Clustering Using the Contrast Model

Daniel J. Navarro and Michael D. Lee
{daniel.navarro,michael.lee}@psychology.adelaide.edu.au
Department of Psychology, University of Adelaide
South Australia, 5005, AUSTRALIA

Abstract

An algorithm is developed for generating featural representations from similarity data using Tversky's (1977) Contrast Model. Unlike previous additive clustering approaches, the algorithm fits a representational model that allows for stimulus similarity to be measured in terms of both common and distinctive features. The important issue of striking an appropriate balance between data fit and representational complexity is addressed through the use of the Geometric Complexity Criterion to guide model selection. The ability of the algorithm to recover known featural representations from noisy data is tested, and it is also applied to real data measuring the similarity of kinship terms.

Introduction

Understanding human mental representation is necessary for understanding human perception, cognition, decision making, and action. Mental representations play an important role in mediating adaptive behavior, and form the basis for the cognitive processes of generalization, inference and learning. Different assumptions regarding the nature and form of mental representation lead to different constraints on formal models of these processes. For this reason, Pinker (1998) argues that "pinning down mental representation is the route to rigor in psychology" (p. 85). Certainly, it is important that cognitive models use principled mental representations, since the *ad hoc* definition of stimuli on the basis of intuitive reasonableness is a highly questionable practice (Brooks 1991, Komatsu 1992, Lee 1998).

One appealing and widely used approach for deriving stimulus representations is to base them on measures of stimulus similarity. Following Shepard (1987), similarity may be understood as a measure of the degree to which the consequences of one stimulus generalize to another, and so it makes adaptive sense to give more similar stimuli mental representations that are themselves more similar. For a domain with n stimuli, similarity data take the form of an $n \times n$ similarity matrix, $\mathbf{S} = [s_{ij}]$, where s_{ij} is the similarity of the ith and jth stimuli. The goal of similarity-based representation is then to define stimulus representations that, under a given similarity model, capture the constraints implicit in the similarity matrix by approximating the data.

Goldstone's (in press) recent review identifies four broad model classes for stimulus similarity: geometric, featural, alignment-based, and transformational. Of these, the two most widely used approaches are the geometric, where stimuli are represented in terms of their values on different dimensions, and the featural, where stimuli are represented in terms of the presence or absence of weighted features. The geometric approach is most often used in formal models of cognitive processes, partly because of the ready availability of techniques such as multidimensional scaling (e.g., Kruskal 1964; see Cox & Cox 1994 for an overview), which generate geometric representations from similarity data. The featural approach to stimulus representation, however, is at least as important as the geometric approach, and warrants the development of techniques analogous to multidimensional scaling.

Accordingly, this paper describes an algorithm that generates featural representations from similarity data. The optimization processes used in the algorithm are standard ones, and could almost certainly be improved. In this regard, we draw on Shepard and Arabie's (1979) distinction between the psychological model that is being fit, and the algorithm that does the fitting. We make no claims regarding the significance of the algorithm itself (and certainly do not claim it is a model of the way humans learn mental representations), but believe that the psychological representational model that it fits has three important properties. First, it allows for the arbitrary definition of features, avoiding the limitations of partitioning or hierarchical clustering. Second, it uses a more general model of featural stimulus similarity than has previously been considered. Third, it generates featural representations in a way that balances the competing demands of data-fit and representational complexity.

Featural Representation

Within a featural representation, stimuli are defined by the presence or absence of a set of saliency weighted

features or properties. Formally, if a stimulus domain contains n stimuli and m features, a featural representation is given by the $n \times m$ matrix $\mathbf{F} = [f_{ik}]$, where

$$f_{ik} = \begin{cases} 1 & \text{if stimulus } i \text{ has feature } k \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

together with a vector $\mathbf{w} = (w_1, \ldots, w_m)$ giving the (positive) weights of each of the features.

