Commonalities and Distinctions in Featural Stimulus Representations

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

Abstract

This paper evaluates four featural models of stimulus similarity using data collected for a set of 16 nations. Algorithms are developed for finding stimulus representations, and the important issue of balancing data-fit against model complexity is addressed by using the Geometric Complexity Criterion. Although the data clearly incorporate both common and distinctive features, Tversky's (1977) Contrast Model seems unable to express these regularities in an appropriate manner. However, we show that a new version of the Contrast Model that treats each feature as either being common or distinctive is better able to capture the essential aspects of the similarity judgments.

Featural Representation

A fundamental issue in psychology regards the appropriate manner in which to represent stimuli in a model of human cognition. As argued by Brooks (1991), it is important to constrain representations to those justified by empirical data, and avoid the questionable practice of specifying representations "by hand" . One well-established technique for pinning down mental representation involves measuring the similarity between pairs of stimuli. The assumption underlying this approach is that the decision process involved in judging similarity is a simple one, and thus the data can be considered to reflect the underlying mental representation to a large extent. While this is not without theoretical difficulties (e.g., Goodman, 1972; Goldstone, Medin, & Halberstadt, 1997), it is substantially superior to the alternative approach of hand-tuning representations, which may not reflect human representational structures in any regard. Goldstone's (1999) recent review identifies four main approaches to similarity modeling: geometric, featural, alignment-based and transformational. In this article we discuss current approaches to featural representation, and provide experimental evidence to support a new approach for modeling featural similarity.

The featural approach to mental representation describes an object in terms of the attributes it possesses. Features may be either perceptual or conceptual in nature: for example, a tiger might possess the features "four legged", "orange", and "predatory". The task of deriving featural representations from similarity data can be stated as follows: if n denotes the number of stimuli in the domain, then given an $n \times n$ matrix of similarity judgments $\mathbf{S}$, find a set of m features that explain these judgments. We can denote this set of features by the $n \times m$ feature matrix $\mathbf{F} = [f_{ik}]$, where f_{ik} is 1 if the ith stimulus possesses the kth feature, and 0 if it does not.

Four Models of Featural Similarity

One well-established approach for extracting featural representations from similarity data involves using additive clustering algorithms (e.g, Shepard & Arabie, 1979; Tenenbaum, 1996). The similarity between two stimuli is estimated as the sum of the weights of their common features (i.e., those that they both possess). That is,

$$\hat{s}_{ij} = \sum_k w_k f_{ik} f_{jk} + c, \tag{1}$$

where w_k denotes the saliency weight of the kth feature, and c is a positive-valued constant added to all similarity estimates. Thus an m-feature common features representation consists of the feature matrix $\mathbf{F}$, the vector of saliency weights $\mathbf{w} = [w_1, w_2, \ldots, w_m]$ and the additive constant. As noted above, additive clustering relies on a purely common features model. This means that the stimuli become more similar only to the extent that they share features.

An alternative featural model is the distinctive features model, under which similarity is measured according to the differences between the features that stimuli have. This means that if one stimulus has a feature and another does not, they become less similar. This can be written as

$$\hat{s}_{ij} = c - \frac{1}{2}\sum_k w_k f_{ik}\,(1 - f_{jk}) - \frac{1}{2}\sum_k w_k\,(1 - f_{ik})\,f_{jk}, \tag{2}$$

which is identical to the symmetric distance metric proposed by Restle (1959), and closely related

to discrete multidimensional scaling (Clouse & Cottrell, 1996; Rohde, in press).

A general framework that interpolates between these two models is Navarro and Lee's (2001) adaptation of Tversky's (1977; Gati & Tversky, 1984) Contrast Model (TCM), consisting of a weighted sum of the common features similarity (Eq. 1) and the distinctive features similarity (Eq. 2). If we let $0 \leq \rho \leq 1$ denote the weighting given to the common features component, then this model is given by

$$\begin{aligned} \hat{s}_{ij} &= \rho \sum_k w_k f_{ik} f_{jk} - \frac{1-\rho}{2} \sum_k w_k f_{ik} (1 - f_{jk}) \\ &\quad - \frac{1-\rho}{2} \sum_k w_k (1 - f_{ik}) f_{jk} + c. \end{aligned} \tag{3}$$

The common features model corresponds to the extreme case $\rho = 1$, and the distinctive features model to the other extreme case $\rho = 0$.

