Representing Stimulus Similarity

Daniel J. Navarro

Department of Psychology

University of Adelaide

December, 2002

Contents

Abstract

The practice of specifying stimulus representations using measures of similarity holds some status in cognitive modelling. Formal theories of cognitive processes such as generalisation, categorisation, identification, and recognition often employ these representations, and therefore rely on theories of stimulus similarity. With this reliance in mind, it is important to limit stimulus representations to those justified by observations of human thought and behaviour, rather than engaging in the questionable practice of specifying stimulus representations on the basis of introspection.

Over the last 50 years, psychologists have developed a range of frameworks for similarity modelling, along with a large number of numerical techniques for extracting mental representations from empirical data. This thesis is concerned with the psychological theories used to account for similarity judgements, as well as the mathematical and statistical issues that surround the numerical problem of finding appropriate representations. Specifically, the thesis discusses, evaluates, and further develops three widely-adopted approaches to similarity modelling: spatial, featural, and tree representation.

The spatial approach locates each stimulus in a multidimensional co-ordinate space, and assumes that the similarity between two stimuli is a function of how close they are to one another. Tree models represent stimuli as the terminal nodes in an acyclic graph, like a genealogical or taxonomic tree. The similarity between two stimuli is considered to be inversely related to the length of the unique path that connects them. Featural representations describe stimuli in terms of a set of characteristics that they either possess or do not possess. Featural similarity is assumed to be increased by shared features and decreased by distinguishing features.

This thesis develops three major themes. The first, discussed in detail in Chapter 3 but reiterated throughout, regards the important question of how to evaluate a representation. Any representation can be considered to be a model purporting to explain the set

of observed similarities, and should be evaluated as such. It is argued that the representation must not only provide a good fit to the data, but do so in the simplest possible manner, and should satisfy such qualitative constraints as interpretability and psychological plausibility. This thesis uses a Geometric Complexity framework to provide an appropriate trade-off between data-fit and model complexity. In doing so, expressions for the statistically principled Geometric Complexity Criterion are derived for several classes of similarity models. Furthermore, an extended investigation of the analytic properties of these measures for featural and tree models is presented, in order to provide an understanding of what makes one representation more complex than another. A briefer discussion of these issues with regard to spatial representations is also provided.

The second main aspect of this thesis is the discussion of featural representation in Chapter 4. Four theories of featural similarity are considered: the common features model, the distinctive features model, Tversky's seminal Contrast Model, and a new theory called the Modified Contrast Model. The Modified Contrast Model differs from Tversky's by assuming that each individual feature is declared to be a commonality or a distinction, rather than a weighted sum of both. These four theories are evaluated with regard to their psychological assumptions, their analytic properties, and their performance when applied to several empirical data sets. In addition to applying these models to pre-existing data, three new data sets are collected in this evaluation. These investigations suggest that the Modified Contrast Model may combine the common and distinctive elements required by a featural theory in a more plausible manner than Tversky's model.

The third theme to this thesis, discussed in Chapter 5, is the development of an approach to spatial representation that allows a single point to represent multiple stimuli. Based on the prototype theory of categorical structure, this approach enables a set of prototypes to be directly inferred from similarity data. The effectiveness of three "prototype scaling" algorithms are evaluated, and the best algorithm is then applied to

several data sets in order to illustrate the potential of the approach.

Overall, it is argued that the different representational frameworks are each best suited to different domains, and that it is therefore important to consider their assumptions, and seek to fit models that are appropriate to the context. As discussed in Chapter 6, future work in this regard might treat similarity judgements as decision processes. Finally, no matter which similarity theory is adopted, the measure of a model should take account of its data-fit, its complexity, and its interpretability.

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by any other person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

Daniel Navarro

December 29, 2002

Acknowledgements

Hidden not-too-deeply behind the elegant facade of any thesis lies years of panic attacks interspersed with the occasional spark of an idea. My deepest thanks go to all those people who helped foster the second and hinder the first. My supervisors, Michael Lee and Doug Vickers, fall into both these categories: what I have learned from them goes well beyond the scope of this thesis, and it is largely thanks to them that I have learned anything worthwhile about psychology. So too do my parents, who taught me to think for myself, instilled in me what sometimes passes for a work ethic, and helped keep things in perspective with the occasional reminder of the probable size of my readership.

I would like to acknowledge the financial assistance I have received these past three years, in the form of an Australian Postgraduate Award, along with additional funding from the Defence Science and Technology Organisation. My thanks go to Chris Woodruff for making the latter happen. Furthermore, without the Psychology Department providing office space, computers, and other assistance, I'd never have started this, much less finished.

Over the years I've been incredibly lucky to have friends and colleagues without whom I'd have ended up completely mad, too many of them to name here. Even so, I should specifically mention Mel Roberts, who has had the misfortune to endure more whingeing from me than any human being should. Her inhuman tolerance should probably be studied, in the interests of world peace.

Finally, I would like to thank Vandelay Industries for all their LaTeX support and hereby resign from my post as Chief Departmental Eeyore.

1. Prelude

Cognitive psychology is in part the science of understanding how cognitive processes operate, and one tool for achieving this goal is the specification of formal models of these processes. It is common for these models to describe processes that act on stimuli: accordingly, a fundamental issue in cognitive modelling regards how to appropriately encode these stimuli. An informational structure that describes a stimulus is called a *stimulus representation*, and the issue at hand can therefore be stated simply: what kinds of stimulus representations are appropriate for a model of human cognition?

As argued by many authors (e.g., Brooks, 1991; Komatsu, 1992; Lee, 1998), 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 is to obtain measures of the similarity between all pairs of stimuli for the domain of interest (e.g., Kruschke, 1992; Nosofsky, 1992b; Shepard & Kannappan, 1991). The assumption underlying this approach is that the process in operation when the measurements are obtained is sufficiently simple that the resulting 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 superior to the alternative approach of hand-tuning representations, which may not reflect human representational structures in any regard.

This thesis addresses the matter of deriving mental representations by modelling

similarity judgements. This chapter begins the account by giving a general overview of mental representation and a background to similarity. Chapter 2 focusses on the psychological theories used to explain similarity judgements and the techniques used to extract representations from data. The chapter also provides a number of original analyses of existing data sets to examine the various representational frameworks. Chapter 3 addresses the important question of how to choose between alternative representations, taking into consideration the issue of model complexity, and giving an account of what makes a representation more or less complex. Next, in Chapter 4, four approaches to the influential featural representation framework are evaluated. The psychological implications of these models are considered, their analytic properties examined, and the models are applied to several empirical data sets, both pre-existing and new. Chapter 5 proposes a new representational formalism based on the spatial approach, that aims to represent categorical information about a set of stimuli in a multidimensional space, and applies it to several data sets. The notion of modelling similarity as a decision process is canvassed in Chapter 6, followed by concluding remarks in Chapter 7.

1.1 The Very Idea of Representation

Psychological models tend to be functionally oriented. Rather than attempting to work out what each part of the brain does (which is more the domain of neuropsychology and neuroscience), cognitive psychologists try to understand how a particular type of behaviour could be produced. Psychological models do not necessarily refer to neurophysiological structures, but rather to logical structures. The two approaches are complementary: a neural-level description must produce psychological phenomena, and psychological models should be consistent with the neural substrate on which the mind rests. This thesis is psychological, not neurological, and so the representational models developed do not refer to literal, neurophysiological structures. When a psychological

model is inferred, it is not assumed that there exists any specific neural circuitry devoted to producing it: instead, it is suggested that people behave *as if* such a thing existed.

The functional nature of psychological models, including mental representations, requires that one assess their validity by looking at the assumptions they entail, their explanatory power, their generalisability and so on. This matter of specifying *appropriate* representations is a major theme in this thesis, discussed in detail in Chapter 3 but reiterated throughout. For the moment, however, it suffices to observe that the very idea of mental representation deserves examination: to justify psychological theories of human mental representation, one should start with a discussion of what a representation is, and how representations fit into the enterprise of modelling cognition. Therefore, this section discusses mental representation in general terms, attempting to provide a broad view of what the term means in different contexts. In doing so, it draws upon three different philosophical ideas about how cognition should be modelled: the classical, connectionist, and dynamical approaches.

1.1.1 Classical Theories

The first approach is the *classical* theory, also known as the computational theory[1], which attempts to provide an account of cognition in terms of the operation of rule-governed processes on discrete symbols (e.g., Fodor, 1983; Haugeland, 1985; Newell, 1980, 1990; Pylyshyn, 1984). Therefore, the classical theory is based on the metaphor of the mind as a computer, or *symbol processor*. Obviously however, the metaphor should not be pushed too far: the brain does not closely resemble a personal computer, and neuroscience is not computer science. Modern classical theories (e.g., Calvin, 1996;

[1] "Computation" is a slippery concept. In a widely used sense, computation refers to any operation that can be performed by a Turing machine, and it is in this sense that the classical theory is computational. However, the intuitive notion of computing can be much broader, and can encompass any mathematical or numerical operations. Therefore, to avoid possible disagreements over what kinds of models are computational, the term "classical models" is used here.

Fodor, 1983; Minsky, 1986; Pinker, 1998) tend to treat the mind more as an arbitrary number of machines working in parallel than as a single serial processor.