The Contrast and Ratio Models

Tversky's (1977) Contrast Model and Ratio Model of stimulus similarity provide a rich range of possibilities for generating featural representations that have been significantly under-utilized. Using the assumption that the similarity between two stimuli is a function of their common and distinctive features, the Contrast Model measures stimulus similarity as:

$$\hat{s}_{ij} = \theta F(\mathbf{f}_i \cap \mathbf{f}_j) - \alpha F(\mathbf{f}_i - \mathbf{f}_j) - \beta F(\mathbf{f}_j - \mathbf{f}_i), \quad (2)$$

where $\mathbf{f}_i \cap \mathbf{f}_j$ denotes the features common to the ith and jth stimuli, $\mathbf{f}_i - \mathbf{f}_j$ denotes the features present in the ith, but not the jth, stimulus, and $F(\cdot)$ is some monotonically increasing function. By manipulating the positive weighting hyper-parameters θ, α and β, different degrees of importance may given to the common and distinctive components. In particular, Tversky (1977) emphasizes the two extreme alternatives obtained by setting $\theta = 1, \alpha = \beta = 0$ (common features only), and $\theta = 0, \alpha = \beta = 1$ (distinctive features only). A different approach is given by the Ratio Model, where similarity takes the form:

$$\hat{s}_{ij} = \frac{\theta F(\mathbf{f}_i \cap \mathbf{f}_j)}{\theta F(\mathbf{f}_i \cap \mathbf{f}_j) + \alpha F(\mathbf{f}_i - \mathbf{f}_j) + \beta F(\mathbf{f}_j - \mathbf{f}_i)}. \quad (3)$$

While the Contrast Model and the Ratio Model provide great flexibility for measuring similarity across featural representations, the only established techniques for generating the representations from similarity data are additive clustering algorithms (e.g., Arabie & Carroll 1980; Lee 1999, in press; Mirkin 1987; Shepard & Arabie 1979; Tenenbaum 1996), which rely exclusively on the common features version of the Contrast Model. This means that only one special case of one of these approaches has been used as the basis of a practical technique for generating representations.

The paucity of available techniques is serious, given the recognition (e.g., Goodman 1972; Rips 1989; see Goldstone 1994 for an overview) that similarity is not a unitary phenomenon, and the way in which it is measured may change according to different cognitive demands. Direct empirical evidence that featural similarity judgments can place varying emphasis on common and distinctive features is provided by the finding that items presented in written form elicit common feature-weighted judgments, whereas pictures tend to be rated more in terms of distinctive features (Gati & Tversky 1984; Tversky & Gati 1978).

A Symmetric Contrast Model

Although the Contrast Model has three hyper-parameters, α and β remain distinct only when $s_{ij} \neq s_{ji}$. While it is certainly the case that real world domains display asymmetric similarity, modeling techniques based on similarity data generally assume that similarity is symmetric. Further, if the similarity ratings are assumed to lie between 0 and 1, the remaining hyper-parameters α and θ can be incorporated into one parameter, $\rho = \theta / (\theta + \alpha)$, which represents the relative weighting of common and distinctive features, with $0 \leq \rho \leq 1$. Setting the functional form $F(\cdot)$ using the same 'sum of saliency weights' approach as additive clustering yields the similarity model

$$\begin{aligned} \hat{s}_{ij} &= \rho \sum_k w_k f_{ik} f_{jk} - \frac{1-\rho}{2} \left[\sum_k w_k f_{ik} (1 - f_{jk}) \right. \\ &\quad \left. + \sum_k w_k (1 - f_{ik}) f_{jk} \right] + c. \end{aligned} \quad (4)$$

It is this symmetric version of the Contrast Model that is used in this paper to develop general featural representations. It allows for any relative degree of emphasis to be placed on common and distinctive features and, in particular, subsumes the additive clustering model ($\rho = 1$) and the distance-based feature-matching similarity model ($\rho = 0$). Technically, it is worth noting that the additive constant c used in additive clustering, which is added to all pairwise similarity estimates in both additive clustering and Contrast Model clustering representations, is not treated as a cluster, and thus is not weighted by ρ.

Limiting Representational Complexity

Shepard and Arabie (1979) have noted that the ability to specify large numbers of features and set their weights allows any similarity matrix to be modeled perfectly by a featural representation using the common features version of the Contrast Model. The same is true for the majority of Tversky's (1977) similarity models, and is certainly true for Eq. (4). While the representational power to model data is desirable, the introduction of unconstrained feature structures with free parameters detracts from fundamental modeling goals, such as the achievement of interpretability, explanatory insight, and the ability to generalize accurately beyond given information (Lee 2001a).

