Common and Distinctive Features in Stimulus Similarity: A Modified Version of the Contrast Model

Daniel J. Navarro and Michael D. Lee*
Department of Psychology
University of Adelaide

Abstract

Featural representations of similarity data assume that people represent stimuli in terms of a set of discrete properties. We consider the differences in featural representations that arise from making four different assumptions about how similarity is measured. Three of these similarity models — the common features model, the distinctive features model, and Tversky's seminal contrast model — have been considered previously. The other model is new, and modifies the contrast model by assuming that each individual feature only ever acts as a common or distinctive feature. Each of the four models is tested on previously examined similarity data, relating to kinship terms, and on a new data set, relating to faces. In fitting the models, we use the Geometric Complexity Criterion to balance the competing demands of data-fit and model complexity. The results show that both common and distinctive features are important for stimulus representation, and we argue that the modified contrast model combines these two components in a more effective and interpretable way than Tversky's original formulation.

Introduction

A central problem for cognitive psychology is to understand the way people mentally represent stimuli. Models of mental representation are necessary building blocks for more general models of perception, cognition, decision making, and action. This means that different assumptions regarding the nature and form of mental representations lead to different constraints on formal models of cognitive processes, and makes it reasonable to argue that "pinning down mental representation is the route to rigor in psychology" (Pinker 1998, p. 85). Certainly, it is important that cognitive models use mental representations justified by the evidence provided by empirical data, and avoid the highly questionable practice of defining stimuli "by hand" on the basis of intuitive reasonableness (see Brooks 1991; Komatsu 1992; Lee 1998 for discussion).

One widely used approach for deriving stimulus representations from data is to base them on measures of stimulus similarity (e.g., Glushko 1975; Lee & Navarro 2002; Nosofsky 1986; Shepard & Arabie 1979; Tversky & Hutchinson 1986; see Shepard 1974, 1980 for overviews).

*Correspondence: Michael D. Lee, Department of Psychology, University of Adelaide, SA 5005, AUSTRALIA. Telephone: +61 8 8303 6096, Facsimile: +61 8 8303 3770, Electronic Mail: michael.lee@adelaide.edu.au

Following Shepard (1987), similarity is naturally understood as a measure of the degree to which the consequences of one stimulus generalize to another, and may be measured using a number of experimental methodologies, including ratings scales (e.g., Kruschke 1993), confusion probabilities (e.g., Shepard 1972), or grouping or sorting tasks (e.g., Rosenberg & Kim 1975).

Modeling the similarities between stimuli involves making assumptions about both the representational structures used to describe stimuli, and the processes used to assess the similarities across these structures. For example, under what Goldstone (1999) terms the 'dimensional' approach, stimuli are represented in terms of continuous values along a number of dimensions, so that each stimulus corresponds to a point in a multi-dimensional space, and the similarity between two stimuli is measured according to the distance between their representative points. Alternatively, under the 'featural' approach, stimuli are represented in terms of the presence or absence of a set of discrete (usually binary) features or properties, and the similarity between two stimuli is measured according to their common and distinctive features.

It is important to understand that the representational assumptions about describing stimuli can maintain a degree of independence from the processing assumptions about how similarity is measured. Within the dimensional approach, for example, the same representations may give rise to different patterns of similarity by using different distance metrics, as is often done (e.g., Potts, Melara, & Marks 1998; Nosofsky 1992; Shepard 1991) to capture the distinction between separable and integral stimulus domains. Similarly, within the featural approach, the same stimuli may be subjected to different similarity models by, for example, changing the relative degree of emphasis given to common and distinctive features.

This paper considers three established similarity models for featural representation: A purely common features model, a purely distinctive features model, and Tversky's (1977) seminal Contrast Model (TCM), which considers both common and distinctive features. We also develop a new similarity model, which is a modified version of the Contrast Model that combines common and distinctive features in a different way, and so we call it the Modified Contrast Model (MCM). By evaluating the four models on one previously studied set of similarity data, and one new data set, we show that the MCM has a number of advantages over the other models.

Four Featural Similarity Models

The following representational notation is used throughout this paper: For a stimulus domain with n stimuli and m features, a featural representation is given by the $n \times m$ matrix $\mathbf{F} = [f_{ik}]$, where $f_{ik} = 1$ if the i-th stimulus has the k-th feature, and $f_{ik} = 0$ if it does not. No restrictions are placed on the matrix $\mathbf{F}$, so that any stimulus can have any combination of features. Each feature also has an associated weight, denoted w_k for the k-th feature, which is a positive continuous number that can be interpreted as the saliency of the feature.

Common Features Model

The common features similarity model assumes that two stimuli become more similar as they share more features in common, and that the extent to which similarity increases is determined by the weight of each common feature. This means that the modeled similarity between the i-th and j-th stimulus, denoted as $\hat{s}_{ij}$, is simply the sum of the weights of the common features, as follows:

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

The positive constant c in Eq. (1) increases the similarity of each stimulus pair by the same amount, and so measures the degree to which all of the stimuli are similar to each other. It can be interpreted as the saliency weight of a 'universal' feature that is common to all stimuli.

The common features model has been the most widely used in cognitive representational modeling (e.g., Lee & Navarro 2002; Shepard & Arabie 1979; Tenenbaum 1996), probably because of the availability of statistical techniques known as additive clustering (e.g., Arabie & Carroll 1980; Chaturvedi & Carroll 1994; Lee 2002a; Mirkin 1987) that fit the model to similarity data.

Distinctive Features Model

The distinctive features approach assumes that two stimuli become more dissimilar to the extent that one stimulus has a feature that the other does not. As with the common features approach, the extent to which similarity is decreased by a distinctive feature is determined by the weight of that feature. This model can be expressed as:

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

One interpretation is that each pair of stimuli start with a high level of similarity, measured by the constant c, which is progressively decreased as they are found to have different features.

The distinctive features model is identical to Restle's (1959) symmetric distance metric, and to the similarity model used in discrete multidimensional scaling algorithms for developing featural representations (e.g., Clouse & Cottrell 1996; Lee 1998; Rohde 2002).

Tversky's Contrast Model (TCM)

The Contrast Model introduced by Tversky (1977) considers both common and distinctive features in measuring similarity, and allows for different degrees of emphasis to be placed on each. As specified by Tversky (1977, p. 332), the Contrast Model measures similarity as

$$\hat{s}_{ij} = \alpha F\left(I \cap J\right) - \beta F\left(I - J\right) - \gamma F\left(J - I\right), \tag{3}$$

where F is a monotonic function, $I \cap J$ are the features common to the i-th and j-th stimuli, $I - J$ are the features the i-th stimulus has but the j-th does not, and α, β, and γ are hyper-parameters that weight the relative contribution of each source of similiarity.

The use of hyper-parameters in the Contrast Model is not important mathematically, because their effects on modeling similarity could easily be absorbed within the loosely constrained function F. Rather, the form of Eq. (3) describes the important *psychological* assumptions on which the Contrast Model is based. According to Tversky (1977, p. 332)

> [t]he major constructs for the present theory are the contrast rule for the assessment of similarity, and the scale $[F(\cdot)]$, which reflects the salience or prominence of the various features.

In other words, Tversky's (1977) stated intent was that F should describe how features combine with one another, while the hyper-parameters establish the balance between common features and distinctive features in assessing similarity.

A number of empirical studies (e.g., Gati & Tversky 1984; Ritov, Gati, & Tversky 1990; Sattath & Tversky 1987) have demonstrated the ability of the Contrast Model to account for observed patterns of stimulus similarity, although this has usually not been achieved by fitting the model directly to data. This is probably because the very loose constraints on F do not specify exactly how the weights associated with common and distinctive features should be combined.