The classical theory assumes that a stimulus is represented by a symbol, or set of symbols, activated directly when the stimulus is encountered, or by proxy when the stimulus is recalled. A symbol is considered to be a manipulable token that bears an arbitrary relationship[2] to the stimulus or concept it represents. Mental representation is crucial within the classical theory, according to which the task facing mental representation research is to discover the symbols used in cognition and determine their structure. That said, it is difficult to state precisely what it is that makes a set of symbols count as a representation. Markman and Dietrich (2000a, 2000b) emphasise the role of a mental representation as an information bearer. That is, a state is a representation by virtue of the fact that it embodies some knowledge about the stimulus being represented. They suggest that, in addition to being discrete, symbolic information bearers, classical mental representations should be enduring (rather than transient) states, as well as abstract (rather than perceptual) states. They should also have a compositional structure (i.e., they can be combined with one another), and be acted upon by rules.

Several caveats attach to this definition. Firstly, it disregards the fact that the representations employed by people are tailored to context (Barsalou, 1989; Hofstadter, 1985, 1995). Although some long-term representations may change slowly (or not at all), other representational structures (e.g., those used in working memory) are transient in nature. Furthermore, Barsalou (1999) has argued that some mental representations at least are modality-specific perceptual states, and not abstract conceptual structures. Examples of such representations would be the sensory homunculi that represent the various sensory surfaces in the brain (see Kolb & Whishaw, 1996 or Kandel, Schwartz, & Jessell, 1995

[2]In this context "arbitrary" simply means that there need not be any particular relationship between stimulus and symbol: the mental symbol for a cat, for example, need not be cat-shaped, cat-coloured, or cat-like in any regard. This does not, however, preclude the symbol from carrying *information* about the shape or colour of a cat.

for example). Another key element missing in their definition is the fact that representation can occur at a number of levels, ranging from representations of raw sensory information to information about the categories to which a stimulus belongs.

1.1.2 Connectionist Theories

A more recent alternative approach to cognitive science is the connectionist paradigm (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986). Connectionist models are built from networks of interconnected nodes, in which processing occurs by passing activation from one node to another. In this manner, the environmental information given to the model is transformed into an appropriate response by virtue of the number, type and strength of connections between nodes. Representation in connectionist systems does not take the form of classical symbols (Smolensky, 1988; but see Fodor & Pylyshyn, 1988): instead, it is implicit in the pattern and strengths of the connections between nodes.

One distinction made within connectionist approaches is between localist and distributed representations. In localist representations, a single node represents a single stimulus and vice versa, whereas in distributed representations, each stimulus is denoted by a pattern of activation across many (possibly all) of the nodes (Hinton, McClelland, & Rumelhart, 1986). Distributed representations are often less easily interpreted than localist representations. However, a well-structured distributed representation can resemble a localist representation constructed at multiple levels of generality. Consider a model such as the schematic processing model proposed by Rumelhart, Smolensky, McClelland, and Hinton (1986) in which various household rooms are represented by patterns of activation across nodes, each of which denotes a single object or feature that might be found in the room. However, it is highly plausible that the cognitive representation of the individual objects could also be distributed over a host of microfeatures.

For instance, the representation of a coffee cup could be a part of the representation of kitchen, but itself be distributed over such microfeatures as "cylindrical", "has a handle", "filled with coffee", and so on (Smolensky, 1988). Thus the representation can be specified at a lower level of generality than the phenomenon of interest, at the "subcognitive level" that Hofstadter (1985) regards as crucial to explanations of cognition.

This discussion of classical and connectionist models suggests that representation is a key issue for both approaches. To some extent, the concerns regarding what constitutes a representation in one theory overlap with the corresponding concerns in the other. For instance, one prominent representational formalism describes stimuli in terms of defining characteristics or features (Tversky, 1977, discussed in detail in Section 2.3 and Chapter 4). The theory is classically framed, in that features are treated as symbols that denote external objects and can be operated on as tokens in a rule-governed system (such as computational psychological models). However, the framework is easily applied in connectionist architectures, by establishing a correspondence between features and nodes. In fact, it is this kind of correspondence that enables some models to be equally expressible as classical or connectionist systems. For instance, the highly successful ALCOVE model of category learning was originally formulated by Kruschke (1992) as a three-layer connectionist network employing spatial representations. However, Lee and Navarro (2002) developed an extension of the model that allows it to accommodate featural representations, and framed the model in terms of rule-governed processes (similarity-to-exemplars followed by a decision rule). In short, although the classical and connectionist theories adopt quite different perspectives on what representations are and how they are used, they can often be encompassed by the same representational formalism. In this instance, it is observed that spatial and featural representations are applicable to both classical and connectionist models.

1.1.3 Dynamical Theories

Dynamical explanation (e.g., Beer, 2000; van Gelder & Port, 1995) constitutes the other main approach to cognition, and is the newest of the three main frameworks. The essence of dynamical theories is the emphasis on the evolution of a cognitive system over time. Given some system which may incorporate both the individual and aspects of their environment, a dynamical model specifies a set of states in which the system may be, called the *state space*, and some function that describes how the state of the system changes over time.

Dynamical theories rarely, if ever, propose classical representations of the kind described by Markman and Dietrich (2000a, 2000b). For example, consider the dynamical explanation of infant reaching behaviour proposed by Thelen (1995). She proposes that the infant's arm can be modelled as if it were a mass-damped spring (a spring with a weight attached to the end, which acts to dampen any motion) with a regular forcing function (a function that introduces force to the spring, causing motion). This theory does not refer to representational states internal to the infant. Nevertheless, dynamical cognitive theories are consistent with the notion of an information bearing state, inasmuch as information passes through the individual and shapes their behaviour within the environment: if they were not consistent in this manner, they would not be psychological theories at all. Returning to the damped-spring reaching model, Thelen observes that the parameters of the model change over time as the infant learns to control his or her movements and that these parameters depend on the internal states of the infant. In a loose sense, these states can be thought of as representational, though not necessarily in the classical sense.

The interplay between different cognitive paradigms as well as the sense in which dynamical systems can be representational are evident in Elman's (1995) treatment of language as a dynamical system. Using the backpropagation of error method (Rumelhart,

Hinton, & Williams, 1986), he trained a recurrent neural network to predict the next word in a series of sentences, and then examined the word representations that the network had learned. Though the word vectors were high dimensional, the variability between them was representable in a low dimensional metric space using principal components analysis (related to multidimensional scaling, discussed in Section 2.2). He argues that lexicon is spatially represented, and that linguistic rules define the dynamics on that space: that is, given an input word, the network moves to a new state in the space according to these rules. A sentence, therefore, is a trajectory in the lexical state space. This approach is consistent with spatial representations (indeed it depends on the spatial representability of the lexicon).

1.1.4 Summary

Mental representation plays an important role in classical, connectionist and dynamical theories of cognition, though in a different manner for each. Furthermore, psychological approaches to mental representation (e.g., spatial or featural) cut across the divide between the three types of cognitive models, suggesting that they have broad applicability as models of human mental representation.

1.2 Types of Similarity

"Similarity", in an intuitive and dictionary sense, refers to some measurement of "likeness" or "resemblance" (though these terms are hardly more informative). Contrastingly, a single technical definition of similarity is difficult, if not impossible, to provide, as the term has been used in a wide range of contexts. Psychological similarity is a complex phenomenon, involving both bottom-up and top-down factors, and it is perhaps more useful to focus on the various kinds of similarity that have been discussed, rather than engage in a fruitless attempt to define a word whose meaning is commonly understood.

1.2.1 Perceptual and Conceptual Similarity

In the bottom-up sense, variously referred to as *perceptual similarity* (Rips, 1989), *surface similarity* (Vosniadou & Ortony, 1989) or *physical similarity* (Brewer, 1989) two stimuli appear alike by virtue of low level processes. An example of this type of similarity would be the similarity between a human being and a mannequin. Although humans and mannequins are conceptually very different, and belong to very different categories, in a simple perceptual sense they are quite similar. According to this kind of similarity process, two stimuli are related through the operation of concrete, low-level processes.

Perceptual similarity is contrasted against the kind of top-down similarity which has been labelled *conceptual* or *deep similarity*. Conceptual similarity invokes a more abstract knowledge about the world and the stimuli being compared. In this sense, two objects that are members of the same category or otherwise conceptually related are more likely to be judged more similar because of the understanding that they belong to the same class of thing. For instance, while a mannequin looks more like a human than does a chimpanzee, a chimp is conceptually more like a human, because of the understanding that chimpanzees and humans belong to many of the same categories (primates, intelligent animals, etc.), whereas mannequins share few such similarities with humans.

Of course, it seems highly unlikely that similarity is either wholly perceptual or wholly conceptual. Perception flows into cognition, and cognition shapes and directs perception (Hofstadter, 1995). Accordingly, similarity judgements employ perceptual and conceptual likenesses as the context demands. Although it is implausible to think that the kind of similarity that makes freedom like liberty is much like the kind of similarity that makes koalas look like teddy bears, both types of similarity may be employed at different times. Indeed, Medin and Ortony (1989) argue that the development of concepts

(and hence conceptual similarity) may initially be based on more concrete perceptual similarities.

1.2.2 Global and Dimensional Similarity

A different distinction is made by Smith (1989), who refers to *global similarity*, in which stimuli are considered in terms of an overall or holistic impression, and *dimensional similarity*, in which two objects are viewed in terms of a set of analysable characteristics or dimensions. She argues that global similarity is a developmental precursor to dimensional similarity, in that until infants and children learn to carve up the world according to perceptually and conceptually relevant dimensions, their impressions of the "likeness" of two things is necessarily holistic. It is worth noting that this distinction is related to that made between integral and separable dimensions (e.g., Garner, 1974; Shepard, 1991; see Section 2.2). Smith's theory suggests that dimensional similarity does not replace global similarity, but overlays it. Accordingly, there should be evidence that the same stimulus may sometimes be evaluated in a holistic (integral) manner, and sometimes in a dimensional (separable) manner. Smith argues that this is the case, though this remains an unresolved question.