This means that techniques for generating featural representations from similarity data must balance the competing demands of maximizing accuracy and minimizing complexity, following the basic principle of model selection known as 'Ockham's Razor' (Myung & Pitt 1997). Data precision must also be considered, since precise data warrants a representation be-

ing made more detailed to improve data-fit, while noisy data does not.

In practice, this means that featural representations should not be derived solely on the basis of how well they fit the data, as quantified by a measure such as the variance accounted for,

$$\text{VAF} = 1 - \frac{\sum_{i<j}(s_{ij} - \hat{s}_{ij})^2}{\sum_{i<j}(s_{ij} - \bar{s})^2}, \tag{5}$$

where $\bar{s}$ is the arithmetic mean of the similarity data. Rather, some form of complexity control must be used to balance data-fit with model complexity. Most established algorithms strike this balance in unsatisfactory ways, either pre-determining a fixed number of clusters (e.g., Shepard & Arabie 1979; Tenenbaum 1996), or pre-determining a fixed level of representational accuracy (e.g., Lee 1999).

Recently, Lee (in press) has applied the Bayesian Information Criterion (BIC: Schwarz 1978) to limit the complexity of additive clustering representations. Unfortunately, an important limitation of the BIC is that it equates model complexity with the number of parameters in the model. While this is often a reasonable approximation, it neglects what Myung and Pitt (1997) term the 'functional form' component of model complexity. For featural representations, parametric complexity is simply the number of features used in a representation. Functional form complexity, however, considers the feature structure $\mathbf{F}$, and is sensitive to the patterns with which stimuli share features (see Lee 2001a), as well as any difference arising from the relative emphasis given to common and distinctive features.

It is important to account for functional form complexity with featural representational models that can vary their emphasis on common and distinctive features. Figure 1 shows the results of fitting featural representations, assuming different levels of ρ, on similarity data that were generated using either entirely common features ($\rho = 1$), entirely distinctive features ($\rho = 0$), or an even balance of the two ($\rho = 0.5$). These results are averaged across five different similarity matrices, each based on a five-feature representation, and show one standard error about the mean level of fit.

As expected, the best-fitting featural representations have ρ values matching those that generated the data. More interestingly, Figure 1 shows that the level of fit for the entirely common features data deteriorates more rapidly than for the entirely distinctive features data when the wrong ρ value is assumed. Similarly, for the evenly balanced data, the fit is greater when too much emphasis is placed on common features in the assumed similarity model. These results imply that common features-weighted models are more able to fit data when they are wrong than are distinctive features-weighted models. In the language of model complexity,

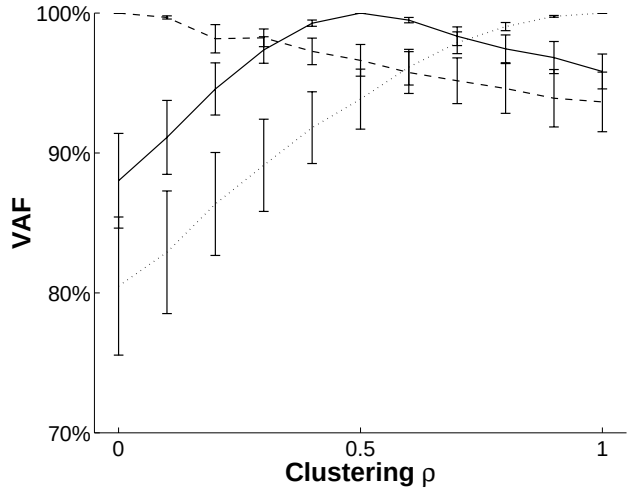

Figure 1: The change in VAF value, as a function of the assumed balance between common and distinctive features, for the entirely common (dotted line), entirely distinctive (dashed line) and balanced (solid line) similarity data.

the common features functional form is more flexible than the distinctive features functional form, and this extra complexity improves the fit of incorrect models. For this reason, it is important to derive featural representations using a measure that is sensitive to functional form complexity.