To make a concrete version of Tversky's (1977) Contrast Model, which we call the TCM, we assume that feature weights are combined through the process of addition. This is perhaps the

Table 1: The featural representation of Anna, Bridget, Charles, and Danielle.

Feature	Anna	Bridget	Charles	Danielle
Is Female	*	*		*
Has Twin	*			*
Plays Chess	*	*	*	
Races Cars	*	*		*
Votes				*
Wears Hats	*	*	*	*
Loves Cricket			*	*
⋮				
Reads Dickens	*	*		*

simplest function that meets Tversky's (1977) requirements, and is consistent with the previously studied common and distinctive features models described by Eqs. (1) and (2) above. Most importantly, it preserves the basic assumptions of the Contrast Model, because it uses global hyper-parameters to weight (or 'contrast') separate measures of common and distinctive feature saliency. We also make the simplifying assumption that similarity is symmetric, so that the similarity between the i-th and j-th stimuli is the same as the similarity between the j-th and i-th stimuli. This last assumption allows us to express the degree of emphasis given to common versus distinctive features using a single value, $0 \leq \rho \leq 1$, where large values give greater emphasis to common features, and smaller values give greater emphasis to distinctive features.

Using these assumptions, the TCM model measures stimulus similarity as follows:

$$\hat{s}_{ij} = c + \rho \sum_k w_k f_{ik} f_{jk} - \frac{1-\rho}{2} \sum_k w_k f_{ik} \left(1 - f_{jk}\right) - \frac{1-\rho}{2} \sum_k w_k \left(1 - f_{ik}\right) f_{jk}. \tag{4}$$

Implications of Global Weighting

Our modification of the TCM is motivated by some of the consequences of the global weighting ρ used to contrast common and distinctive features. If ρ does not take the extreme values 0 or 1, every feature possessed by a stimulus acts both to increase and decrease the similarities it has to other stimuli, depending on whether or not the second stimulus in the comparison has the feature. In this sense, every feature acts sometimes as a common feature, and sometimes as a distinctive feature. The TCM assumes that every feature gives exactly the same relative emphasis to these two roles, as quantified by the value of ρ. We believe this assumption is problematic. Some features seem to give the same emphasis to their common and distinctive roles. That is, the increase in similarity between stimuli sharing the feature is about the same as the decrease in similarity between stimuli distinguished by the feature. Other features seem to give greater emphasis to their common role than their distinctive role. These features greatly increase the similarity between stimuli sharing the feature, but have less effect in decreasing the similarity between distinguished stimuli. This pattern of effects is not well described by a fixed global weighting of the contrast between common and distinctive features in measuring stimulus similarity.

To demonstrate these concerns concretely, consider an example involving a group of four people called Anna, Bridget, Charles, and Danielle. Anna, Bridget and Danielle are female, while Charles is male. Anna and Danielle are identical twins. When thinking about the similarity between the four people, these facts are obviously important, and so "identical twin" and "is female" features are part of the representation. Other features, of course, will also be important in judging similarities, including hobbies, political views, and so on. Suppose, in this respect, that Bridget happens to have exactly the same hobbies, political views, and so on, as Anna. This situation is summarized by Table 1.

With these background assumptions, consider how the features affect the similarities between the four people. The fact that Anna and Danielle share the identical twin feature makes them much more similar. Anna and Danielle have an important and defining genetic relationship that would increase significantly how similar they feel to one another. The fact that Anna and Charles are distinguished by the identical twin feature has, we would argue, less impact. To see this, compare the similarity between Charles to Anna with the similarity between Charles and Bridget. There is the same pattern of match and mismatch between the hobbies, political views, and so on, for the two comparisons. On this basis, Charles would probably feel almost as much like Anna as he does like Bridget, even though neither he nor Bridget has an identical twin. Overall, therefore, the identical twin feature acts to increase the similarity of stimuli that share it by a large amount (call this increase c_1), but decreases the similarity between stimuli that are distinguished by a much smaller amount (call this decrease d_1). In other words, the identical twin feature gives greater emphasis to its common features than its distinctive features roles (i.e., $c_1 > d_1$).

Now consider how the female feature influences the same similarity comparisons. This is most easily done by imagining the people are only known as A, B, C and D, and then thinking about how the similarities change once their names and genders are known. The similarity between Anna and Danielle, who are both female, will increase, while the similarity between Charles and Anna, who are distinguished by gender, will decrease. We believe the size of the increase within genders (call this increase c_2) will be about the same as the decrease between genders (call this decrease d_2). Effectively, the female feature partitions the people into two groups who are similar to one another, but different from each other. Overall, therefore, the female feature gives about the same emphasis to its common and distinctive feature roles (i.e., $c_2 \approx d_2$).

To the extent that these intuitive arguments are accepted, our example mounts a case against the appropriateness of the TCM. The differences in emphasis on common and distinctive features cannot be accommodated by a global weight. Since it relies on the single parameter ρ to establish a balance between common and distinctive features, the TCM makes the assumption[1] that $c_1/d_1 = c_2/d_2$. However, our example has the relationship $c_1/d_1 \gg c_2/d_2$.

Psychologically, our general point is that all of the features in a domain may sometimes act as common features, and sometimes act as distinctive features, but that they will not all do so with equal degrees of emphasis. Admittedly, the choice of "identical twins" and "is female" features constitutes an extreme example. The identical twins feature emphasizes common feature effects over distinctive feature ones, while the female feature is far more balanced. For this reason, the example highlights the variation in relative emphasis possible for different features. In general, however, we believe that almost all features will give different degrees of emphasis to their common and distinctive roles. At the very least, it seems unlikely that the common-to-distinctive balance is the same for every feature in a domain, yet this is what the TCM implies.

[1] Algebraically, if the weights for the identical twins and female features are w_1 and w_2 respectively, the TCM requires $w_1\rho = c_1$, $w_1(1-\rho)/2 = d_1$, $w_2\rho = c_2$, and $w_2(1-\rho)/2 = d_2$, implying that $c_1/d_1 = c_2/d_2$.

Modified Contrast Model (MCM)

As an alternative to the global weighting assumed by the TCM, we propose a similarity model in which *each individual feature* is declared to be either a purely common feature (which increases the similarity of pairs of stimuli that share it) or a purely distinctive feature (which decreases the similarity of a pair of stimuli if one has it and the other does not). This means that our Modified Contrast Model (MCM) measures similarity by examining the two subsets of declared common features and declared distinctive features separately. The weights of all of the declared common features that are shared by two stimuli are added together, and the weights of all of the declared distinctive features that distinguish between two stimuli are added together. The final similarity measure is then found by subtracting the distinctive features total from the common features total, as follows:

$$\hat{s}_{ij} = c + \sum_{k \in CF} w_k f_{ik} f_{jk} - \frac{1}{2} \sum_{k \in DF} w_k f_{ik} \left(1 - f_{jk}\right) - \frac{1}{2} \sum_{k \in DF} w_k \left(1 - f_{ik}\right) f_{jk}, \tag{5}$$

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

So, for example, in Table 1, suppose that the identical twin feature is declared to be a common feature, and the female feature is declared to be a distinctive feature. Then the identical twin feature would increase the similarity of Anna to Danielle, but would not decrease the similarity between Anna and Bridget. By contrast, the female feature would decrease the similarity between Anna and Charles, but would not increase the similarity of Anna to Bridget.