1.2.3 Summary

This discussion makes it clear that similarity is not a monolithic, homogeneous entity, being composed of several types of likeness that may interact with one another in various ways. Given this, a pragmatic view of similarity is adopted here. Similarity is not treated as a cognitive primitive: rather, similarity is understood to be the result of processes that act on mental representations.

1.3 Is Similarity Indeterminate?

A philosophical criticism of the use of similarity as an explanatory principle (Goodman, 1972) claims that similarity is indeterminate, since it is always possible to define an infinite number of characteristics shared by a pair of stimuli. For example, both an elephant and a neutron star have the characteristic of "weighing more than a tonne". According to this argument, by defining such a feature shared by both, their similarity should increase. Since the number of features that can be defined in this manner is limitless, the similarity between all pairs of things should be infinite. However, this is based on a psychologically implausible notion of how stimuli are compared, in that irrelevant characteristics do not enter into the process. As Medin and Ortony (1989) and Hahn and Chater (1997) argue, similarity judgements are made using some representation: "A first step is to define similarity not in terms of all logically possible shared predicates but in the more restricted sense of shared *represented* predicates. For example, both tennis balls and shoes share the predicate *not having ears*, but it is unlikely that this predicate is part of our representation of either tennis balls or shoes." (Medin & Ortony, 1989, pp. 180). Furthermore, if the representation employed is influenced by the demands of context (e.g., Barsalou, 1989; Goldstone et al., 1997; Hofstadter, 1985, 1995) then similarity should be expected to vary accordingly. Crucially, however, it should not be expected to vary arbitrarily. Unless prompted by some highly contrived reasoning, no-one seriously believes that a neutron star is very much like an elephant.

1.4 The Role of Similarity in Cognition

Similarity has been employed as an explanatory device in a number of areas in cognitive modelling. Although this thesis is primarily concerned with representational explanations of similarity, it is instructive to consider how these notions of similarity are relevant

within the wider context of the study of cognition.

1.4.1 Identification

The Similarity Choice Model (Luce, 1963; Shepard, 1957) of stimulus identification was proposed to predict the probability with which people misidentify a stimulus. Formally, the probability that stimulus i is identified as stimulus j is given by,

$$p(j|i) = \frac{b_j s_{ij}}{\sum_k b_k s_{ik}}$$

where s_{ij} denotes the similarity between stimuli i and j, and b_j is a response bias parameter for stimulus j such that $0 \leq b_j \leq 1$ and $\sum_k b_k = 1$. The SCM is also discussed by authors such as Nosofsky (1986, 1992b) and Getty, Swets, Swets, and Green (1979). Importantly, Nosofsky (1986) observes that this model requires the similarities to be constrained by some theory of similarity: otherwise, the model has more free parameters than data points. In particular, he argues that the spatial approach employed by Shepard (1957) is a simple and effective way to do so, but there is no *a priori* reason why other representational formalisms should not be appropriate for this purpose.

1.4.2 Generalisation

In a now-classic paper, Shepard (1987, see also Myung & Shepard, 1996) formulated the "universal law of generalisation", which predicts the probability with which an organism will assume that a novel stimulus j has the same environmental consequences as a previously encountered stimulus i. This law emerges from a rational analysis of the structure of psychological spaces. In particular, Shepard argues that – for evolutionary reasons – a stimulus space is structured so that stimuli with the same real-world consequences will be near one another in the space. Correspondingly, a set of stimuli with the same consequences form a "consequential region" in the space. Shepard demonstrates

that, with little knowledge about the size of the consequential region that contains i, the probability that j is also in the region is given by an exponential decay function of the distance between them, $\hat{d}_{ij}$. That is,

$$g(j|i) \propto \exp(-\hat{d}_{ij}).$$

Although Shepard's theory is inherently spatial in nature, Russell (1986) has argued that the same law should emerge from appropriately formulated featural representations. Furthermore, the extension to the theory proposed by Tenenbaum and Griffiths (2002a) shows that featural representations can be accommodated in this way.

1.4.3 Categorisation

Categorisation is an area in which similarity has been widely employed. In fact, stimulus similarity is so fundamental to the prototype and exemplar views of conceptual structure that both have been labelled "similarity" or "resemblance" views (Komatsu, 1992; Medin, 1989; Rips, 1989). According to the exemplar view, typified by models such as ALCOVE (Kruschke, 1992; Lee & Navarro, 2002) and the General Contrast Model (GCM; Nosofsky, 1986), a category is represented by a set of past instances ("exemplars"). When assigning a stimulus to a category, ALCOVE and the GCM both calculate the (attention-weighted) similarity of the presented stimulus to each stored exemplar as a means of estimating the similarity to each category. ALCOVE allows exemplars to have ambiguous category memberships: that is, a single exemplar may be associated with multiple categories, indicating that ALCOVE does not "know" which category it belongs to. The category assignment is made using the SCM-rule.

According to the prototype view (see Rosch & Mervis, 1975 for example), a category is represented by the typicality or central tendency information about its instances. In formal models the similarity between a stimulus and a category is calculated by comparing the stimulus representation to the prototype representation (see Nosofsky, 1992a

for an overview). For example, Nosofsky (1992a) notes that Massaro and Friedman's (1990) Fuzzy Logic Model of Perception (FLMP) can be characterised as such a model, using a multiplicative distance metric.

1.4.4 Recognition

Nosofsky, Clark, and Shin (1989) also provide a similarity-based account of recognition. If C_k denotes the set of stimuli belonging to the kth category, and m_j denotes the strength with which the jth stimulus is stored in memory, then the familiarity of the ith stimulus is given by,

$$f(i) = \sum_k \sum_{j \in C_k} m_j s_{ij}.$$

This rule is related very closely to the SCM-rule used by ALCOVE and the GCM, but differs in that the SCM-rule is a relative measure, whereas the recognition rule is an absolute measure (Nosofsky, 1992b).

1.5 Summary & General Discussion

The importance of similarity as an explanatory tool in psychology is highlighted by the extent and effectiveness with which it has been used. At a fundamental level, the ability to observe that "this thing is like that thing", and use that information appropriately is crucial to intelligent behaviour, and may even lie at the core of cognition itself (Hofstadter, 2000). If so, it is crucial to understand how similarity operates. As philosophers such as Goodman (1972) have pointed out, it is not sufficient to assume that similarity is cognitively primitive, in the sense of having no constituent structure itself. Psychological theories of similarity are required: similarity may not be taken as an assumed thing. Two things are similar for a reason, not "just because". Throughout this thesis, it is these theories of similarity that are evaluated. It is important to realise that when

collecting and analysing similarity measures, the resulting representations can only ever reflect the information used by participants when the experiment took place. Similarity measures are always contextually bound, and the representations derived should never be taken as the entirety of a participant's knowledge. Nevertheless, an understanding of the information that they did use when making some similarity judgement goes a long way towards providing psychologically appropriate theories of similarity.

2. Theories of Similarity

The previous chapter introduced similarity and mental representation in general terms. In this chapter the discussion turns to the specific psychological theories proposed to account for similarity judgements. The task of similarity modelling can be stated succinctly, as follows: given an $n \times n$ matrix of similarity judgements $\mathbf{S} = [s_{ij}]$ (or dissimilarity judgements $\mathbf{D} = [d_{ij}]$), what underlying representation most probably gave rise to the data? Goldstone's (1999) recent review identifies four main approaches to similarity modelling: spatial, featural, alignment-based, and transformational. To these four trees and networks might also be added. Therefore, this chapter is structured as follows: several sets of similarity data are described, which are used to provide concrete examples throughout the thesis, and then each of these six approaches to similarity modelling are reviewed. Special emphasis is placed on spatial, tree and featural representations, as these are the most fully developed approaches and are the focus of the research reported in Chapters 3, 4 and 5.

2.1 Similarity Data Sets

As the concern in this thesis is the analysis of similarity data, it is useful to have data sets to analyse. Three sets of empirical similarity data are collected in this thesis (all in Chapter 4). Additionally, several data sets collected by other researchers are used to make a range of theoretical points. This section briefly describes these data sets.

Where possible, precision estimates have been made: that is, estimates of the inherent uncertainty associated with the data. Precision estimates are important when analysing data sets, as noisy data do not warrant as detailed a model as highly precise data (Lee, 2001a). The aim when modelling similarity data is to represent the structure in the data, not the noise. Noisy data provide less evidence about the underlying or "true" similarity, and are therefore less able to support elaborate representations (see Chapter 3).

2.1.1 Risks

Johnson and Tversky (1984) investigated people's perception of risks using a number of tasks, one of which was to rate the similarity of one risk to another. The domain they looked at was the risk of death from 18 causes: stroke, heart disease, stomach cancer, lung cancer, leukaemia, toxic chemical spills, nuclear accidents, war, terrorism, homicide, airplane accidents, traffic accidents, accidental falls, floods, tornados, fire, lightning and electrocution. Each of their 245 participants were presented with all pairs of stimuli and asked to rate their similarity on a nine-point scale ranging from "very dissimilar" (1) to "very similar" (9). The similarity matrix is presented in their paper, but empirically derived precision estimates are unavailable. Analyses of this data set are presented in Section 2.4 and 5.5.

2.1.2 Colours

In a classic experiment Ekman (1954) asked 31 participants with normal colour vision to rate the similarity of 14 colours, produced by placing various coloured filters in front of a light source. These filters transmitted light of wavelengths 434, 445, 465, 472, 490, 504, 537, 555, 584, 600, 610, 628, 651, and 674nm. The five-point rating scale ranged from "no similarity at all" (0) to "identity" (4). Again, empirical precision estimates are not available. An analysis of this data set is presented in Section 2.2.