However, this model is not the only way of striking a balance between common and distinctive features. Alternatively, we propose a new featural similarity model in which *each individual feature* is declared to be either a common feature (which increases the similarity of pairs of stimuli that share it) or a distinctive feature (which decreases the similarity of a pair of stimuli if one has it and the other does not). This Modified Contrast Model (MCM) is thus:

$$\begin{aligned} \hat{s}_{ij} &= \sum_{k \in CF} w_k f_{ik} f_{jk} - \frac{1}{2} \sum_{k \in DF} w_k f_{ik} (1 - f_{jk}) \\ &\quad - \frac{1}{2} \sum_{k \in DF} w_k (1 - f_{ik}) f_{jk} + c, \end{aligned} \tag{4}$$

where $k \in CF$ implies that the sum is taken over common features, and $k \in DF$ means that only distinctive features are considered.

Psychologically speaking, the argument is that a feature embodies some kind of regularity about the world, which may be that a set of stimuli all have something in common, or alternatively, that two groups of stimuli are in some way different from each other. A common feature instantiates the idea of "similarity within", whereas a distinctive feature represents the notion of "difference between". While it may be the case that the saliency of a feature can change, a commonality does *not* suddenly become a distinction, nor vice versa. In the MCM, the overall balance between commonality and distinctiveness emerges as a function of the relative number and saliency of common and distinctive features, rather than being specified by the parameter ρ, as it is in the TCM. That is, where the TCM assumes that common and distinctive features are weighted during the decision process, the MCM considers the commonality or distinctiveness of a feature to be a regularity inherent in the environment, and so embeds it in the representation itself. In this way, the MCM assumes that featural regularities can be either commonalities or distinctions, but never a bit of both. When a group of stimuli have both common and distinctive aspects, the MCM models these two aspects as two distinct featural regularities.

Model Fitting

It is useful to distinguish between the psychological problem of modeling featural similarity and the numerical problem of finding features (Shepard & Arabie, 1979). The psychological problem is: given a set of features $\mathbf{F}$, how should similarities be estimated? This is the question addressed by the four featural models discussed in the previous section. The numerical problem is a data fitting problem: given a set of similarity data $\mathbf{S}$, and assuming a particular psychological model, what set of features $\mathbf{F}$ most probably gave rise to the data? A variety of approaches have been adopted in fitting the additive clustering model, ranging from mathematical programming (Arabie & Carroll, 1980) to expectation maximization (Tenenbaum, 1996) and stochastic hillclimbing (Lee, in press). The process by which such algorithms operate is relevant to the psychological problem of similarity modeling only inasmuch as we require that they derive good answers to the numerical problem. While none of the above methods is perfect, it is fair to say that each approach performs well enough to allow interpretation and discussion of the derived representations. The representations derived here used a stochastic hillclimbing approach to fit the featural models similar to that adopted by Lee (in press) and Navarro and Lee (2001).

The fitting procedure adopted the Geometric Complexity Criterion (GCC: Myung, Balasubramanian, & Pitt, 2000) as the measure to be minimized by the successful representation. As has been remarked upon previously (e.g., Myung, 2000; Roberts & Pashler, 2000), achieving a good data-fit is not the sole criterion of a good model. Other considerations such as generalizability, simplicity and interpretability must be taken into account. From a quantitative standpoint, one can operationalize the trade-off between fit and complexity in a kind of formal version of Ockham's razor. The GCC is based on the notion that the complexity of a model is given by the number of distinguishable parametric distributions indexed by the model. Informally, this can be thought of as a measure of how many different

similarity matrices could be produced by a given feature structure under all possible choices of saliency weights. The more distributions a model indexes, the more complex it is. This measure is superior to the Akaike Information Criterion (Akaike, 1977) or the Bayesian Information Criterion (Schwarz, 1978), which estimate model complexity by counting the number of free parameters. As Lee (2001) has pointed out, the number of parameters is not a good indicator of the complexity of featural representations, since the way in which features are assigned to stimuli has a considerable influence on model complexity. Furthermore, Navarro and Lee (2001) have demonstrated that common features representations are more complex on average than distinctive features representations. These systematic differences in what Myung and Pitt (1997) call the functional form complexity of a model require a more discriminating measure such as the GCC. The derivation of GCC measures for the four featural models is straightforward, and follows the approach outlined by Navarro and Lee (2001).