A Geometric Complexity Criterion

Myung, Balasubramanian, and Pitt (2000) have recently developed a measure called the Geometric Complexity Criterion (GCC) that constitutes the state-of-the-art in accounting for both fit and complexity in model selection. The basic idea is to define complexity in terms of the number of distinguishable data distributions that the model can accommodate through parametric variation, with more complicated models being able to index more distributions than simple ones. Using Tenenbaum's (1996) probabilistic formulation of the data-fit of a featural model, and extending Lee's (2001a) derivation of the Fisher Information matrix for the common features case of the Contrast Model, it is a reasonably straightforward exercise to derive a GCC for the current similarity model. The final result is:

$$\begin{aligned}\text{GCC} &= \frac{1}{2s^2}\sum_{i<j}(s_{ij} - \hat{s}_{ij})^2 + \frac{m+1}{2}\ln\left(\frac{n(n-1)}{4\pi s^2}\right) \\ &\quad + \frac{1}{2}\ln\det G, \end{aligned} \tag{6}$$

where s denotes an estimate of the inherent precision of the data (see Lee 2001b), m is the number of features,

n is the number of stimuli, and G denotes the $m \times m$ *complexity matrix* for the feature structure. The xy-th cell of the complexity matrix is given by,

$$\sum_{i<j} e_{ijx} e_{ijy} \tag{7}$$

where e_{ijx} equals ρ if x is a common feature, $-(1-\rho)/2$ if x is a distinctive feature, and 0 if neither i nor j possesses the feature x.

An interesting aspect of the complexity matrix, and the GCC measure as a whole, is that it is independent of the parameterization of the model. That is, the complexity of a featural representation is dependent only on the feature structure, and not the saliencies assigned to the features. We should make two technical points about the GCC. First, this derivation is based on the assumption that ρ is a fixed property of a model, and not a free parameter. An alternative would be to modify the GCC so that it accommodated ρ as a model parameter. Second, since the additive constant is not weighted by ρ, the terms in the complexity matrix corresponding to the additive constant behave as if $\rho = 1$.

Algorithm

In developing an algorithm to fit featural representations using the Contrast Model, we were guided by the successful additive clustering algorithm reported by Lee (submitted). Basically, the algorithm works by 'growing' a featural representation, starting with a one-feature model, and continually adding features while this leads to improvements in the GCC measure. For any fixed number of features, the search for an appropriate assignment of stimuli to features is done using stochastic hill-climbing, with the best-fitting weights being determined using a standard non-negative least squares algorithm (Lawson & Hanson 1974). The algorithm terminates once the process of adding features leads to representations with GCC values that are more than a pre-specified constant above the best previously found, and the featural representation with the minimum GCC value is returned.

To test the ability of this optimization algorithm to fit similarity data, we examined its ability to recover a known featural representation. This representation had seven stimuli and four features, and included partitioning, nested, and overlapping clusters, as shown in Figure 2. Using this representation, similarity data were generated assuming entirely common features, entirely distinctive features, or an even balance between the two. Feature weights were chosen at random subject to the constraint that they resulted in positive similarity values. Each of the similarity values was perturbed by adding noise that was independently drawn from a Normal distribution with mean 0 and standard deviation 0.05.

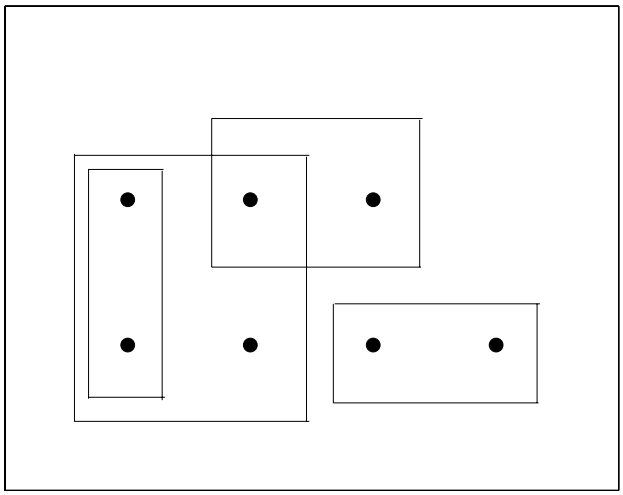

Figure 2: The artificial featural representation containing seven stimuli and four features.