2.1.3 Drug Use

Huba, Wingard, and Bentler (1981) conducted a large scale study of the patterns of drug use of 1634 students in grades 7 through 9 in the greater Los Angeles area. Each participant was asked the frequency with which they used each of the following drugs: cigarettes, beer, wine, liquor, cocaine, tranquilisers, prescription medication (for recreational use), heroin (and other opiates), marijuana, hashish, inhalants (glue or petrol), hallucinogenics (LSD, psilocybin or mescaline) and amphetamines. The rating scale ranged from "never tried" (1) through "only once" (2), "a few times" (3), "many times" (4) and "regularly" (5). A measure of the similarity between the patterns of use for two drugs is obtained through the correlation between responses for the two drugs. No empirical precision estimate is available. Analyses of this data appear in Sections 2.2, 2.3 and 2.4.

2.1.4 Arabic Numerals

Shepard, Kilpatric, and Cunningham (1975) collected similarity data for the numbers 0 through 9, using four participants. They presented the numbers in a variety of ways, including Arabic and Roman numerals, regular polygons with different numbers of sides and spoken words. Participants responded on a continuum, but the responses were then binned into a 21-point scale. Section 2.3 re-presents Tenenbaum's (1996) analysis of the Arabic numerals data. No empirical precision estimate is available.

2.1.5 Kinship Terms

Rosenberg and Kim (1975) measured the similarity of 15 English kinship terms: aunt, brother, cousin, daughter, father, granddaughter, grandfather, grandmother, grandson, mother, nephew, niece, sister, son and uncle. 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. Following Lee (2001a), the empirical precision of this data set was estimated by calculating, for each stimulus pair, the sample standard deviation of their similarity values for the six groups. The overall precision was estimated at 0.09 by taking the average of these sample standard deviations. This data set is analysed in Section 4.6.

2.1.6 Plants, Animals and Colours

This data set, reported by Cooke, Durso, and Schvaneveldt (1986), consists of the following set of 25 concepts: living thing, animal, mammal, hairs, dog, deer, bats, blood, bird, feathers, robin, chicken, antlers, hooves, frog, plant, leaves, tree, cottonwood, flower, rose, daisy, colour, green and red. Participants were asked to judge the relatedness of pairs of concepts on a 10-point scale (from 0-9). The individual subjects' ratings are not available, so empirical precision estimates cannot be obtained for the data. This data set is analysed in Section 5.5.

2.1.7 Animals

O'Doherty and Lee's (2002) "animals" data set consists of the following 21 stimuli, presented pictorially and in written form: koala, chimpanzee, elephant, camel, cow, zebra, horse, lion, cat, dog, chicken, eagle, bat, dragon, snake, scorpion, butterfly, bee, frog, shark and goldfish. Participants were asked to rate similarity on either a 5-point or 11-point scale (the scale manipulation did not affect the overall similarity estimates, so the data is aggregated across the conditions). The empirical precision estimate for the written form stimulus data which is analysed in Section 5.5 was obtained by taking the standard deviation of the individual participants' similarity ratings for a given pair of stimuli, and then averaging across the pairs to yield the estimated precision of 0.18.

2.2 Spatial Representation

The origins of spatial representation lie with the notion of a stimulus dimension, in which a number of stimuli are ordered according to some criterion. This idea of assigning a value to each stimulus on a dimension is one fundamental to measurement theory, and has attracted some attention over the years (e.g., Stevens, 1946; Michell, 1986; Krantz, Luce, Suppes, & Tversky, 1971). An example of a stimulus dimension that applies to the perception of people is "height". One can assign to each person a height-value, and use the distance between people on the height-dimension when making judgements. Spatial representation incorporates this notion of a stimulus dimension, but allows stimuli to be assigned values on multiple stimulus dimensions (e.g., "height", "weight", "age", "intelligence" and so forth). Multiple dimensions yield a space in which stimuli are represented by a point: spatial representations of similarity data therefore model the *dis*similarity between two stimuli as a function of the distance between them in the psychological space. As a result, spatial representations are subject to the metric axioms: minimality ($\hat{d}_{ii} = 0$), symmetry ($\hat{d}_{ij} = \hat{d}_{ji}$), and the triangle inequality ($\hat{d}_{ik} \leq \hat{d}_{ij} + \hat{d}_{jk}$). Examples of spatial representations are displayed in Figures 2.1 and 2.2. Note that the issue of dimensionality determination (and corresponding concerns in other similarity theories) involves a trade-off between data-fit and model complexity. Though disregarded in this chapter, the matter is discussed in detail in Chapter 3.

2.2.1 Distance in Psychological Spaces

Early work on spatial representation highlighted the importance of choosing an appropriate distance metric. Attneave (1950) provided evidence that the "city block" metric provided a good account of certain data sets. Under a city-block metric, the separation between two points is measured by

$$\hat{d}_{ij} = \sum_k |p_{ik} - p_{jk}|,$$

where p_{ik} denotes the value of the ith object on the kth dimension, and $\hat{d}_{ij}$ refers to the dissimilarity estimate predicted by the model. Contrastingly, Torgerson (1958) highlighted the utility of the familiar Euclidean metric,

$$\hat{d}_{ij} = \sqrt{\sum_k (p_{ik} - p_{jk})^2}.$$

Attneave observed that the city-block metric implies a unique set of axes, whereas the Euclidean metric is unaffected by arbitrarily rotating the axes, and argued that the city-block makes sense when the uniquely specified axes correspond to easily identifiable stimulus dimensions. Torgerson in turn acknowledged this property of the city-block metric, but observed that the Euclidean metric is appropriate when the stimulus dimensions are not so obvious. Garner (1974) refers to "obvious" dimensions as being "separable" (in the sense that they may be attended to separately from one another), and dimensions that cannot be separately attended to as "integral". As it is often acknowledged (e.g., Garner, 1974; Shepard, 1991; Torgerson, 1958) that dimensions may be partly separable, it is commonplace to use one of the Minkowski r-metrics,

$$\hat{d}_{ij} = \left(\sum_k |p_{ik} - p_{jk}|^r \right)^{\frac{1}{r}} \tag{2.1}$$

where $r = 1$ yields the city-block metric and $r = 2$ gives the Euclidean metric. Metrics lying between these extremes (i.e., $1 < r < 2$) correspond to dimensions that are only partly separable. While Shepard (1991) observes that metrics with $r < 1$ have a psychological interpretation as dimensions that compete for attention, there is no obvious correspondence with metrics where $r > 2$, since it is unclear what "more than totally integral" means (Lee, 2001a).

Shepard (1991) provides a powerful theoretical justification for the identification of integral and separable dimensions with Euclidean and city-block metrics. In his theory of generalisation (Shepard, 1987), stimuli are considered to belong to natural kinds that correspond to connected, "consequential" regions in the psychological space. Individuals will therefore tend to treat two stimuli as more similar according to the probability with which they are assumed to belong to the same consequential region. Shepard (1991) therefore argues that if two dimensions are perfectly positively correlated in the environment – that is, a value on one dimension exactly predicts the value on the other dimension – people will assume that the consequential regions have the same extension in the psychological space. If so, the generalisation gradient – and therefore contours of equal similarity in the space – will be circular in shape, which is the characteristic of the Euclidean metric. Accordingly, if the two dimensions are assumed to be uncorrelated, the generalisation gradient becomes diamond-shaped, which is the signature of the city-block metric. Similarly, the assumption of an intermediate positive correlation implies an intermediate metric, and a negative correlation yields a "competitive" metric ($r < 1$).

The numerical problem of extracting a spatial representation from similarity data is known as multidimensional scaling (MDS; see Cox & Cox, 1994; Borg & Lingoes, 1987; Davison, 1983; Carroll & Arabie, 1980 for overviews), which will now be discussed.

2.2.2 Classical MDS

Although it is the oldest of the multidimensional scaling techniques, classical scaling (Young & Householder, 1938; Torgerson, 1958) is of some interest as it is still widely used, and is less computationally-expensive than most MDS approaches. If $\mathbf{D} = [d_{ij}]$ denotes the proximity matrix, then classical scaling works by "doubly centering" $\mathbf{D} \cdot \mathbf{D} = [d_{ij}^2]$ (subtracting the row and column means from each element in the matrix, then adding in the grand mean), to obtain the matrix of scalar products $\mathbf{B} = [b_{ij}]$, where

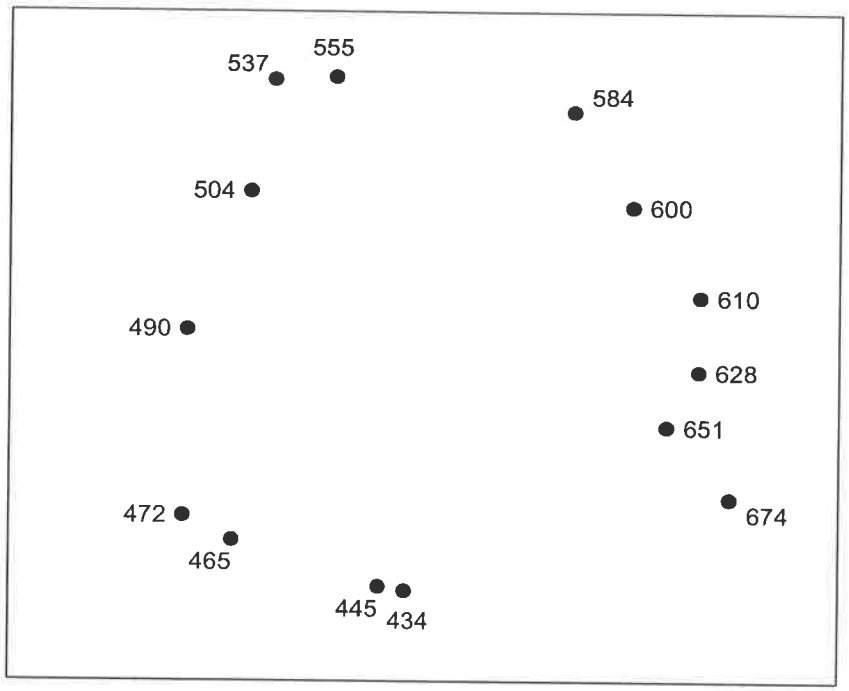

Figure 2.1: Best fitting two dimensional configuration for Ekman's (1954) colour data, explaining 81.3% of the variance. This representation was found using a gradient descent approach to multidimensional scaling, and is an original analysis, though it closely resembles other MDS analyses of this data set.