Experiment

In order to provide an empirical test of the four featural similarity models, similarity data were collected for a set of 16 nations identified by name. The nature of this domain made it less than satisfactory to present people with a pair of countries and ask them to rate their similarity. It seems likely that this task would be ambiguous, in that the initial reaction of participants may be to ask, "Compared to what?" Even when the similarity between a pair of nations does not need a context, participants are unlikely to bring to this task a pre-existing numerical scale of nation-similarity upon which to rate it. An alternative approach is to provide participants with a context in which to make judgments. The task we used was to present people with a list of four countries, and ask them to select from the list the pair of nations most similar to one another.

Method

Participants Participants in the study were 16 university students (4 male, 12 female) aged 17 to 36, with a median age of 24, who took part in the experiment for course credit.

Materials The list of nations used was: China, Cuba, Germany, Indonesia, Iraq, Italy, Jamaica, Japan, Libya, Nigeria, the Philippines, Russia, Spain, United States, Vietnam, and Zimbabwe. They were chosen to suggest a variety of possible classification schemes (e.g., political system vs geographical location), and involve a variability in overall saliency (e.g., Italy and Germany were better known to most of our participants than Zimbabwe and Nigeria).

Procedure On each trial a list of four countries was displayed (via computer) to the participant, who was asked to pick out the two countries most similar to each other. The 16 nations yield $\binom{16}{2} = 120$ distinct pairs of nations, and a total of $\binom{16}{4} = 1820$ possible lists of four. Given that the similarity ratings are sensitive to all four presented stimuli, it was important to exhaust exactly the set of 1820 quadruples. To that end, the 1820 items were partitioned into 20 subsets of 91 quadruples. Most participants provided responses to one of these subsets, though a few of the participants provided responses to multiple subsets. Since each quadruple involves the presentation of 6 of the 120 nation-pairs, each pair appeared a total of $\frac{1820\times 6}{120} = 91$ times across the entire data set.

Results

Calculating the mean empirical similarity involved operationalizing the similarity of a pair of countries as the expected probability of selecting that pair in an arbitrary trial containing both stimuli. Using a standard result in Bayesian statistics (Gelman, Carlin, Stern, & Rubin, 1995, p. 31), if a particular pair is chosen k times out of n (n being 91), then the empirical similarity is given by $s_{ij} = \frac{k+1}{n+2}$. In using the GCC to control model complexity, it is important to know the precision of the similarity values (Lee, in press), which is basically a measure of the extent to which participants agreed in their judgments. Precision is important because more precise data justify more complex models. We estimated the precision to be moderate, by using the full distribution of the similarity judgments. Details of this estimation procedure, as well as that used to calculate similarity values, are given by Navarro (2002).

Using our stochastic hillclimbing algorithms, representations of the nations similarity data that minimized the GCC were found for each of the four similarity models. Of these four representations, the GCC values for the common features, distinctive features, and MCM representations are virtually indistinguishable, with the TCM performing slightly better. However, the qualitative characteristics of these representations are important in terms of model interpretability, and we discuss each in turn.

The best common features representation is shown in Figure 1, and contains seven features that explain 78.1% of the variance in the data. The features are highly interpretable, containing features for western European nations, Caribbean nations, south-

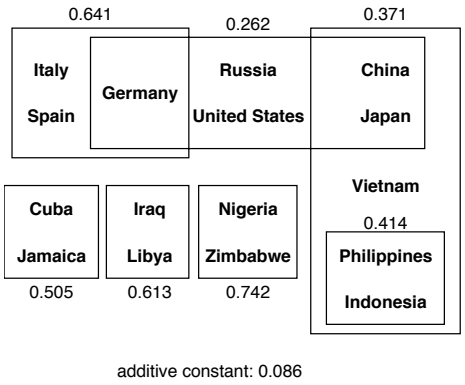

Figure 1: Common features representation of the nations similarity data, accounting for 78.1% of the variance (GCC=41.2). Common features are depicted by rectangles that encompass the nations that possess them.