In what way does the MCM differ from the TCM? From the general perspective of modeling mental representation, a feature should embody some kind of regularity about the world. This 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 captures the idea of 'similarity within', whereas a distinctive feature captures the notion of 'difference between'. When a group of stimuli have both common and distinctive aspects, the MCM treats them as distinct regularities, introducing common features and distinctive features as appropriate. In this way, within 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 imposed by the value of ρ, as it is in the TCM. In this sense, a natural analogue of ρ for the MCM is the proportion of common feature weights in a representation, given by $\widetilde{\rho} = \sum_{k \in CF} w_k / \left(\sum_{k \in CF} w_k + \sum_{k \in DF} w_k\right)$.

Another way of understanding the difference between the MCM and the TCM is to note that the TCM places different degrees of emphasis on common and distinctive features only after the relative strength of each has been assessed. This can be seen algebraically in Eq. (4), where the ρ and $(1 - \rho)$ values are applied after the sum of weights for the representational features have been formed. This two stage process is naturally interpreted as having a representational component, where the measures of stimulus similarity implied by the featural representations are calculated, and a decision component, where the two sources of stimulus similarity are combined. The MCM, in contrast, blends the distinction between common and distinctive features into the way in which feature weights are originally assessed, and so involves only a one stage process that is entirely based on the featural representations themselves. By declaring features to be either common or distinctive, the MCM embeds this information in the way stimuli are represented.

Relationships Between the Models

In the context of evaluating the four featural models of stimulus representation, there are some important relationships between them that need to be considered. Most obviously, the TCM reduces to the common features model when $\rho = 1$, and to the distinctive features model when

$\rho = 0$. The MCM can also reduce to the purely common or distinctive feature models by specifying only common or distinctive features.

More subtly, it is possible for both the common features model and the TCM to mimic the MCM if strong additional assumptions are made. If a common features model (a) has the same common features as a MCM, (b) has both the feature and its complement for each MCM distinctive feature, and (c) it the weights of each of these complementary pairs are fixed to be identical, then the similarity estimates of the common features model are indistinguishable from those of the MCM. For the TCM to mimic the MCM, besides these three conditions, it also needs to be assumed that $\rho = 1$, so that the TCM is first reduces to a common features model.

The assumptions needed for the common features and TCM to behave in the same way as the MCM make it clear that there are important theoretical differences between the models. They are built to support different types of representations, with different psychological interpretations. The possibility of formal equivalence under special circumstances, however, means that some sophistication is required to evaluate the models. In particular, the model evaluation methodology needs to be able to account for the possibility that, through complicated and implausible assumptions, the common features and TCM models could accomplish a pattern of similarity estimates that the MCM can do in a natural and straightforward way.

Evaluating the Models

Accounting for Model Complexity

Good models in psychology, as in other empirical sciences, should aim for at least three goals. They should provide good descriptions of the available data (accuracy), they should allow good predictions for new or different situations where data are not available (generalization), and they should convey meaning or offer substantive insight into the phenomena being investigated (explanation). Often these three goals conflict, and it is a basic challenge of modeling to strike the right balance between them.

Models that specify featural representations for similarity data certainly require a balance between accuracy, generalizability, and explanatory power. In particular, they need to be wary of achieving high levels of accuracy at the expense of generality and explanation. As noted by Shepard and Arabie (1979, p. 98), a common features representation able to specify arbitrary features, and freely manipulate its weights, can fit any similarity data perfectly[2]. Consider a representation that has a feature for each pair of stimuli, and sets the weight of each feature to the empirical similarity of its stimulus pair. Although this model has perfect accuracy, its featural stimulus representations is very unlikely to make good predictions about new similarity data. The representation also will not convey any new meaning, since it is effectively just a re-statement of the original data.

Scientific modeling often addresses the competing requirements of data-fit, generalizability, and explanatory power using the principle known as 'Ockham's Razor', which (loosely speaking) argues that the simplest sufficiently accurate model should be chosen. By seeking simple models that provide a reasonable account of the data, only the 'essence' of the information in the data is modeled, and so the model is applicable in new situations, and is amenable to interpretation. It is through the application of this principle that the potential problems of model mimicry can be overcome, because it is sensitive to the complexity involved in tuning a much more general model to behave like a more specific one, when the data require only the simpler account.

In practice, Ockham's Razor requires that models are developed by simultaneously maximizing a measure of data-fit and minimizing a measure of model complexity. Measuring the

[2]The same observation obviously applies to the TCM and MCM, since they have special cases that reduce to the common features model. The distinctive features model does not have the same property, since it is constrained by the metric axiom known as the triangle inequality.

accuracy of a featural representational model is conceptually straightforward. It is determined by how closely the modeled similarities approximate the observed empirical similarities. Measuring the complexity of a featural representational model is more difficult. We have previously (Lee 2001b, 2002a; Navarro & Lee 2001) considered a number of different types of measures, including the Akaike Information Criterion (AIC: Akaike 1974), the Bayesian Information Criterion (BIC: Schwarz 1978), the Stochastic Complexity Criterion (SCC: Rissanen 1996), and the Geometric Complexity Criterion (GCC: Myung, Balasubramanian, & Pitt 2000)[3]. While the AIC and BIC are conceptually and computationally the most straightforward, they are sensitive only to what Myung and Pitt (1997) term the 'parametric complexity' of the models, and so assess complexity solely according to the number of features that a representation uses. The SCC and GCC, in contrast, are also sensitive to 'functional form' complexity, which is determined by the exact way in which features are assigned to stimuli. As demonstrated by Lee (2002a), this is important, since it is possible for two different featural representation to have the same number of features and fit the data equally well, so that only complexity measures such as the SCC and GCC are able to select the simpler (and more interpretable) model.

For any of these complexity measures, it is important to consider how the precision of the empirical data influences the appropriate balance between data-fit and model complexity. Precision is basically a measure of the confidence that can be placed in empirically gathered similarity values reflecting the "true" underlying similarity of the stimuli. For example, if similarity data are collected by averaging across a large number of people, who all agree closely in their assessments, it is reasonable to be confident about the average values, and regard the data as being precise. As the level of disagreement increases, however, the data should be regarded as being less precise. When data are precise, the introduction of additional complexity into a model to achieve a greater level of descriptive accuracy may well be warranted. When data are imprecise, however, the same increase in complexity will not be warranted. This means that exactly the same observed averaged data values may provide different levels of evidence for competing models depending upon their level of precision. As argued by Lee (2001a, p. 155), it follows that a quantitative estimate of data precision is needed to determine the appropriate balance between data-fit and model complexity. The data sets we consider were obtained by averaging across a number of similarity matrices, associated with different subjects or groups of subjects, and so the approach for estimating precision described by Lee (2001a, 2002a) is applicable. Basically, this approach estimates precision as the average variation in similarity judgments across all subjects and all pairs of stimuli. In addition, this approach to estimating precision has been demonstrated in a closely related context to overcome many of the problems inherent in averaging across data that may involve significant individual differences (Lee & Pope 2003).

