Lee, 2002b). We then placed uniform priors on the feature weights

$$w_k, c \sim \text{Uniform}\big(0, 1\big), \tag{5}$$

Table 1: A feature representation of four animals using three 'true' features, three 'additional' features, and a similarity constant.

Feature	Weight	Dog	Cat	Elephant	Donkey
Kept as pet	0.5	•	•		•
Is hunted	0.2		•	•	
Can be dangerous	0.1		•	•	•
Prime number of letters	—	•	•		
US political mascot	—			•	•
Does not end in "t"	—	•			•
Constant	0.05	•	•	•	•

and a prior on the latent indicators

$$z_k \sim \text{Bernoulli}(\frac{1}{2}), \tag{6}$$

consistent with the assumption that each feature is a priori equally likely to be a true or additional feature. We implemented this model in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000).

An Illustrative Application

In this section we provide a concrete illustrative application of the model. In this application we created a 'toy' domain with a few stimuli and features, and constructed similarity data using the known features and weights. We then test the ability of the model to recover the known true features and their weights.

Features and Similarities

Table 1 gives a feature representation for four animals—a dog, cat, elephant, and monkey—in terms of seven features. Three of these features, "kept as a pet", "is hunted", and "can be dangerous", are given non-zero weights in the representation, and so correspond to true features that are an integral part of the representation of the animals. Three other features, "has a prime number of letters", "is a US political mascot", and "does not end in the letter t" are given zero weights, to indicate that they are additional features. These additional features might be able to be produced in a generation task, but are not an integral part of how animals are represented. The final feature is the constant, with a weight that gives the level of similarity all animals share.

Using the representation in Table 1, we generated artificial similarity data, using the similarity judgment process described earlier. That is, we added the weights of the common features for each pair of animals, and added Gaussian noise with a variance of 0.5^2, to produce the similarity matrix

$$S = \begin{bmatrix} - & 0.541 & 0.043 & 0.547 \\ & - & 0.361 & 0.652 \\ & & - & 0.155 \\ & & & - \end{bmatrix}.$$

Modeling Results

We then applied our model to the similarities in S and the feature matrix F given by Table 1, recording 2,000 posterior samples from 4 chains after a burn-in period of 3,000 samples. The recorded samples are treated as draws from the full joint posterior distribution of the weights $\mathbf{w} = (w_1, \ldots, w_k, c)$ and indicator variables $\mathbf{z} = (z_1, \ldots, z_k)$.

Table 2: The seven patterns of indicator variable assignments found in the posterior samples, together with their proportion in the sample.

Pattern	Proportion	Pet	Hunted	Dangerous	Prime	Mascot	Not end "t"
1	0.94	•	•	•			
2	0.02	•	•	•		•	•
3	0.02	•	•	•			•
4	0.01	•	•	•	•		
5	< 0.01	•	•	•	•	•	
6	< 0.01	•	•	•	•		•
7	< 0.01	•	•	•		•	•

The key analysis of the model's inferences involves the joint posterior over the indicator variables. Of the 2^6 possible combinations that could be sampled (i.e., all possible patterns of features being true versus additional features), only seven were ever sampled. These seven patterns are detailed in Table 3, together with their observed proportion in the sample, which approximates their posterior mass. Each pattern corresponds to a different inference about which features are true and which are additional, and the mass measures the certainty with which each combination is the correct pattern.

The MAP assignments (i.e., the assignments with the most posterior mass) given by pattern 1 dominate the posterior, and have the right structure. In particular, the first three features—pet, hunted and dangerous—are assigned as true features, while the others are assigned as additional features, following the design we used to generate the data.

The posterior distribution of the weights conditional on the assignments given by pattern 1 also show the model is making the right inferences. The marginal expected values for the weights of the true features and constant were $w_1 = 0.499$, $w_2 = 0.207$, $w_3 = 0.109$, and $c = 0.043$, all of which are very close to the original values in Table 1.

This simple example illustrates how our model can identify just those generated features that play a role in the judgment of stimulus similarity. While it is possible to characterize animals, or any other stimuli, in terms of an endless number of features, only some features are important for representing and understanding. In this example, the spurious features like "is hunted" were inferred to be important in explaining similarity, while features like "has a prime number of letters" were inferred to be unimportant.

Application to Human Data

In this section we apply the model to large-scale feature generation and similarity data collected experimentally. The basic form of the data, and the way in which we apply our model, is identical to the illustrative example. We just work with many more stimuli and features.

Feature and Similarity Data

Our data come from the Leuven natural language concepts database (De Deyne et al., 2008), which includes feature generation task and similarity judgment task data for a number of semantic categories[1]. We considered just the "mammal" category, which gives the feature representation of 30 mammals in terms of 764 features. The feature list was produced by one set of participants in a generation task, and the assessment of whether each mammal had each feature was done by a different set of participants. A third set of participants made similarity judgments for every pair of mammals on a 20-point Likert scale.