$$b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \frac{\sum_a d_{aj}^2}{n} - \frac{\sum_b d_{ib}^2}{n} + \frac{\sum_a \sum_b d_{ab}^2}{n^2}\right)$$

Note that this double centering technique, introduced by Torgerson (1958), is useful in the analysis of proximity matrices containing noise, since the centroid is less likely to be perturbed than any individual point (because the centroid moves as a function of the mean perturbation of all points, and is thus less sensitive to random error). If $\mathbf{B}$ is positive semi-definite, then $\mathbf{D}$ can be perfectly represented in a Euclidean space (Young & Householder, 1938). Eigenvalues of $\mathbf{B}$ are calculated by singular value decomposition, yielding $\mathbf{B} = \mathbf{UVU}'$. The co-ordinates are given by the $n \times r$ matrix $\mathbf{P} = [p_{ik}] = \mathbf{UV}^{\frac{1}{2}}$, where r is the rank of $\mathbf{B}$. The m-dimensional solution is found by retaining only the

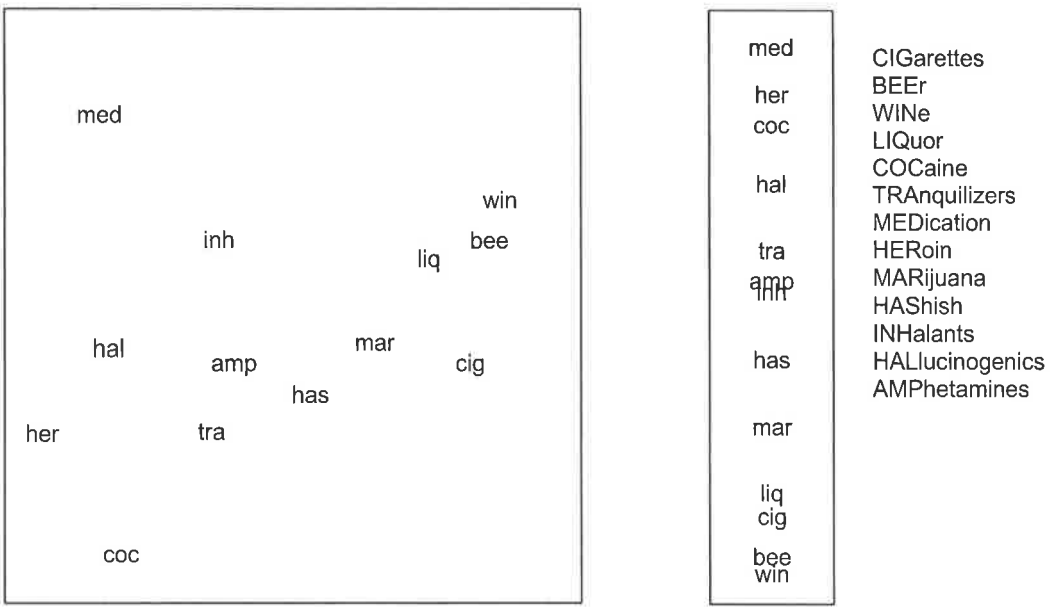

Figure 2.2: Best fitting one (right panel) and two (left panel) dimensional configuration for the drug use data (Huba et al., 1981), explaining 81.6% and 94.0% of the variance respectively. Again, this representation was derived using a gradient descent approach to multidimensional scaling.

largest m eigenvalues during decomposition rather than all r eigenvalues, yielding an $n \times m$ matrix $\mathbf{P}$.

In many cases it is preferable to assume only interval scale data (Stevens, 1946), although the technique detailed above requires ratio scale data. Thus the model implies that the observed dissimilarities differ from the true dissimilarities by a positive constant c, called the additive constant, as well as the measurement error. One approach to the additive constant is to choose the smallest value for c that makes $\mathbf{B}$ positive semi-definite, for which there exists an analytic solution (see Cox & Cox, 1994 for details).

2.2.3 Least Squares Scaling

Least squares scaling (e.g., Greenacre & Underhill, 1982; Lee, 1999) is a more recent development in metric multidimensional scaling, in which an error measure is defined

taking the form,

$$E \propto \sum_{i<j} \left(d_{ij} - \hat{d}_{ij}\right)^2 .$$

This relates to the commonly used Variance Accounted For (VAF) measure of data fit,

$$\text{VAF} = 1 - \frac{\sum_{i<j} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i<j} \left(d_{ij} - \bar{d}\right)^2},$$

where $\bar{d}$ denotes the mean of the observed dissimilarities. Least squares algorithms generally distribute the co-ordinates randomly in the multidimensional space, and then minimise an error measure (or correspondingly, maximise some data fit measure such as the VAF) using numerical optimisation techniques. Gradient descent methods (e.g., More, 1977; Powell, 1977) have generally been used for this purpose, but there is no reason in principle why other methods such as trust region optimisation or quasi-Newton methods (see Nocedal & Wright, 1999, for instance) could not be used. However, as a continuous optimisation problem, metric multidimensional scaling represents a comparatively simple task, and most optimisation methods yield reasonable solutions. An exception to this rule occurs when there is only one dimension: therefore, unidimensional scaling requires that the *ordering* of the stimuli along the dimension be treated as a discrete optimisation problem, separate to the continuous problem of finding co-ordinates (Hubert, Arabie, & Meulman, 1997). Least squares scaling can accommodate an additive constant by redefining the error measure as

$$E \propto \sum_{i<j} \left(d_{ij} - \left(\hat{d}_{ij} + c\right)\right)^2 .$$

This modification allows an optimal value of c to be recovered.

2.2.4 Non-Metric Multidimensional Scaling

The rationale for non-metric multidimensional scaling, and a great many non-metric procedures besides, is provided by an analysis of the process of data collection. If one

assumes that the underlying psychological representation of a given domain is spatial, the relationship between the distances in the space and the dissimilarities in the associational data is not immediately obvious (Shepard, 1962a). Although Shepard (1987) has provided good theoretical grounds for assuming an exponential relationship between distance and dissimilarity, he also observes that the data-gathering process may involve other transformations to the data (Shepard, 1962a). In short, in a number of situations one may wish to represent only the ordinal information present in the data. Non-metric multidimensional scaling is a term referring to the collection of techniques available for deriving spatial representations that fit only the ordering of the dissimilarities, not the magnitude of the differences. That is, if $d_{ij} < d_{kl}$ then a model is only required to ensure that $\hat{d}_{ij} < \hat{d}_{kl}$ to achieve a perfect fit.

Given that monotonicity is the aim of non-metric scaling techniques, it is no longer appropriate to measure fit in terms of the Variance Accounted For, which uses metric information in the data. The first measure of departure from monotonicity was proposed by Shepard (1962a, 1962b). If the dissimilarity ratings and spatial distances are rank ordered, the match between the two orderings can be calculated, as follows: let ij denote a pair of objects that gave rise to the dissimilarity rating d_{ij} and let r denote its rank. Furthermore, let $\hat{ij}$ denote the stimulus pair with spatial distance that also has rank r (which may not be the same pair as ij). Shepard's measure is then given,

$$\frac{2\sum_{r=1}^{n(n-1)/2}\left(d_{ij}-d_{\hat{ij}}\right)^2}{n(n-1)}.$$

One difficulty with this measure is that, although it measures the departure from monotonicity, it uses the metric information in the empirical dissimilarities to do so (Kruskal, 1964a). That is, the error is proportional to $d_{ij} - d_{\hat{ij}}$, and is not merely ordinal. However, the scaling procedure proposed by Shepard (1962a) did not use this error function to minimise the departure from monotonicity, but only to terminate the procedure. In

broad terms, Shepard's algorithm distributes the stimulus locations as an $n-1$ dimensional simplex (the higher dimensional analogue of a tetrahedron), perturbs them slightly to provide a perfect account of the ordering of the observed dissimilarities, and attempts to force the points into a coplanar set without unduly disturbing the ordering, thereby minimising the dimensionality of the solution.