Geometric Complexity Criteria for the Models

The GCC evaluates probabilistic models by taking into account both their data-fit and complexity. The data-fit is measured by the maximum log likelihood of the model, $-\ln p\left(D \mid \theta^*\right)$, where $p\left(\cdot\right)$ is the likelihood function, D is a data sample of size N, and θ is a vector of the k model parameters which take their maximum likelihood values at θ^*. The complexity of the model is measured in terms of the number of distinguishable data distributions that the model indexes through parametric variation. The innovative geometric approach developed by Myung, Balasubramanian and Pitt (2000) leads to the following four term expression:

[3]Readers familiar with the model selection literature may notice a slight terminological ambiguity here. To be precise: we use SCC to denote Rissanen's (1996) approximation to the normalized maximum likelihood codelength, and GCC to refer to Balasubramanian's (1997) geometric criterion, where terms smaller than $O(1/N)$ are ignored.

$$\mathrm{GCC} = -\ln p\left(D \mid \theta^*\right) + \frac{k}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int d\theta\sqrt{\det \mathbf{I}\left(\theta\right)} + \frac{1}{2}\ln\left(\frac{\det \mathbf{J}\left(\theta^*\right)}{\det \mathbf{I}\left(\theta^*\right)}\right),$$

where

$$\mathbf{I}_{ij}\left(\theta\right) = -E_\theta\left[\frac{\partial^2 \ln p\left(D \mid \theta\right)}{\partial\theta_i\partial\theta_j}\right]$$

is the expected Fisher Information Matrix of the model parameters, and

$$\mathbf{J}_{ij}\left(\theta^*\right) = -\left[\frac{\partial^2 \ln p\left(D \mid \theta\right)}{\partial\theta_i\partial\theta_j}\right]_{\theta=\theta^*}$$

is the observed Fisher Information Matrix (Schervish 1995, p. 226).

We follow Tenenbaum (1996) in treating each of the empirical similarities as coming from a Gaussian distribution with common variance. For the i-th and j-th stimuli, the mean of this Gaussian is given by the similarity value s_{ij}. The common variance measures the precision of the data, and has a sample estimate $\hat{\sigma}^2$.

Under the Gaussian formulation, the probability of similarity data $\mathbf{S}$ arising for a particular featural representation $\mathbf{F}$, using a particular weight parameterization $\mathbf{w}$, is given by

$$\begin{aligned} p\left(\mathbf{S} \mid \mathbf{F}, \mathbf{w}\right) &= \prod_{i<j}\frac{1}{\left(\hat{\sigma}\sqrt{2\pi}\right)}\exp\left(-\frac{\left(s_{ij}-\hat{s}_{ij}\right)^2}{2\hat{\sigma}^2}\right) \\ &= \frac{1}{\left(\hat{\sigma}\sqrt{2\pi}\right)^{n(n-1)/2}}\exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i<j}\left(s_{ij}-\hat{s}_{ij}\right)^2\right), \end{aligned}$$

and so the log-likelihood takes the familiar form of the sum of squared differences between the empirical data and model predictions, as scaled by the estimated precision of the data. The first term of the GCC, which measures data-fit, maximizes this log-likelihood (by minimizing the negative log-likelihood), using the best fitting modeled similarities $\hat{s}^*_{ij}$, as follows:

$$-\ln p\left(\mathbf{S} \mid \mathbf{F}, \mathbf{w}^*\right) = \frac{1}{2\hat{\sigma}^2}\sum_{i<j}(s_{ij}-\hat{s}^*_{ij})^2 + \text{constant}. \tag{6}$$

The second term of the GCC for a featural representation with m features may be found by noting that it uses $m+1$ parameters (including the additive constant), and that the N data being modeled are the $n\left(n-1\right)/2$ unique observations in a symmetric $n \times n$ similarity matrix excluding self-similarities, giving

$$\frac{m+1}{2}\ln\left(\frac{n(n-1)}{4\pi}\right). \tag{7}$$

For the common features, distinctive features, and MCM, using the similarity models given in Eqs. (1), (2), and (5), the calculation of the second-order partial derivatives

$$\frac{\partial^2 \ln p\left(\mathbf{S} \mid \mathbf{F}, \mathbf{w}\right)}{\partial w_x \partial w_y}$$

is straightforward[4], and allows the expected and observed Fisher Information Matrices $\mathbf{I}\left(\mathbf{w}\right)$ and $\mathbf{J}\left(\mathbf{w}\right)$ to be specified. As it turns out, these two matrices are identical for any given featural

[4]See Lee (2001b) and Navarro (2003) for a detailed derivation and interpretation of the common features case.

representation, and so the fourth term of the GCC vanishes. This means that, for these models, the GCC is identical to the Stochastic Complexity measure, based on the Minimum Description Length approach to model selection, developed by Rissanen (1996).

In fact, the two matrices assume a constant value that is independent of the weight parameters, and is determined entirely by the feature structure $\mathbf{F}$, which also simplifies the third term of the GCC. This constant value is conveniently written as the determinant of an $(m+1)\times(m+1)$ "complexity matrix", $\mathbf{G} = [g_{xy}]$, defined as

$$g_{xy} = \sum_{i<j} e_{ijx} e_{ijy},$$

where $e_{ijk} = f_{ik} f_{jk}$ for the common features model, $e_{ijk} = -\frac{1}{2} f_{ik}(1-f_{jk}) - \frac{1}{2}(1-f_{ik}) f_{jk}$ for the distinctive features model, and $e_{ijk} = f_{ik} f_{jk}$ for the MCM when the k-th feature is a declared common feature, and $e_{ijk} = -\frac{1}{2} f_{ik}(1-f_{jk}) - \frac{1}{2}(1-f_{ik}) f_{jk}$ for the MCM when the k-th feature is a declared distinctive feature.

Using the complexity matrix, and assuming that a reasonable range for each of the weight parameters is over the interval $[0,1]$, the third term of the GCC is given by:

$$\begin{aligned} \ln \int \sqrt{\det \mathbf{I}(\mathbf{w})}\, d\mathbf{w} &= \ln \int_0^1 \int_0^1 \ldots \int_0^1 \sqrt{\det\left(\frac{1}{\hat{\sigma}^2}\mathbf{G}\right)}.dw_1.dw_2 \ldots dw_{m+1} \\ &= \frac{1}{2}\ln\det\mathbf{G} - \frac{m+1}{2}\ln\hat{\sigma}^2. \end{aligned} \tag{8}$$

Putting together the results in Eqs. (6), (7) and (8), the GCC for the common features, distinctive features and MCM representations may be finally given as

$$\mathrm{GCC} = \frac{1}{2s^2}\sum_{i<j}(s_{ij} - \hat{s}^*_{ij})^2 + \frac{m+1}{2}\ln\left(\frac{n(n-1)}{4\pi\hat{\sigma}^2}\right) + \frac{1}{2}\ln\det\mathbf{G} + \text{constant}.$$

The TCM, because of its additional ρ parameter, does not have a correspondingly simple analytic result. Technically, the TCM defines a family of models, indexed by different ρ values. This means that a TCM representation is comparable to an MCM, common features, or distinctive features representation only once a ρ value is chosen. Because of these circumstances, finding the complexity matrix $\mathbf{G}$ for the TCM is more easily achieved using an alternative definition for the expected Fisher Information Matrix based on covariances (Schervish 1995, p. 111), to obtain

$$\mathbf{G}(\mathbf{w},\rho) = \begin{bmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b^T} & c \end{bmatrix},$$

where $\mathbf{A}$ is an $(m+1)\times(m+1)$ matrix with entries

$$a_{xy} = \sum_{i<j}\left(f_{ix}f_{jx} - \frac{1-\rho}{2}(f_{ix}+f_{jx})\right)\left(f_{iy}f_{jy} - \frac{1-\rho}{2}(f_{iy}+f_{jy})\right),$$

$\mathbf{b}$ is an $(m+1)$-dimensional column vector with entries

$$b_x = \frac{1}{2}\sum_{i<j}\left(f_{ix}f_{jx} - \frac{1-\rho}{2}(f_{ix}+f_{jx})\right)\sum_k w_k(f_{ik}+f_{jk}),$$

and

$$c = \frac{1}{4}\sum_{i<j}\left(\sum_k w_k(f_{ik}+f_{jk})\right)^2.$$

From this result, we used the simple Monte Carlo estimate

$$\int \det \mathbf{G}\left(\mathbf{w}, \rho\right) d\mathbf{w} d\rho \approx \frac{1}{N} \sum_{i=1}^{N} \det \mathbf{G}\left(\mathbf{w}_i, \rho_i\right),$$

where $N = 10^6$ parameter combinations were chosen by selecting each parameter value independently from the uniform distribution on $[0, 1]$. Despite the moderately high dimensionality of these integrals for the stimulus domain examined later in this paper (nine for the kinship terms and five for the faces), repeated runs indicated that accurate convergence was being achieved, and so more sophisticated Monte Carlo techniques like importance sampling or Gibbs sampling (e.g., Gilks, Richardson, & Spiegelhalter 1996) were not required. The final term of the GCC for the TCM, involving the evaluation of the covariance matrix at the maximum likelihood parameterization, was also found by standard numerical methods.