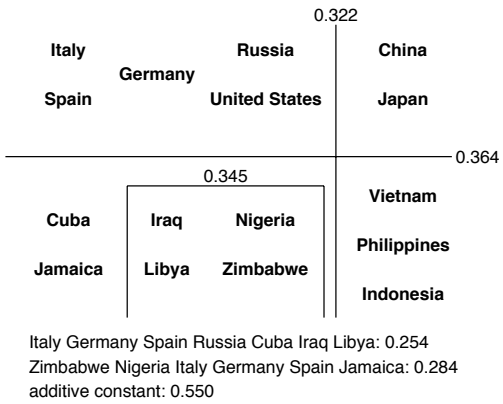

Figure 2: Distinctive features representation of the nations similarity data, accounting for 71.0% of the variance (GCC=41.5). Distinctive features are depicted by lines that partition the stimulus set.

ern African nations and Asian nations. It seems probable that the feature containing only the Philippines and Indonesia reflects perceived cultural and geographical similarities, and that the Iraq–Libya feature reflects the fact that both are Arabic nations and that both have been considered "rogue states". The final feature consists of Germany, Russia, United States, China and Japan, and could be said to denote the "world powers" among the stimulus set.

The best distinctive features representation is shown in Figure 2, and contains five features that explain 71.0% of the variance in the data. Three of the five features have a natural interpretation: one feature separates the African and Middle-Eastern nations from the rest of the world, and another separates the Asian nations from the others. The top-weighted feature in Figure 2 makes an intuitively plausible distinction that might be labeled "developed vs undeveloped". It is interesting to note that the placement of China within the developed nations is equivocal, since the GCC increases only marginally when China is placed in the other category. This makes sense, given China's status as a quickly developing nation. Significantly, however, the remaining two features do not appear to reflect any kind of interpretable structure. From a psychological standpoint, this is highly undesirable, since a central aim of representational modeling is to find interpretable structures in the data.

Table 1 displays the six feature representation derived using the TCM with $\rho = 0.7$, which explains 80.8% of the variance. Since the TCM specifies a balance of common and distinctive features, there is no simple way of depicting this representation graphically. The high ρ value indicates that commonalities are weighted more heavily than differences. All of the clusters in Table 1 appear in either the common features or distinctive features representations, which is not surprising. The features themselves make sense, although it is not easy to see exactly what $\rho = 0.7$ means when providing an overall interpretation. It is noteworthy that the distinctive feature that separated the developing from developed world in Figure 2 does not appear in this representation, despite being the most heavily weighted of the distinctive features. The reason for this may be that the feature does not make sense as anything other than a purely distinctive feature, since any common features component makes one half (either developed or developing) more salient than the other.

The seven feature MCM representation shown in Figure 3 explains 81.2% of the variance, and picks out a number of features from the common features representation: the western Europe, Caribbean, southern Africa and Asian features are all present, as is the "world powers" feature. The feature containing Cuba, Iraq and Libya is interesting, in that the inclusion of Cuba is a marginal case as judged by the GCC. With Cuba included, the feature has a "rogue states" interpretation, whereas without Cuba it would reflect the Arabic nations. Finally, the model also includes the "developed vs developing" regularity from the distinctive features representation. The comparison between this distinctive feature and the "world powers" common feature is also worth making. The fact that these two related but

Table 1: TCM representation of the nations similarity data, employing a common features bias ($\rho = 0.7$), accounting for 80.8% of the variance (GCC=36.9).

Stimuli in Cluster	Weight
Germany, Italy, Spain.	0.682
Nigeria, Zimbabwe.	0.495
China, Indonesia, Japan, Philippines, Vietnam.	0.453
Indonesia, Philippines.	0.374
China, Germany, Japan, Russia, United States.	0.316
Iraq, Libya, Nigeria, Zimbabwe.	0.288
additive constant	0.236

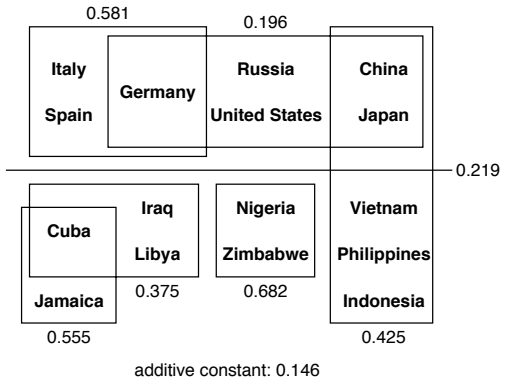

Figure 3: MCM representation of the similarity-condition countries data, accounting for 81.2% of the variance (GCC=41.0)

different features emerge in these different frameworks suggest that people make a distinction between the developed nations and the undeveloped nations, but that there is also something shared by the world powers of the stimulus set. Not all developed nations share this status, so it is appropriate that both features emerge. Moreover, the fact that two nations are developed does not necessarily imply that they are alike, but if one is developed and the other developing, it does make them different, so the "developed vs. developing" feature *should* be a distinctive feature. Correspondingly, two nations *are* alike if they are major world powers, but this does not say anything about their similarity if they are not.