We performed the following pre-processing steps. First any feature belonging to none of the stimuli, or just one stimulus, was removed, because it can have no impact on similarity under the additive clustering model we assume. Secondly, any feature that belonged to all stimuli was removed, because it can be conceived as part of the overall similarity constant. Finally, if more than one feature had exactly the same pattern of belonging to stimuli, we retained only the most frequently generated of those features. Making these changes reduced the total list to 516 features.

Modeling Results

True and Additional Features Figure 2 summarizes the posterior distribution over the latent indicator variables. It shows the marginal posterior mass for each of the 516 features, ordered in terms of decreasing posterior mass. There are three visually distinct classes of features. The first 69 features are nearly always classified as true features, with posterior mass near one. Features 70–180 have posterior mass nearer 0.5, showing that there is uncertainty as to whether they are true or additional features. The remaining features from 181–516 have almost no posterior mass, and so are almost always classified as additional features.

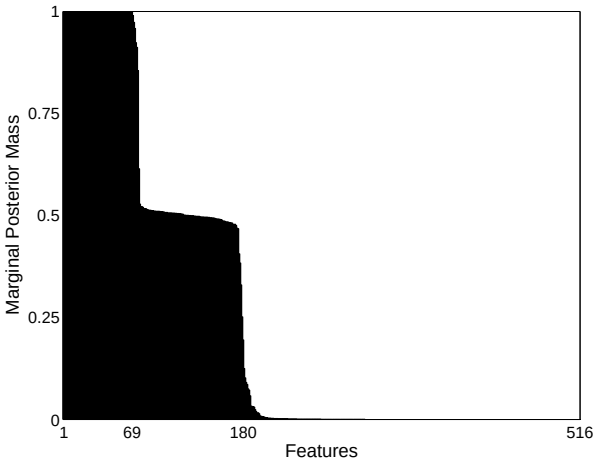

Figure 2: Marginal posterior mass for the latent indicators of the 516 features, ordered in terms of decreasing mass.

Table 3: Examples of features always classified as true features, and always classified as additional features.

True features	Additional features
Is big	Appears in comics
Is noisy	Is smaller than a horse
Lives in a herd	Is spectacular
Has a mane	Does not have 1,000 paws
Stands in meadows	Stands in the crib at Christmas
...	...

It is not possible to list all 516 features, but an examination of which features were consistently classified as true features, and which were consistently classified as additional features, gives generally easily interpreted results. Table 3 gives some representative examples of both cases. All of the true features seem to correspond to important semantic properties needed to describe the relationships between mammals. The additional features seem much less important.

The results in Figure 2 and Table 3 show that the latent indicators identify three classes of features, and that the differences between the classes are generally interpretable.

Capturing Similarities We also examined how our model helps us account for observed patterns of similarities between the mammals, by considering two feature

[1]The database is available at http://ppw.kuleuven.be/concat/

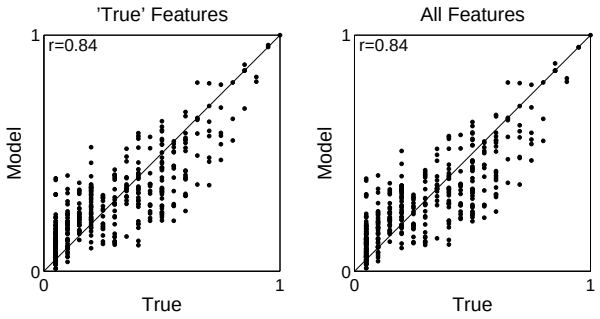

Figure 3: The posterior performance of the model account for the similarity judgments, using only inferred true features (left panel) and using all features (right panel).

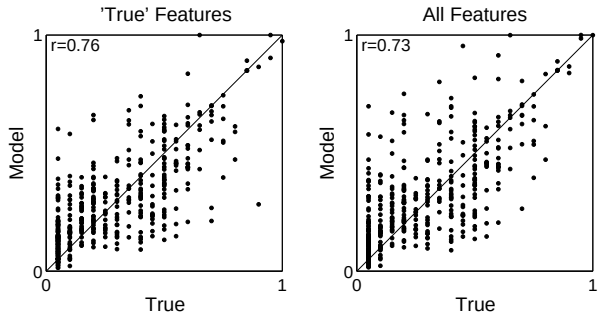

Figure 4: The posterior performance of the model account for the similarity judgments, when approximately 70% of the judgments were removed, using only inferred true features (left panel) and using all features (right panel).

representations. One simply uses all 516 features. The second uses only those 69 features are model identified as almost certainly being true features.

We had both feature representations learn appropriate weights for all of their features from the similarity judgments, under the additive clustering assumptions, and examined how well they were able to fit the pairwise similarities. In our Bayesian context, the fits are assessed by generating posterior predictions (i.e., the distribution of similarities a feature representation generates integrating over its posterior distribution for the weight parameters), and comparing these predictive distributions to the observed data.