Fitting Algorithm

It has often been noted that fitting featural representations to similarity data is a difficult combinatorial optimization problem (e.g., Shepard & Arabie 1979; Tenenbaum 1996), because of the discrete (binary) nature of the feature membership variables f_{ik}. Additive clustering algorithms, which fit the common features model, have used a number of general optimization approaches, including mathematical programming (Arabie & Carroll 1980; Chaturvedi & Carroll 1994), qualitative factor analysis (Mirkin 1987), probabilistic expectation maximization (Tenenbaum 1996), genetic algorithms (Lee 2002b), and stochastic hillclimbing (Lee 2002a). Despite the difficulty of the problem, most of these approaches have been successful enough to find useful representations, and there does not seem to be any compelling reason to prefer one approach over another.In fitting the models reported in this paper, we used the stochastic hillclimbing algorithm developed and evaluated by Lee (2002a). The basic idea is to start with a representation that has only the universal feature, and then to add features successively until the GCC measure indicates that the increase in complexity is no longer warranted by the improvement in data-fit.

Evaluation on Kinship Data

Data

As a first comparison of the four similarity models, we fitted them to a previously analyzed data set. These data measured the similarities between 15 common kinship terms, such as 'father', 'daughter', and 'grandmother', as collected by Rosenberg and Kim (1975) and published in Arabie, Carroll and DeSarbo (1987, pp. 62–63). The similarity values were based on a sorting procedure performed by six groups of 85 participants, where each kinship term was placed into one of a number of groups, under various conditions of instructions to the participants. We considered a slightly modified version of this data set that excluded the 'cousin' stimulus. This was done because we were interested in examining how the different similarity models dealt with the concept of gender, and 'cousin' was the only ambiguous term in this regard.

Representations

The featural representations generated using the common features model, distinctive features model, TCM, and MCM, are detailed in Tables 2, 3, 4, and 5 respectively. Each feature corresponds to a row, where the stimuli that have the feature are listed, and the weight of the feature is given. The weight of the universal feature shared by all stimuli is also given.

Table 2: Common features representation of the kinship stimuli, explaining 93.1% of the variance with a GCC value of 59.6.

Stimuli with Feature	Weight
Brother, Sister	0.305
Father, Mother	0.290
Granddaughter, Grandfather, Grandmother, Grandson	0.288
Aunt, Uncle	0.286
Nephew, Niece	0.283
Aunt, Daughter, Granddaughter, Grandmother, Mother, Niece, Sister	0.223
Brother, Father, Grandfather, Grandson, Nephew, Son, Uncle	0.221
Aunt, Nephew, Niece, Uncle	0.219
Brother, Daughter, Father, Mother, Sister, Son	0.193
Daughter, Granddaughter, Grandson, Son	0.128
Universal Feature	0.226

The common features representation explains 93.1% of the variance in the data, and has a GCC value of 59.6. It includes features describing the concepts siblings (brother, sister), parents (father, mother), and nuclear family (brother, daughter, father, mother, sister, son). The only clear weakness of the representation is that it has separate features for the concepts male (brother, father, grandfather, grandson, nephew, son, uncle) and female (aunt, daughter, granddaughter, grandmother, mother, niece, sister). These features have almost identical weights, which means that the representation is effectively treating the difference between male and female as a distinctive feature that separates the stimuli into two classes.

The distinctive features representation explains 94.7% of the variance in the data, and has a GCC value of 50.5. It does capture the concept of gender using the first feature listed in Table 3, where the seven male stimuli are explicitly listed. Many of the other features in the representation, however, are far less interpretable. It is difficult, for example, to understand what the feature that distinguishes between (brother, granddaughter, grandson, sister) and the remaining terms means.

The TCM representation shown in Table 4 explains 91.4% of the variance, and has a GCC value of 68.3. It uses a ρ value of 0.2, and so allows a balance between common and distinctive features, but still has shortcomings. It has a feature for the concept of nuclear family, but does not have one for the concepts of parents or siblings. In addition, it uses two features to capture to concept of gender, in the same way as the common features model. These deficiencies can be explained in terms of the ρ value. Because $\rho = 0.2$ emphasizes the role of distinctive features over common features in explaining stimulus similarity, concepts like parents and siblings that rely on a common features interpretation are not found. Because $\rho > 0$, however, it is not possible to have a purely distinctive feature that gives equal weight to both the male and female parts of the gender concept, and so each part must be represented explicitly by its own feature. Of course, by using $\rho = 0$, the TCM would reduce to the distinctive features representation, and also explain 94.7% of the variance in the data. However, the GCC for this TCM is about 71.7 rather than 50.5, because of the different complexity of the two models. In particular, recall that

Table 3: Distinctive features representation of the kinship stimuli, explaining 94.7% of the variance with a GCC value of 50.5.

Stimuli with Feature	Weight
Brother, Father, Grandfather, Grandson, Nephew, Son, Uncle	0.451
Aunt, Brother, Nephew, Niece, Sister, Uncle	0.249
Aunt, Brother, Daughter, Father, Mother, Sister, Son, Uncle	0.242
Aunt, Granddaughter, Grandfather, Grandmother, Grandson, Uncle	0.238
Aunt, Daughter, Granddaughter, Grandson, Nephew, Niece, Son, Uncle	0.213
Aunt, Father, Mother, Nephew, Niece, Uncle	0.203
Brother, Daughter, Father, Granddaughter, Grandson, Mother, Sister, Son	0.164
Brother, Granddaughter, Grandson, Sister	0.091
Universal Feature	0.902

the complexity of the TCM, unlike the distinctive features model, involves integrating across all of the the possible values of the parameter ρ, and so it is possible that other feature structures $\mathbf{F}$ could lead to lower GCC values. It is the minimality of the GCC value for the $\rho = 0.2$ representation that makes it preferred, despite explaining less of the variance in the data.

The MCM representation shown in Table 5 explains 93.5% of the variance in the data, and has a GCC value of 56.1. It includes easily interpreted common and distinctive features, with the ratio of common feature weights to the total being $\widetilde{\rho} = 0.6$. It has four distinctive features, dividing males from females, once removed terms (aunt, nephew, niece uncle) from those not once removed, extreme generations (granddaughter, grandfather, grandmother, grandson) from middle generations, and the nuclear family (brother, daughter, father, mother, sister, son) from the extended family. It also has six common features, which generally capture meaningful subsets within the broad distinctions, such as parents, siblings, grandparents and grandchildren. These concepts are appropriately designated as common features since, for example, a brother and sister have the similarity of being siblings, but this does not make those who are not siblings, like an aunt and a grandson, more similar.