The algorithm was applied to this similarity data under different assumptions regarding the balance between common and distinctive features, using ρ values of 0, 0.25, 0.5, 0.75 and 1. In calculating the GCC measure, a data precision value of $\sigma = 0.05$ was assumed, in accordance with the known level of noise. Figure 3 summarizes the results of 10 runs of the algorithm for each of the three similarity conditions, across all of the assumed ρ values. The mean GCC value of the 10 derived representations is shown, together with error bars showing one standard error in both directions.

Figure 3 shows that the GCC is minimized at the correct ρ value for all three similarity conditions. An examination of the derived representation revealed that the correct featural representation was recovered 25 times out of 30 attempts: nine times out of ten for the entirely distinctive data, and eight times out of ten for the evenly balanced and the entirely common data. It is interesting to note that Figure 3 is far more symmetric than Figure 1, suggesting that the GCC has successfully accounted for the differences in functional form complexity between the common and distinctive feature approaches to measuring similarity.

Additional Monte Carlo simulations with other featural representations, based on particular structures reported by Tenenbaum (1996, Table 1) and Lee (1999, Table 5), also suggested that the algorithm is capable of recovering known configurations when more stimuli or more features are involved, although problems with local minima are encountered more frequently.

An Illustrative Example

To demonstrate the practical application of the algorithm, we used the averaged similarity data reported

Table 1: Representation of Rosenberg and Kim's (1975) kinship terms domain.

Stimuli in Cluster	Weight
aunt uncle niece nephew cousin	0.319
granddaughter grandson grandmother grandfather	0.291
mother daughter grandmother granddaughter aunt niece sister	0.222
sister brother cousin	0.221
father son grandfather grandson uncle nephew brother	0.208
mother father daughter son sister brother	0.163
mother father daughter son	0.136
daughter son granddaughter grandson niece nephew sister brother	0.128
mother father grandmother grandfather aunt uncle sister brother	0.091
additive constant	0.563
Variance Accounted For	92.7%

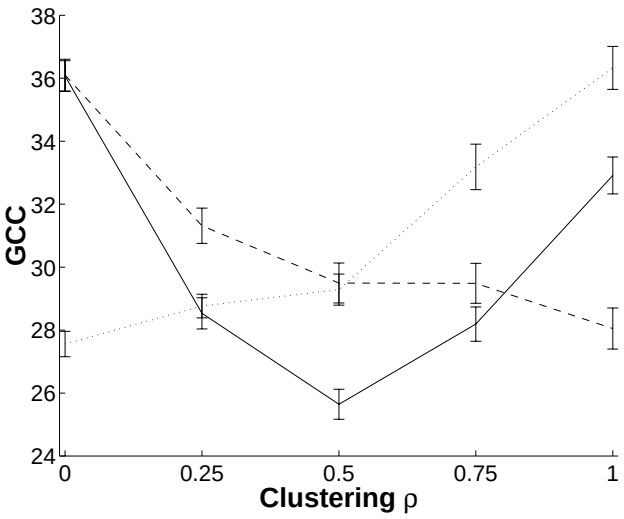

Figure 3: The change in GCC value, as a function of the assumed balance between common and distinctive features, for the entirely common (dotted line), entirely distinctive (dashed line) and balanced (solid line) similarity data.

by Rosenberg and Kim (1975), which measures similarity of English kinship terms. A data precision estimate of $s = 0.09$ was made based on the sample standard deviation of the individual matrices. Since the data was obtained by having participants sort items into different stacks, we might expect a model that provides a weighting of common and distinctive features to provide a better fit than one allowing only for common features. Using ρ values of 0, 0.1, 0.2, ..., 1.0, the representation with the minimum GCC was found at $\rho = 0.4$.

This representation contained the nine features detailed in Table 1, and explained 92.7% of the variance in the data. Interpreting most of the features in Table 1 is straightforward, since they essentially capture concepts such as 'male', 'female', 'nuclear family', 'extended family', 'grandparents', 'descendants', and 'progenitors'. While this representation is very similar to the nine-feature representation generated by additive clustering (Lee submitted, Figure 2), it explains more of the variance in the data, suggesting that participants did indeed use both common and distinctive features in assessing similarity.