Discussion

As previously noted, a major aim of featural similarity modeling is to capture simple, interpretable regularities present in the data. In this case, the common features representation is easily interpreted, yet the distinctive features representation is not. Furthermore, both the TCM and MCM representations evidence a bias towards commonalities. Overall, therefore, it appears that the participants' judgements were more heavily influenced by common features than distinctive features. Nonetheless, both the TCM and MCM representations include a distinctive features component, suggesting that distinctive features are not irrelevant in the data.

It is worth noting that the "developed vs developing" feature included in the distinctive features and MCM representations is the single most prominent regularity in the data set: when restricted to a single feature, all featural models except the common features model (which is incapable of expressing this feature) yield this feature. It accounts for more variance and has a substantially lower GCC than any other single feature. Though we do not wish to draw overly strong conclusions from a single experiment, the prominence of this regularity calls into question a central assumption of the TCM. As argued in the results section, this "developed vs developing" feature *only* makes sense as a purely distinctive feature. The TCM could only incorporate it by setting $\rho = 0$. However, as observed above, common features are more prominent than distinctive features in the data, and thus a high ρ value is preferred. In doing so, the TCM is able to give a good account for the data in a quantitative sense, but only by discarding this qualitatively important regularity. Notably, since the TCM assumes that all features are subject to the single decision variable ρ, it cannot simultaneously accommodate a commonality and a distinction under *any* parameterization. In comparison, the MCM is not merely capable of accommodating this phenomenon; it is typical behavior for the model.

Future Work

As noted above, a single data set provides only mild evidence: it is important to demonstrate that similar effects may be observed in other data sets. Though space does not permit further analyses here, work currently underway in this area seeks to generalize these findings in precisely this way. Further research could also extend the MCM in order to allow other types of distinctions. A distinctive feature in the current model partitions the stimulus set into two groups, such as "male" and "female". It could be argued that there are regularities that have the hallmarks of a distinction (in that differences-between matter, but similarities-within do not) but involve

more than two groups. For example, the notion that a thing is "animal", "mineral" or "vegetable" could be considered to be distinctive feature that partitions the stimuli into three groups. Moreover, there is a case to be made for representational formalisms that involve both discrete aspects (such as features) and continuous aspects (such as spatial dimensions). Accordingly, another avenue for research would be to pursue hybrid models that involve spatial as well as featural components.

Acknowledgements

This research was supported by Australian Research Council Grant DP0211406, and by a scholarship to DJN from the Australian Defence Science and Technology Organisation. We wish to thank several referees for helpful comments.

References

Akaike, H. (1977). On entropy maximization principle. In P. R. Krishnaiah (Ed.), *Applications of Statistics* (p. 27-41). Amsterdam: North-Holland.

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.

Clouse, D. S., & Cottrell, G. W. (1996). Discrete multi-dimensional scaling. In *The 18th Cognitive Science Conference* (p. 290-294). San Diego, CA.

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

Goldstone, R. L. (1999). Similarity. In R. Wilson & F. C. Keil (Eds.), *MIT encylopedia of the cognitive sciences* (p. 763-765). Cambridge, MA: MIT Press.

Goldstone, R. L., Medin, D. L., & Halberstadt, J. (1997). Similarity in context. *Memory and Cognition*, *25*(2), 237-255.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and Projects* (p. 437-447). Indianapolis: Bobbs-Merrill.

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

Lee, M. D. (in press). Generating additive clustering models with limited stochastic complexity. *Journal of Classification*.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190-204.

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 aproach. *Psychonomic Bulletin and Review*, *4*(1), 79-95.

Navarro, D. J. (2002). *Representing Stimulus Similarity.* Unpublished phd thesis, University of Adelaide.

Navarro, D. J., & Lee, M. D. (2001). Clustering using the contrast model. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, 686-691.

Restle, F. (1959). A metric and an ordering on sets. *Psychometrika*, *24*(3), 207-220.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358-367.

Rohde, D. L. T. (in press). Methods for binary multidimensional scaling. *Neural Computation*.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

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.* Cambridge, MA: MIT Press.

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