Discussion

The GCC values for the models measures of their ability to maximize data-fit and minimize complexity, and so quantify their performance on the accuracy and generalizability criteria. As Myung, Balasubramanian and Pitt (2000) explain, the difference between the GCC values for two models are interpretable on a log-odds scale, with smaller GCC values corresponding to more likely models. Under the broad interpretive framework suggested by Kass and Raftery (1995, p. 777), differences less than two are "not worth more than a bare mention", differences between two and six are "positive", differences between six and ten are "strong", and differences greater than ten are "very strong". These guidelines are not, of course, intended to be prescriptive, because odds have an inherent meaning defined by betting, and so do not need calibration for interpretation. Rather, the Kass and Raftery (1995) guidelines give useful suggested standards for scientific evidence in choosing between models.

Using these guidelines, there is some indication that the distinctive features representation

Table 4: TCM representation of the kinship stimuli, using $\rho = 0.2$, and explaining 91.4% of the variance with a GCC value of 65.5.

Stimuli with Feature	Weight
Brother, Daughter, Father, Mother, Sister, Son	0.392
Brother, Granddaughter, Grandfather, Grandmother, Grandson, Sister	0.250
Daughter, Father, Granddaughter, Grandfather, Grandmother, Grandson, Mother, Son	0.220
Aunt, Daughter, Granddaughter, Grandmother, Mother, Niece, Sister	0.219
Brother, Father, Grandfather, Grandson, Nephew, Son, Uncle	0.213
Brother, Daughter, Granddaughter, Grandson, Nephew, Niece, Sister, Son	0.168
Aunt, Brother, Father, Mother, Nephew, Niece, Sister, Uncle	0.123
Universal Feature	0.679

Table 5: MCM representation of the kinship stimuli, explaining 93.5% of the variance with a GCC value of 56.1 and $\widetilde{\rho} = 0.6$. Each feature is declared to be a common feature (CF) or a distinctive feature (DF).

Stimuli with Feature	Weight
DF: Brother, Father, Grandfather, Grandson, Nephew, Son, Uncle	0.452
CF: Aunt, Uncle	0.298
CF: Nephew, Niece	0.294
CF: Brother, Sister	0.291
CF: Grandfather, Grandmother	0.281
CF: Father, Mother	0.276
CF: Granddaughter, Grandson	0.274
DF: Aunt, Nephew, Niece, Uncle	0.230
DF: Granddaughter, Grandfather, Grandmother, Grandson	0.190
DF: Brother, Daughter, Father, Mother, Sister, Son	0.187
Universal Feature	0.660

is more likely than the MCM representation which, in turn, is marginally more likely than the common features representation. The GCC values also indicate that the data provide very little evidence for the TCM representation. While these results are suggestive, they only address the quantifiable criteria of accuracy and generalizability. To provide a complete evaluation of the merits of the four models, other modeling criteria must be considered. As Pitt, Myung and Zhang (2002, p. 486) acknowledge, measures like the GCC "contribute only a few pieces of evidence to the model selection process" and so should be used in conjunction with other qualitative criteria like "explanatory adequacy".

The four representations show clear differences in terms of interpretability. The common features representation is easily interpreted, but does not explicitly represent the male versus female regularity as a distinctive feature. The TCM representation also fails to represent the gender distinction in a single feature, and does not include some useful common features. The distinctive features model captures gender, but many of its remaining features are difficult to interpret. The MCM representation combines the best of the common features and distinctive features approaches, providing an easily interpreted representation.

Since this interpretative analysis is not amenable to quantification, it is not possible to weight the superior GCC of the distinctive features account against the superior interpretability of the MCM account. Indeed, it probably does not make sense to try to achieve this unification. Under different circumstances, the relative importance of accuracy, generalizability and explanation as modeling goals will vary, and so model selection involves subjective decision making conditioned on the evidence provide by both measures like the GCC and interpretive analysis[5]. We would argue, however, that the problem of learning stimulus representations from similarity data places a heavy emphasis on interpretability, explanation and understanding, and so there are grounds to argue the MCM representation should be preferred.

Evaluation on Faces Data

As a second evaluation of the four models, we considered the similarities between cartoon face stimuli, collected by pairwise ratings. These choices were intended to help achieve generality and robustness in the overall evaluation of the models, by using perceptual rather than conceptual stimuli, and using a different methodology for collecting the similarity data.

Data

Subjects Subjects in the study were 10 university students (six female, four male) aged 24 to 49, with a median age of 26.

Stimulus Domain We designed ten cartoon face stimuli, labeled face 'a' through face 'j', varying in their hairstyle (male or female), hair color (black, brown, burgundy, grey or bright blue), shape of glasses (round or square), and color of glasses (dark blue or pink). Table 6 gives a verbal description of each of the faces to supplement the black-and-white presentations in Figures 1, 2 and 3.

Procedure Participants were shown (via computer) all $\binom{10}{2} = 45$ pairs of faces in a random order, and asked to rate the similarity of each pair on a seven-point scale, ranging from "completely different" (1) to "completely identical" (7). The similarity of each pair of faces, arithmetically averaged across all participants, and normalized to lie in the unit interval, is given in Table 7.

[5]It is important to realize that "subjective" is not synonymous with "arbitrary". Subjective decisions follow from degrees of belief conditioned on evidence, and are part of rational decision making. All probabilities, for example, are subjective in this sense (Cheeseman 1999). Arbitrary decisions, on the other hand, are those that are all as good (or bad) as each other, and do not provide a rational basis for decision making.

Table 6: Verbal description of the face stimuli.

Face a: Female with burgundy hair and round pink sunglasses
Face b: Female with brown hair and round blue sunglasses
Face c: Female with black hair and round blue sunglasses
Face d: Female with burgundy hair and square blue sunglasses
Face e: Female with blue hair and square blue sunglasses
Face f: Male with grey hair and round pink sunglasses
Face g: Male with black hair and round blue sunglasses
Face h: Male with brown hair and square blue sunglasses
Face i: Male with grey hair and square blue sunglasses
Face j: Male with blue hair and square blue sunglasses

Table 7: Similarity data for the face stimuli.

	a	b	c	d	e	f	g	h	i	j
a	-									
b	0.70	-								
c	0.64	0.83	-							
d	0.70	0.70	0.63	-						
e	0.51	0.53	0.60	0.71	-					
f	0.54	0.36	0.43	0.30	0.31	-				
g	0.43	0.54	0.60	0.40	0.36	0.67	-			
h	0.37	0.63	0.41	0.56	0.50	0.53	0.66	-		
i	0.30	0.40	0.40	0.53	0.47	0.67	0.67	0.83	-	
j	0.26	0.33	0.37	0.44	0.61	0.47	0.59	0.74	0.74	-

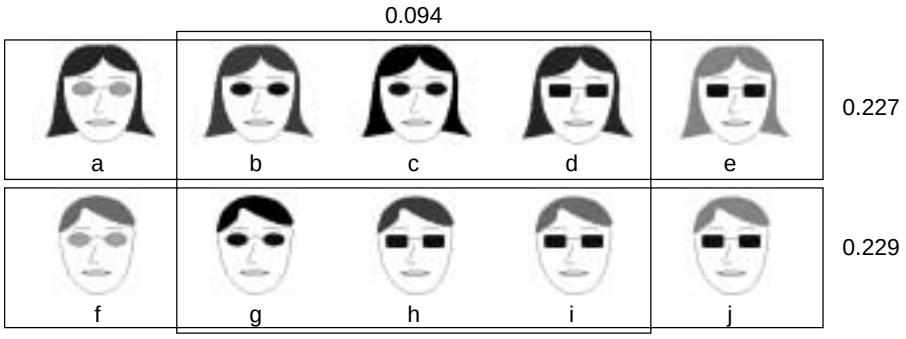

Figure 1: The common features representation of the faces data, explaining 63.9% of the variance with a GCC value of 23.3, with a universal feature weight of 0.40. The weights of the other three features are shown.