Conclusion

We have developed, tested, and demonstrated an algorithm that generates featural stimulus representations from similarity data. Unlike previous additive clustering approaches, the algorithm uses a symmetric version of Tversky's (1977) Contrast Model that measures similarity in terms of both common and distinctive features. A particular strength of the algorithm is its use of the Geometric Complexity Criterion to guide the generation process, which allows the desire for data-fit to be balanced with the need to control representational complexity. Importantly, this criterion is sensitive to the functional form complexity of the similarity model, preventing an over-emphasis on the inherently more complicated common features approach.

In terms of future work, it should be acknowledged that the symmetric version of the Contrast Model is certainly not the only possibility for combining common and distinctive features approaches to measuring similarity. Tenenbaum and Griffiths (in press) provide a compelling argument for the use of the Ratio Model in the context of their Bayesian theory of generalization. It would also be worthwhile to examine featural representations where each feature is assumed to operate using entirely an distinctive or an entirely common approach. The distinctive similarity features would be those that globally partition the entire stimulus set, as

for the feature 'male', which implies the existence of the complementary feature 'female'. The (more prevalent) common similarity features would be those that captured shared properties, such as eye or hair color, where no broader implications are warranted.

Acknowledgments

This article was supported by a Defence Science and Technology Organisation scholarship awarded to the first author. We wish to thank several referees for helpful comments on an earlier version of this paper.

References

Arabie, P., & Carroll, J.D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika 45*(2), 211–235.

Brooks, R.A. (1991). Intelligence without representation. *Artificial Intelligence 47*, 139–159.

Cox, T.F., & Cox, M.A.A. (1994). *Multidimensional scaling.* London: Chapman and Hall.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgments. *Cognitive Psychology 16*, 341–370.

Goldstone, R.L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition 52*, 125–157.

Goldstone, R.L. (in press). Similarity. In R.A. Wilson & F.C. Keil (Eds.) *MIT Encyclopedia of the Cognitive Sciences.* Cambridge, MA: MIT Press.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and Projects*, pp. 437–446. New York: Bobbs-Merrill.

Lawson, C.L., & Hanson, R.J. (1974).*Solving Least Squares Problems.* Englewood Cliffs, NJ: Prentice-Hall.

Lee, M.D. (1998). Neural feature abstraction from judgments of similarity. *Neural Computation, 10* (7), 1815-1830.

Lee, M.D. (1999). An extraction and regularization approach to additive clustering. *Journal of Classification, 16* (2), 255-281.

Lee, M.D. (2001a). On the complexity of additive clustering models. *Journal of Mathematical Psychology, 45* (1), 131-148.

Lee, M.D. (2001b). Determining the dimensionality of multidimensional scaling models for cognitive modeling. *Journal of Mathematical Psychology, 45* (1), 149-166.

Lee, M.D. (in press). A simple method for generating additive clustering models with limited complexity. *Machine Learning.*

Lee, M.D. (submitted). *Generating additive clustering models with limited stochastic complexity.* Manuscript submitted for publication.

Komatsu, L.K. (1992). Recent views of conceptual structure. *Psychological Bulletin 112*(3), 500–526.

Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29* (1), 1–27.

Mirkin, B.G. (1987). Additive clustering and qualitative factor analysis methods for similarity matrices. *Journal of Classification 4*, 7–31.

Myung, I.J., Balasubramanian, V., & Pitt, M.A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA 97*, 11170–11175.

Myung, I.J., & Pitt, M.A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review 4*(1), 79–95.

Pinker, S. (1998). *How the mind works.* Great Britain: The Softback Preview.

Rips, L.J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning*, pp. 21–59. New York: Cambridge University Press.

Rosenberg, S., & Kim, M.P. (1975). The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research 10*, 489–502.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

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

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

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*, Volume 8, pp. 3–9. Cambridge, MA: MIT Press.

Tenenbaum, J.B., & Griffiths, T.L. (in press). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences 24*(2).

Tversky, A. (1977). Features of similarity. *Psychological Review 84*(4), 327–352.

Tversky, A., & Gati, I. (1978). Studies of similarity. In E. Rosch and B.B. Lloyd (Eds.), *Cognition and Categorization*, pp. 79–98. Hillsdale, NJ: Wiley.