Figure 3 summarizes these tests by showing the expected posterior predicted similarities of each feature representation against the observed data. It can be seen that both representations perform extremely similarly, and that both perform well, having a $r = 0.84$ correlation with the data. The important point is that the feature representation using only the true features identified by our model shows no decrement in descriptive adequacy. That is, Figure 3 provides strong evidence the 69 features we identified are adequare to account for the patterns of similarities between the mammals.

Generalization A feature representation of stimuli should serve not just to summarize what is known about relationships between stimuli, but also to generalize to new situations where observational data are not yet available. To test how the two feature representations performed in this situation, we re-ran the posterior predictive assessments, supplying only a fraction of the original similarity judgments. This means that the weights for the features must be inferred from fewer pairwise similarity comparisons, and what is learned must be used to make predictions about the similarity relationships between mammal pairs that are not provided.

Figure 4 shows the performance of both representations when approximately 70% of the similarity judgments were withheld. The representation using only true features performs slightly better, because the more complicated representation using all of the features over-fit the available data, and so generalized less accurately (Pitt & Myung, 2002). While the difference is correlations is not large, it is still impressive that the true feature representation performs better, because it did not use the large set of features 70–180 in Figure 2 that are likely to play some role in similarity.

We also note that the impressive absolute level of correlation for both representations with so many similarity data withheld supports our assumptions that the feature generation and similarity judgment data are closely linked, and that the additive clustering model is a useful and appropriate one.

Discussion

In this paper, we developed a model that relates the data from feature generation and similarity judgment tasks, and is able to use the relationship to determine a set of 'true' features, and their salience in assessing similarity. As a practical method for generating feature representations, our model has a number of attractive properties. Compared to methods based solely on similarity data, our model needs to solve a relatively simple inference problem, and so scales easily to large problem sizes. It also automatically provides a semantic label for each feature. Unlike methods based solely on feature generation, our model is able to determine which features are appropriate, how many there should be, and how they should be weighted.

It is straightforward to extend our model in a number of different complementary directions. Alternative assumptions about how features are generated or how similarity judgments are made will automatically lead to alternative models. For feature generation, one obvious possibility is to weaken the assumption that all true features are included in the generated list, and allow for the possibility some (relatively small) set of features and their patterns of belonging to stimuli must still

be inferred. For similarity assessment, the reliance on common features could be relaxed, to allow distinctive features to play a role in how stimuli are related (e.g., Navarro & Lee, 2004).

A more dramatic extensions would involve alternatives to feature representation, perhaps allowing for the more structured accounts provided by trees or other graph structures, or allowing for continuous dimensions to underlie some aspects of how stimuli are represented mentally (e.g., Kemp, Bernstein, & Tenenbaum, 2005; Navarro & Lee, 2002). It would also be possible to change or extend the types of tasks considered, including others that are driven by feature representations of stimuli, such as category learning or analogy making tasks.

Finally, we argue that the basic idea of using behavior in multiple tasks to understand a latent psychological variable manifest in all of the tasks, is a general and potentially very powerful one. It is unlikely any one task, or any one model of human performance on that task, will be sufficient to characterize complicated psychological constructs. The approach we have demonstrated, using the relationship between observed data across different related tasks to make inferences that neither task alone supports, is one promising way of addressing this fundamental challenge.

References

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., et al. (2008). *Exemplar by feature applicability matrices and other dutch normative data for semantic concepts.*

Dennis, S. J., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*(2), 452–478.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Hintzman, D. L. (1984). Minerva-2 - a simulation-model of human-memory. *Behavior Research Methods Instruments & Computers*, *16*(2), 96-101.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, *19*, 140–155.

Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Lee, M. D. (2002a). Generating additive clustering models with limited stochastic complexity. *Journal of Classification*, *19*(1), 69–85.

Lee, M. D. (2002b). A simple method for generating additive clustering models with limited complexity. *Machine Learning*, *49*, 39–58.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15.

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

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

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS: A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

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

Navarro, D. J., & Griffiths, T. L. (in press). Latent features in similarity judgment: A nonparametric Bayesian approach. *Neural Computation*.

Navarro, D. J., & Lee, M. D. (2002). Commonalities and distinctions in featural stimulus representations. In W. G. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th annual conference of the cognitive science society* (pp. 685–690). Mahwah, NJ: Erlbaum.

Navarro, D. J., & Lee, M. D. (2004). Common and distinctive features in stimulus similarity: A modified version of the contrast model. *Psychonomic Bulletin & Review*, *11*(6), 961–974.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1990). *The adaptive decision maker*. New York: Cambridge University Press.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(2), 421–425.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

Shepard, R. N., & Arabie, P. (1979). Additive clustering representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87–123.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Tenenbaum, J. B. (1996). Learning the structure of similarity. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems* (Vol. 8, pp. 3–9). Cambridge, MA: MIT Press.