Although the technique proposed by Shepard (1962a) established the qualitative criterion of monotonicity and a procedure for achieving it in a space of minimal dimensionality, it is the work of Kruskal that is generally considered to have established a solid foundation for non-metric multidimensional scaling. In the first of a pair of papers (Kruskal, 1964a), he proposed an explicit measure of the goodness of a non-metric MDS representation, analogous to the VAF, known as *stress*. The stress of a configuration can be intuitively understood by consideration of Figure 2.3, which plots a hypothetical set of dissimilarities against the corresponding distances in an MDS configuration (depicted by stars). It is clear from looking at the stars that the relationship between distances and dissimilarities is not monotonic. There exists a set of numbers x_{ij} that one could substitute for the distances so that $\sum_{i<j}\left(\hat{d}_{ij}-x_{ij}\right)$ is minimal, providing that the relationship is monotonic (this substitution is depicted by the plotted circles). Leaving aside for the moment the method by which the monotonic numbers x_{ij} are calculated, it is clear that the discrepancy between these numbers and the configuration distances $\hat{d}_{ij}$ represents a measure of the extent to which the relationship between the observed dissimilarities d_{ij} and configuration distances $\hat{d}_{ij}$ departs from the ideal of perfect monotonicity. In order that the stress be invariant under dilation of the space, the measure Kruskal proposed is given by,

$$\text{stress} = \min_{x} \sqrt{\frac{\sum_{i<j}\left(\hat{d}_{ij}-x_{ij}\right)^2}{\sum_{i<j}\hat{d}_{ij}^2}} \quad \text{such that } x_{ij} \text{ satisfies monotonicity,}$$

and is the most commonly used non-metric measure of fit. Simple algorithms for finding

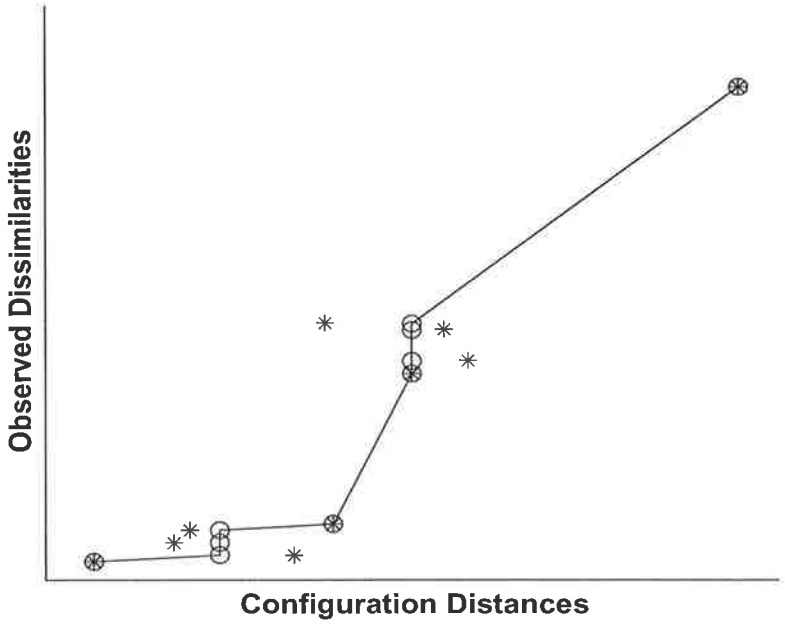

Figure 2.3: Scatterplot of a hypothetical set of dissimilarities against the corresponding distances in an MDS configuration (*). The set of monotonic points used for calculating the stress of this configuration are also shown (o). Based on Kruskal (1964a, fig. 4).

the distances $\hat{d}_{ij}$ that minimise the stress, along with the monotonic numbers x_{ij} are given by Kruskal (1964b). Briefly, Kruskal's scaling algorithm distributes the co-ordinates randomly in a space of some dimensionality, and then uses gradient descent optimisation (e.g., Powell, 1977) to minimise the stress in the same manner that least squares scaling minimise the sum squared error.

The adoption of non-metric measures such as stress introduces substantial changes to the properties of the models. Importantly, the representation need not predict that similarities (or dissimilarities) satisfy the triangle inequality. In fact, the ordinal information in any proximity matrix can always be perfectly accommodated by an $n-1$ dimensional space. However, it has been suggested that non-metric spatial representations should still satisfy (or closely approximate) two constraints known as intradimensional subtractivity and interdimensional additivity (Beals, Krantz, & Tversky, 1968; Tversky & Krantz,

1970). Intradimensional subtractivity requires that the dissimilarity between two stimuli i and j be approximated by some monotonically increasing function of the distance between them on the m dimensions. That is,

$$d_{ij} \approx f\left(|p_{i1} - p_{j1}|, |p_{i2} - p_{j2}|, \ldots, |p_{im} - p_{jm}|\right)$$

where f is the same function for all stimulus pairs. Correspondingly, interdimensional additivity implies that the dissimilarity between two stimuli is a function of the sum of each dimensional component, denoted $\phi(p_{ik}, p_{jk})$ (where $\phi(p_{ik}, p_{jk}) = |p_{ik} - p_{jk}|$ if intradimensional subtractivity is satisfied). Therefore,

$$d_{ij} \approx f\left(\sum_k \phi(p_{ik}, p_{jk})\right).$$

Such requirements act as basic representational assumptions that may be empirically tested when evaluating a non-metric spatial representation.

As a final remark regarding non-metric MDS, Shepard's (1974) observation is important, that non-metric methods are more subject to local minima problems than metric methods. Furthermore, he notes that when the stress of a representation is very low, non-metric representations suffer from degeneracy problems: that is, the optimal solution is non-unique. Therefore there may be many representations that have equivalent and minimal stress but involve different stimulus configurations. If these configurations are substantially different, it may be possible to draw quite different conclusions from these equally good representations.

2.2.5 Other Scaling Techniques

As MDS is a widely-used technique, there are a wide variety of techniques available, only some of which are discussed in detail here. Other approaches allow for spaces with different topologies, such as spheres (Cox & Cox, 1991) or circles in the unidimensional case (Hubert et al., 1997). Alternatively, Lindman and Caelli (1978) argue

that Riemannian spaces with constant curvature may provide psychologically appropriate models in some circumstances, and detail an algorithm for deriving such representations. Finally, in some circumstances it may be appropriate to take individual differences into account. Carroll and Chang (1970) describe a metric MDS model called INDSCAL that extracts both a stimulus space and a subject space. The similarity estimate for a pair of stimuli i and j and the xth participant is given,

$$\tilde{s}_{ijx} = \left(\sum_k w_{xk} |p_{ik} - p_{jk}|^r \right)^{\frac{1}{r}}$$

where w_{xk} is the co-ordinate value for the xth participant on the kth dimension. The subject space, therefore, gives the weightings applied to each dimension for each participant. It is thus possible to characterise the INDSCAL model as one in which there exists a single spatial representation, in which allowance is made for different attention-weights for each individual.

2.3 Featural Representation

Under a featural approach to mental representation, a stimulus is described by a set of attributes or characteristics that it possesses. For example, the features that describe a building might include "tall", "topped by spires" and "has crosses on roof", suggesting it is a church rather than an igloo. Classically, features are treated as all-or-nothing properties: that is, a stimulus either possesses the feature or it does not (e.g., Tversky, 1977). However, some approaches allow features to take on continuous values (e.g., Shiina, 1988), in a manner that resembles dimensions in multidimensional scaling.

Although all the features in the previous example are perceptual in nature, there is no reason why featural stimulus representations cannot incorporate conceptual characteristics. Following the church example, it makes sense to think that a feature such as "place of worship" could form part of people's mental representations of a church, even though

this feature more closely resembles a categorical judgement than a perceptual regularity. As outlined in Chapter 1, this pragmatic view of similarity is acceptable so long as it is acknowledged that, to the extent that it involves high-level processes, similarity is neither cognitively primitive nor a unitary phenomenon.

2.3.1 Notes on Notation

Since there is no standard notation for the mathematics underlying featural representation, a consistent feature notation has been used here. Throughout the thesis, $\mathbf{F}$ denotes a set of features: algebraically, $\mathbf{F}$ is an $n \times m$ matrix whose ikth cell is one if the ith stimulus possesses the kth feature. By adopting this standard, it is convenient to denote the set of features possessed by the ith stimulus as $\mathbf{f}_i$, indicating that this refers to the ith vector of $\mathbf{F}$. Correspondingly, f_{ik} denotes a scalar quantity that is 1 if the ith stimulus belongs to the kth cluster, and 0 if it does not. This system of nomenclature is a useful way of keeping track of a feature structure. In various places $\mathbf{w}$ is used to refer to the vector of feature saliency weights, w_k to refer to a specific saliency weight, and the Greek letters α, β, θ and ρ to refer to hyper-parameters. In keeping with the nomenclature used elsewhere, m and n denote the total number of features and stimuli respectively. The subscripts i and j generally denote an arbitrary stimulus, and k usually denotes an arbitrary feature.

One source of conflict with this notation is that is common to refer to the functional forms of various models using f and g. In order to avoid any potential confusion, this convention has been abandoned. Lower case Greek letters are avoided for similar reasons. Since there is some relationship between the hyper-parameters α, β, θ and ρ and these functional forms, there is a weak case to be made in favour of using uppercase Greek letters for the functions. This is in fact the approach adopted, using Λ and Υ for this purpose, but the true reason for this is that these symbols are visually distinct from

the other symbols used to denote featural representation.