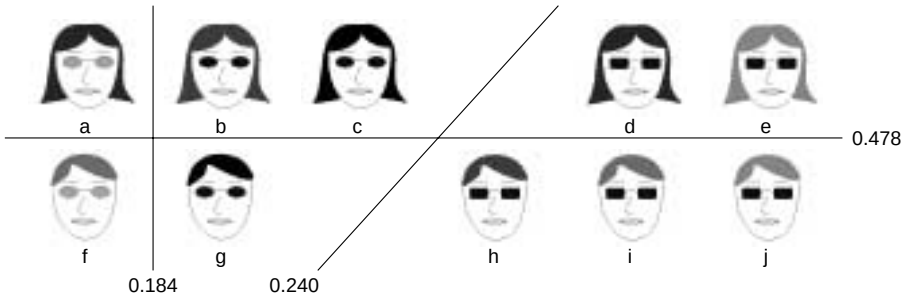

Figure 2: The distinctive features and TCM representation of the faces data, explaining 82.1% of the variance with a GCC value of 18.1 for the distinctive features model, and 25.7 for the TCM. The representation has a universal feature weight of 0.77, and the weights of the other three features are shown.

Representations

The common features representation, shown in Figure 1, explains 63.9% of the variance and has a GCC value of 23.3. It uses only three features, two of which correspond to the concepts male and female. The third feature corresponds to something like the concept 'unremarkable', since it includes all of those people with 'conservative' hair colors and sunglass designs.

The distinctive features representation is shown in Figure 2, explains 82.1% of the variance with a GCC value of 18.1 and also includes three features. One of these divides the males from the females, another divides people with square sunglasses from those with round sunglasses, and the third divides people with pink sunglasses from the remaining people.

As it turns out, the TCM representation used $\rho = 0$, which corresponds to a complete emphasis on distinctive features in measuring similarity. In fact, the TCM representation is identical to the distinctive features representation, and so is also described by Figure 2. It explains 82.1% of the variance but has a larger GCC value of 23.6. As discussed earlier, this difference in GCC values arises because the TCM is a more complicated model, able to index more data distributions than the distinctive features model by varying the ρ parameter. Accordingly, with the same level of data-fit, the TCM model becomes less likely, and has a larger GCC value.

The MCM representation is shown Figure 3, explains 84.2% of the variance with a GCC

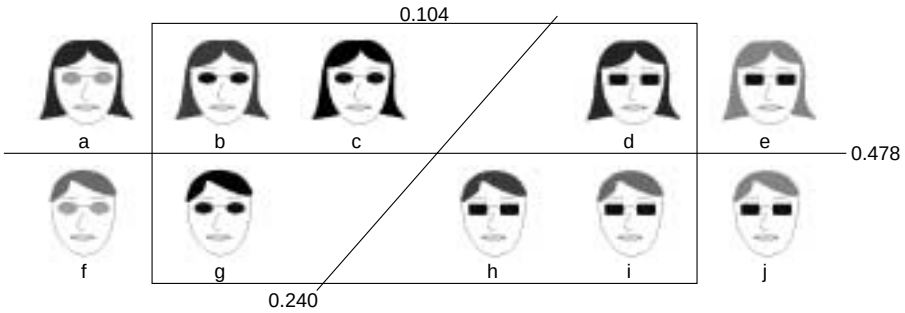

Figure 3: The MCM representation of the faces data, explaining 84.2% of the variance with a GCC value of 18.4 and $\widetilde{\rho} = 0.1$. The weight of the universal feature is 0.71, and the weights of the other three features are shown.

value of 18.4 and $\widetilde{\rho} = 0.1$. It also includes three features. There are two distinctive features, one of which divides males from females, with the other dividing people with square sunglasses from those with round sunglasses. The lone common feature corresponds to the 'unremarkable' people concept.

Discussion

The GCC values for the four models provide some evidence that the distinctive features and MCM representations are superior to the common features and TCM representation accounts. The poor data-fit of the common features representation indicates that it fails on the criterion of model accuracy. This is because many of the important regularities in the data, relating to differences between males and females, and those with round or square glasses, are best treated as distinctive features. The TCM, by using a distinctive features representation, achieves accuracy, but has a large GCC value because of its relatively greater complexity.

When interpreting the features in the distinctive features and MCM representations, it is difficult to find compelling reasons to prefer one over the other. The distinctive features representation divides those with pink sunglasses from those with blue sunglasses, which seems reasonable. The MCM establishes a common feature for unremarkable people that is also convincing. Our conclusion is that these data provide roughly equivalent evidence for both the distinctive features and MCM similarity models.

More generally, however, the face stimuli highlight an important theoretical difference, which is that only the MCM could represent the 'unremarkable' concept as a common feature. This is important because, while unremarkable things are similar to each other, this does not imply unusual things are similar. As an example, think of marriages[6]. While all good marriages probably have much in common, and so are similar to each other, it is not the case that all bad marriages are similar, because, unfortunately, there are so many way for marriages to go wrong. Returning to face stimuli, we expect people would judge unremarkable Western faces (e.g., Al Gore and Tom Cruise) to be more similar than unusual Western faces (e.g., Sid Vicious and John Malkovich), and so the 'unremarkable' concept needs to be treated as a common feature.

[6]We thank Nick Kingsbury for introducing us to this analogy, and Joshua Tenenbaum for directing us to the first line of Tolstoy's Anna Karenina: "Happy families are all alike; every unhappy family is unhappy in its own way".

General Discussion

One of the clear conclusions that can be drawn from the two evaluations is that neither common nor distinctive features alone are sufficient for a good representation, yet both are necessary. The kinship terms and cartoon faces show regularities consistent with both the common and distinctive features approaches for assessing similarity. The concept of gender in the kinship terms is a distinctive feature, but parent and sibling are common features; the shape of sunglasses on the faces is a distinctive feature, but the concept of unremarkable people is a common feature. For this reason, the motivation behind the TCM — to allow for both common and distinctive features to contribute to stimulus similarity — is well founded.

Comparing the two TCM and MCM representations, however, provides substantial empirical evidence for the superiority of the MCM. In both cases, the GCC favors the MCM, suggesting that it better maximizes data-fit while minimizing complexity on the key modeling criteria of accuracy and generalizability. Treating the two evaluations as independent pieces of equally weighted evidence gives an overall difference in GCC values of 17.4. This translates to odds of about 4,230 to 1 against the TCM.

The TCM representations are also more difficult to interpret, and often fail to include one or more of the important domains features. Using principled model theoretic constraints to control its complexity, the TCM did not express the parents or siblings concepts for the kinship terms, nor the 'unremarkable' concept for the faces. More generally, the TCM is unlikely, under complexity constraints, to include salient common features when it places great overall emphasis on distinctive features, or salient distinctive features when it places great overall emphasis on common features. Indeed, as the concrete examples of TCM representations make clear, the basic approach of having each feature act as both a common and distinctive feature in some mixture can make their psychological interpretation difficult.