2.3.2 Featural Similarity Models

Tversky (1977) proposed two general featural models that describe the similarity between two stimuli in terms of the features they share (common features) and the features that distinguish between them (distinctive features). If $\mathbf{f}_i \cap \mathbf{f}_j$ denotes the features common to the ith and jth stimuli, and $\mathbf{f}_i - \mathbf{f}_j$ denotes the features possessed by the ith but not the jth stimulus, then the first of these models, known as the Contrast Model, is given by

$$\hat{s}_{ij} = \theta\Lambda(\mathbf{f}_i \cap \mathbf{f}_j) - \alpha\Lambda(\mathbf{f}_i - \mathbf{f}_j) - \beta\Lambda(\mathbf{f}_j - \mathbf{f}_i), \tag{2.2}$$

where Λ is a monotonically decreasing function and θ, α, and β are non-negative parameters that assign weights to each of the terms. The Contrast Model assumes that similarity is a linear function of the common and distinctive features components. Tversky's Contrast Model is discussed in some detail in Chapter 4, and a Modified Contrast Model is considered, that has some advantages over Tversky's. Alternatively, Tversky proposed the Ratio Model,

$$\hat{s}_{ij} = \frac{\Lambda(\mathbf{f}_i \cap \mathbf{f}_j)}{\Lambda(\mathbf{f}_i \cap \mathbf{f}_j) + \alpha\Lambda(\mathbf{f}_i - \mathbf{f}_j) + \beta\Lambda(\mathbf{f}_i - \mathbf{f}_j)}, \tag{2.3}$$

in which similarity is given by the ratio of the common features term to the sum of the common and distinctive features terms. However, neither the Contrast Model nor the Ratio Model are generally employed by the clustering algorithms used to analyse similarity matrices. Some clustering algorithms do not explicitly fit any well-specified psychological model, whereas others implement special cases of Tversky's models. The following sections discuss hierarchical clustering methods and additive clustering algorihms.

2.3.3 Hierarchical Clustering

A large number of clustering procedures exist for finding hierarchically organised feature structures (see Hartigan, 1975 or Johnson, 1967, for instance). In these representations, any two clusters are either disjoint (do not have any common stimuli) or nested (one is a strict subset of the other). It is a simple matter to demonstrate that this is sometimes an unduly restrictive constraint. For example, the features of "red" and "big" are both psychologically plausible, yet are clearly neither disjoint nor nested: there are big red things, small red things, big non-red things and small non-red things. However, hierarchical clustering procedures are less computationally expensive than non-hierarchical procedures. Furthermore, there are many situations in which the hierarchical constraint is highly plausible (discussed shortly with regard to trees).

Most hierarchical clustering procedures begin by assigning each stimulus to a trivial "cluster" to which it is the sole member. At each stage of the algorithm the two most similar clusters are united. This process continues until only one cluster remains, containing all of the stimuli. The final clustering solution consists of all the clusters found during the procedure (excluding the trivial clusters containing one stimulus). The difference between the various procedures regards how the similarity between two clusters is calculated. In single-link clustering (Sneath, 1957; Sokal & Sneath, 1963), the similarity between two clusters is given by the greatest similarity of any member of one cluster to any member of the other cluster (a maximum similarity rule). Contrastingly, in complete-link clustering (see Johnson, 1967) the similarity between two clusters is equal to smallest similarity between stimuli in different clusters. Other hierarchical clustering schemes use the average or median similarity between stimuli in different clusters (see D'Andrade, 1978; Hartigan, 1975).

2.3.4 Additive Clustering

Additive clustering (Shepard & Arabie, 1979) is a framework for deriving featural representations in which the similarity between two stimuli is given by the sum of the weights of their shared features. That is,

$$\hat{s}_{ij} = \sum_k w_k f_{ik} f_{jk}, \tag{2.4}$$

where f_{ik} is a binary-valued cluster membership variable and w_k is the cluster weight. If $\hat{\mathbf{S}} = [\hat{s}_{ij}]$ denotes the $n \times n$ matrix of similarity-estimates under the model, $\mathbf{F} = [f_{ik}]$ denotes the $n \times m$ matrix whose binary valued cells indicate which objects belong to which clusters, and $\mathbf{W}$ is an $n \times n$ matrix whose main diagonal entries contain the cluster weights $\mathbf{w} = \{w_1, w_2, \ldots, w_m\}$ and whose other elements are zero, then the additive clustering model can be written,

$$\hat{\mathbf{S}} = \mathbf{F}'\mathbf{W}\mathbf{F}$$

where $\mathbf{F}'$ denotes the matrix transpose of $\mathbf{F}$. The framework has considerable status as a model of human conceptual structure as a special case of Tversky's (1977) Contrast Model, namely a common features model ($\theta = 1$, $\alpha = \beta = 0$) with an additive functional form for Λ.

Additive clustering differs from hierarchical clustering in that there is no requirement that the cluster structure be nested. This flexibility affords representational possibilities that hierarchical clustering schemes lack, as demonstrated in the two additive clustering representations shown in Tables 2.1 and 2.2. In both tables the Variance Accounted For (VAF) is reported. When analysing similarity matrices, this takes the form,

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

Table 2.1: The six-feature additive clustering representation of the drug use data (Huba et al., 1981) preferred by the Stochastic Complexity measure (see Section 3.1) for most reasonable precision assumptions. This representation was found using a stochastic hillclimbing algorithm resembling Lee's (in press) approach, and is an original analysis.

Feature	Weight
Cigarettes, Beer, Wine, Liquor, Marijuana, Hashish	0.257
Beer, Wine, Liquor	0.175
Cocaine, Tranquilizers, Medication, Heroin, Marijuana, Hashish, Inhalants, Hallucinogenics, Amphetamines	0.151
Tranquilizers, Hashish, Hallucinogenics, Amphetamines	0.138
Cigarettes, Beer, Wine, Liquor, Tranquilizers, Marijuana, Inhalants, Amphetamines	0.122
Liquor, Cocaine, Tranquilizers, Heroin, Hashish, Inhalants, Hallucinogenics, Amphetamines	0.073
Additive Constant	0.026
Variance Accounted For	92.2%

where $\bar{s}$ denotes the arithmetic mean of the empirical similarities. Both of these representations account for most of the variance in the data, which is made possible by allowing the features to overlap arbitrarily. For example, five of the eight features in Table 2.2 capture magnitude-related characteristics of the ten numbers (e.g., "big numbers", "small numbers", etc.). The remaining features capture mathematical properties, such as "powers of two" and "multiples of three". However, in order to represent these two types of characteristics, the features must overlap arbitrarily since, for instance, the powers of two are neither disjoined from nor nested within the magnitude-related features.

Table 2.2: The eight-feature additive clustering representation of the Arabic numeral domain (Shepard et al., 1975) data reported by Tenenbaum (1996).

Feature										Weight
		2		4				8		0.444
0	1	2								0.345
			3			6			9	0.331
						6	7	8	9	0.291
		2	3	4	5	6				0.255
	1		3		5		7		9	0.216
	1	2	3	4						0.214
				4	5	6	7	8		0.172
Additive Constant										0.148
Variance Accounted For										90.9%

ADCLUS

The original additive clustering algorithm (ADCLUS; Shepard & Arabie, 1979) employed a heuristic method to reduce the space of possible cluster structures to be searched. Shepard and Arabie observed that a subset of the stimuli in the domain is most likely to constitute a feature if the pairwise similarities of the stimuli in the subset are high. They define the s-level of a subset, C, to be the lowest pairwise similarity rating for two stimuli within the subset. Further, the subset C is *elevated* if and only if every larger subset that contains C has a lower s-level than C. The ADCLUS algorithm consists of two distinct stages. In the first step, all elevated subsets are found. In the second step, the saliency weights are found and the subset further reduced. The weight initially assigned to each potential cluster is proportional to its *rise*, defined as the difference between the s-level of the subset and the minimum s-level of any subset containing the

original subset. The weights are iteratively adjusted by using the first partial derivatives of the VAF with respect to each elevated subset, denoted $\frac{\partial v}{\partial w_k}$, and given by the formula,

$$\frac{\partial v}{\partial w_k} = 2\frac{\sum_{i<j,\in C_k}(s_{ij} - \hat{s}_{ij})}{\sum_{i<j}(s_{ij} - \bar{s})}$$

where C_k denotes the set of all pairings of the objects that comprise subset k. At each iteration a small step is taken in the direction of $\frac{\partial v}{\partial w_k}$, which is then updated. Deletion of subsets also occurs iteratively, where any subset whose weight falls below a preset criterion is removed from the list of candidate clusters. This iterative process continues until the length of the gradient vector falls below a specified level.

Although the elevated subsets heuristic has become unnecessary due to advances in computing technology, it is interesting to observe that the ADCLUS algorithm performs fairly well in comparison to modern algorithms, and produces reasonable solutions to overlapping clustering problems.

MAPCLUS

The MAPCLUS algorithm (Arabie & Carroll, 1980) employs a mathematical programming approach to the clustering optimisation problem, by embedding the discrete optimisation in a continuous one. The cluster membership matrix $\mathbf{F}$ is initially allowed to assume continuously varying values, rather than the binary membership values required in the final solution. An error function is defined as the weighted sum of two parts, the first being the sum squared error and the second being a penalty function designed to push the elements of $\mathbf{F}$ towards 0 or 1. MAPCLUS requires the number of features to be specified in advance.

Expectation Maximisation

A statistically principled approach suggested by Tenenbaum (1996) is Expectation Maximisation (EM). Under this probabilistic approach, the observed similarities are assumed

to be drawn from Gaussian distributions with a common variance σ. The EM algorithm for additive clustering requires the number of clusters to be specified in advance, and consists of an alternating two-step procedure. In the E-step, the saliency weights are held constant, and the expected sum squared error is estimated, with the aim of calculating the "energy function",

$$-\frac{1}{2\sigma^2} E\left[\sum_{i<j}(s_{ij} - \hat{s}_{ij})^2\right].$$

Since the observed similarities and the saliency weights are constant in the E-step, the expectation is taken only over the elements of the feature matrix $\mathbf{F}$. Furthermore, the probability of a given feature matrix is proportional to

$$\exp\left(-\frac{1}{2\sigma^2}\sum_{i<j}(s_{ij} - \hat{s}_{ij})^2\right).$$

Therefore, as Tenenbaum (1996) observes, at small σ values, only the more probable feature matrices make a substantial contribution to the expected sum squared error. Using the expected values for the cluster membership values that are calculated during the E-step, the M-step finds a new set of saliency weights that minimise the expected sum squared error. As the EM algorithm iterates, the value of σ is reduced, and the expected values for the cluster membership values f_{ik} converge on 0 or 1, yielding a final feature matrix $\mathbf{F}$ and saliency weights $\mathbf{w}$.