The MCM provides a useful alternative to the TCM by embedding the balance between common and distinctive features within its basic representational assumptions. Through declaring each individual feature to be either entirely common or entirely distinctive in the way it contributes to stimulus similarity, the MCM incorporates common and distinctive features in a straightforward way. Stimulus similarity is conceived as increasing according to the weights of shared common features, and decreasing according to differences on distinctive features. Empirically, the two MCM representations largely contain a collection of the highly weighted features found using the purely common and purely distinctive feature models, and are easily interpreted.

Our general conclusions about the four models are as follows. The distinctive features model lacks representational flexibility, because it is constrained by metric axioms. This means, even when it is able to fit data well, and do so with low complexity, its representations are often difficult to interpret. The common features model has the representational flexibility to represent domains well, but has a weakness with respect to distinctive features. Where features divide stimuli into two meaningful groups, common feature representations have to use two features, and set the weights of these features to be the same. The TCM has the right motiviation of combining common and distinctive features, but its implementation is problematic. Contrasting common and distinctive features using a single global weight does not seem to be the most effective way of combining the two sources of stimulus similarity. The MCM takes a different approach, essentially augmenting the common features approach to allow distinctive features to be included. The MCM basically says that two complementary common features with identical weights are not a coincidence, but are best understood as a single distinctive feature. Accordingly, to the extent that stimulus domains have distinctive features[7], the MCM provides a useful

[7]Just how prevalent distinctive features are in realistic stimulus domains is debatable. For example, we had to remove the "cousin" term to create the gender distinctive feature. If it were included, both male and female common features would be needed to replace the gender distinctive feature, so that the cousin stimulus could belong to both. This would mean the MCM would be reduced to a common features model.

extension to common features representations, and an alternative to Tversky's Contrast Model.

While the MCM provides a very general account of featural stimulus representation — placing no restrictions on how stimuli share features, nor on whether features are declared to be common or distinctive — there remain a number of avenues for future work. The additive approach to similarity used by all four of the models is not the only possibility. It may prove useful to consider alternative mechanisms for combining the weights of features, or for contrasting the combined common and distinctive contributions to stimulus similarity. One possibility is to use a ratio approach for contrasting, so that similarity is measured as the proportion of common to distinctive features, rather than the difference between the two. Tenenbaum and Griffiths (2001) have provided a compelling argument for the ratio approach in the context of their Bayesian theory of generalization, and it would be interesting to see how featural representations differed by making this change.

Another obvious avenue for future work is to consider asymmetric similarity data. One of the important theoretical contributions of Tversky's (1977) featural models is their ability to model similarity judgments that depend on the ordering of stimuli. The basic MCM approach can naturally accommodate this possibility, in the same way as the General Contrast Model, by using parameters to weight the two distinctive components in Eq. (5). Whether the observed advantages of the MCM over the TCM are also observed for asymmetric similarities is an important topic for future work. The effects of context on similarity could also be considered along these lines. It may be that there is a need to reintroduce the hyper-parameters in Tversky's original formulation to give different emphasis to common and distinctive features in different contexts. A final issue that needs to be addressed is whether increases in self-similarities as stimuli share additional features are easily accommodated by the MCM, particularly if these features are otherwise best treated as distinctive.

Moving beyond featural representation, there is the challenge of developing representational models that incorporate both continuous dimensions and discrete features. The distinction between dimensional and featural approaches to mental representational modeling has been a classic one in cognitive psychology. It has frequently been suggested (e.g., Carroll 1976, p. 440, Tenenbaum 1996, p. 3, Tversky 1977, p. 328) that dimensional representations are better suited to low-level and continuous perceptual domains, such as tones and colors, while the featural representations are more appropriate for modeling high-level and abstract conceptual domains. Most environmental stimuli, however, would seem to require both representational mechanisms to express completely their similarities to other. As Carroll (1976, p. 462) concludes: "Since what is going on inside the head is likely to be complex, and is equally likely to have both discrete and continuous aspects, I believe the models we pursue must also be complex, and have both discrete and continuous components". We have recently (Navarro & Lee 2003) developed a model that combines common features with one or more continuous dimensions. A final important goal for future research is to extend this idea, using the MCM so that both common and distinctive features are combined with the dimensions.

Acknowledgments

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 Helen Braithwaite, Tim Johnson, In Jae Myung and Yong Su for helpful comments, and Robert Nosofsky and Joshua Tenenbaum, whose detailed and critical reviews greatly improved this paper.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

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

Arabie, P., Carroll, J. D., & DeSarbo, W. S. (1987). *Three-Way Scaling and Clustering*. Newbury Park, CA: Sage.

Balasubramanian, V. (1997). Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions. *Neural Computation 9*, 349–368.

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

Carroll, J. D. (1976). Spatial, non-spatial and hybrid models for scaling. *Psychometrika 41*, 439–463.

Chaturvedi, A., & Carroll, J. D. (1994). An alternating combinatorial optimization approach to fitting the INDCLUS and generalized INDCLUS models. *Journal of Classification 11*, 155–170.

Cheeseman, P. (1999). Foundations of probability. In R. A. Wilson & F. C. Keil (Eds.), *MIT Encyclopedia of the Cognitive Sciences*, pp. 673–674. Cambridge, MA: MIT Press.

Clouse, D. S., & Cottrell, G. W. (1996). Discrete multidimensional scaling. In *Proceedings of the Eighteenth Cognitive Science Conference*, San Diego, CA, pp. 290–294. Mahwah, NJ: Erlbaum.

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

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Glushko, R. J. (1975). Pattern goodness and redundancy revisited: Multidimensional scaling and hierarchical cluster analysis. *Perception & Psychophysics 17*(2), 158–162.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

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

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science 5*, 3–36.

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

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

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

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

Lee, M. D., & Pope, K. J. (2003). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical Psychology 47*, 32–46.

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 97*, 11170–11175.

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

Navarro, D. J. (2003). Regarding the complexity of additive clustering models: Comment on Lee (2001). *Journal of Mathematical Psychology 47*, 241–243.

Navarro, D. J., & Lee, M. D. (2001). Clustering using the contrast model. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pp. 686–691. Mahwah, NJ: Erlbaum.

Navarro, D. J., & Lee, M. D. (2003). Combining dimensions and features in similarity-based representations. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*, pp. 59–66. Cambridge, MA: MIT Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General 115*(1), 39–57.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology 43*, 25–53.

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

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review 109*(3), 472–491.

Potts, B. C., Melara, R. D., & Marks, L. E. (1998). Circle size and diameter tilt: A new look at integrality and separability. *Perception & Psychophysics 60*(1), 101–112.

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

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory 42*(1), 40–47.

Ritov, I., Gati, I., & Tversky, A. (1990). Differential weighting of common and distinctive components. *Journal of Experimental Psychology: General 119*(1), 30–41.

Rohde, D. L. T. (2002). Methods for binary multidimensional scaling. *Neural Computation 14*(5), 1195–1232.

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

Sattath, S., & Tversky, A. (1987). On the relation between common and distinctive feature models. *Psychological Review 94*(1), 16–22.

Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer.

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

Shepard, R. N. (1972). Psychological representation of speech sounds. In E. E. David & P. B. Denes (Eds.), *Human communication: A unified view*, pp. 67–113. New York, NY: McGraw Hill.

Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika 39*(4), 373–422.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science 210*, 390–398.

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

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In J. R. Pomerantz & G. L. Lockhead (Eds.), *The perception of structure: Essays in honor of Wendell R Garner*, pp. 53–71. Washington, DC: American Psychological Association.

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. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences 24*(4), 629–640.

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

Tversky, A., & Hutchinson, J. W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review 93*(1), 3–22.