Stochastic Hillclimbing

Lee (in press) has proposed a simple stochastic hillclimbing algorithm that "grows" an additive clustering model. The algorithm initially specifies a single-cluster representation, which is optimised by "flipping" the elements of $\mathbf{F}$ (i.e., $f_{ik} \rightarrow 1 - f_{ik}$) one at a time, in a random order. Every time a new feature matrix is generated, best-fitting saliency weights $\mathbf{w}^*$ are found by solving the corresponding non-negative least squares problem (see Lawson & Hanson, 1974), and the solution is evaluated. Whenever a

better solution is found, the flipping process restarts. If flipping f_{ik} results in an inferior solution, it is flipped back. If no element of $\mathbf{F}$ can be flipped to provide a better solution, a local minimum has been reached. Since, as Tenenbaum (1996) observed, additive clustering energy landscapes tend to be riddled with local minima, Lee's algorithm allows the locally optimal solution to be "shaken", by randomly flipping several elements of $\mathbf{F}$ and restarting, in order to find a globally optimal solution. Once this process terminates, a new (randomly generated) cluster is added, and this solution is used as the starting point for a new optimisation procedure. Importantly, Lee evaluates a solution using the Stochastic Complexity measure (Rissanen, 1996, see Section 3.1), which provides a statistically-principled method for determining the number of clusters to include in the representation, since the Stochastic Complexity will deteriorate as the representation becomes too complex.

2.4 Tree Representation

Tree models of similarity judgements represent each stimulus by terminal nodes in a connected, acyclic graph (see Figures 2.4 and 2.5, for instance) and measures the dissimilarity between two stimuli by the length of the unique path that connects them. The basic representational assumption made by tree models is that the stimuli are hierarchically organised. For instance, Pinker (1994, p. 469) argues that human judgements regarding natural kinds (e.g., plants and animals) are inherently hierarchical, reflecting an adaptation to a hierarchical structure in the environment. Therefore, as Corter (1996, p. 52) remarks, "any set of objects that has arisen through an evolutionary process of 'splitting' or successive differentiation is likely to be modeled successfully by some sort of tree model".

Along similar lines, Tversky and Hutchinson (1986) define two measurements for a similarity matrix, called centrality and reciprocity. Centrality is high when most stimuli

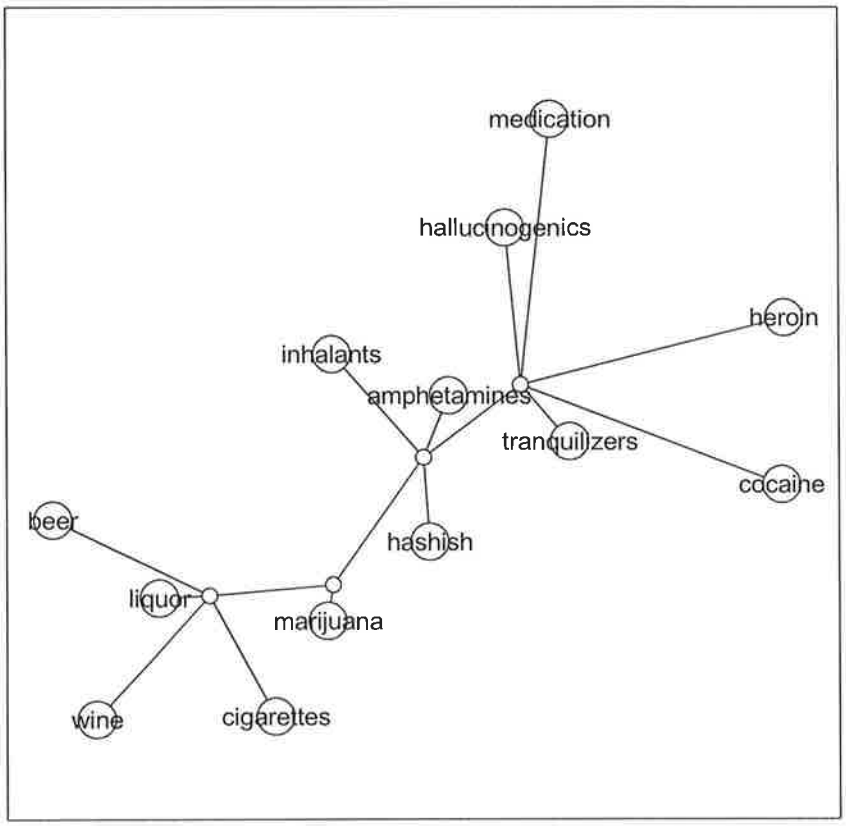

Figure 2.4: The four-node tree preferred by the BIC for the drug use data (Huba et al., 1981) under reasonable precision estimates, explaining 86.8% of the variance. This representation was found using Lee's (submitteda) successive differentiation algorithm and is an original analysis.

share the same "nearest neighbour" (i.e., the stimulus that they are most similar to). Reciprocity is high when the nearest neighbour relation tends to be symmetric (i.e., if i is j's nearest neighbour, then j is i's nearest neighbour). They note that tree representations typically display high centrality and low reciprocity, and argue that trees are therefore appropriate representations for stimulus domains displaying the same pattern. Finally, Pruzansky, Tversky, and Carroll (1982) suggest that additive trees appear more appropriate for conceptual domains, whereas spatial models tend to be more suited to perceptual stimuli.

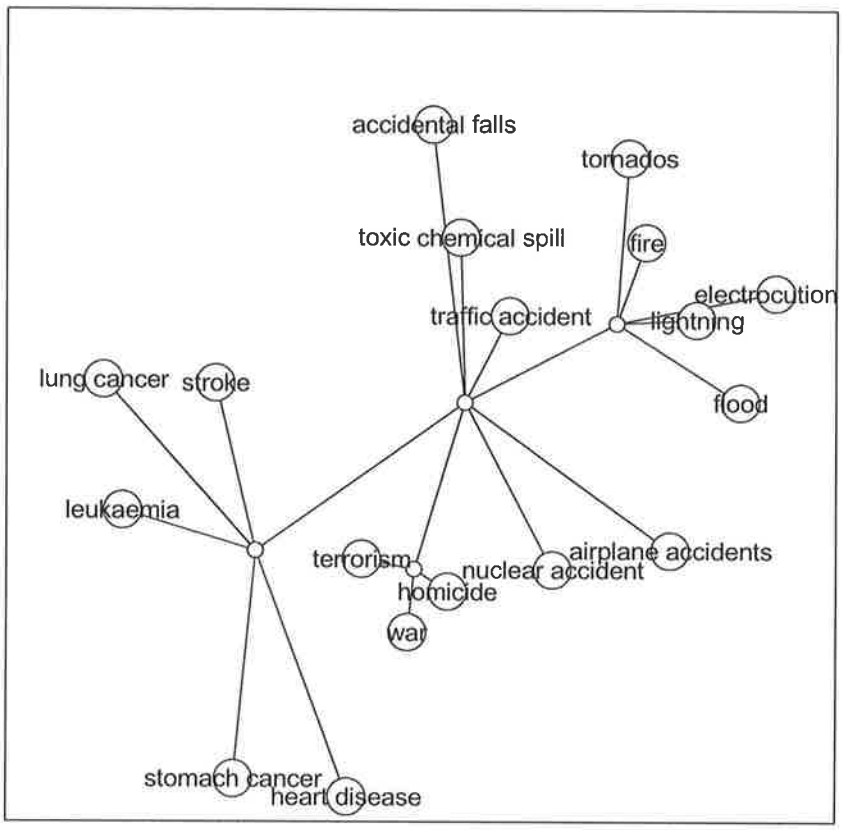

Figure 2.5: Best fitting four-node tree for Johnson and Tversky's (1984) risk data, explaining 62.5% of the variance. This representation was found using Lee's (submitteda) successive differentiation algorithm, and is an original analysis. Compare Lee's (1999) five-node tree for the same data set.

2.4.1 Types of Trees

The simplest type of tree is the star tree, also called the singular tree. In a star tree, all stimuli are represented as terminal, or *leaf*, nodes connected to a single internal (*stem*) node. The arcs connecting each stimulus to the stem node may be of any length. Star trees satisfy the following relationship:

$$\hat{d}_{ij} + \hat{d}_{kl} = \hat{d}_{ik} + \hat{d}_{jl} = \hat{d}_{il} + \hat{d}_{jk}. \ \forall i, j, k, l.$$

Consequently, star trees are highly constrained, and rarely constitute useful models of

mental representation. In contrast, ultrametric trees may include multiple internal nodes. However, ultrametric trees are constrained so that every stimulus is equidistant from one of the internal nodes, called the *root*. The corresponding *ultrametric inequality* is given

$$\hat{d}_{ij} \leq \max\left(\hat{d}_{ik}, \hat{d}_{jk}\right) \ \forall i, j, k.$$

It is noteworthy that ultrametric trees are formally equivalent to hierarchical clustering schemes, and therefore can be interpreted as common features representations according to Tversky's (1977) Contrast Model (Corter & Tversky, 1986).

The most widely-used tree-based similarity model is the additive tree model (Sattath & Tversky, 1977). Additive trees relax the ultrametric constraint, allowing stimuli to be an arbitrary distance from the root node. Accordingly, additive trees are constrained by the additive inequality (Buneman, 1971),

$$d_{ij} + d_{kl} \leq \max\{d_{ik} + d_{jl}, d_{jk} + d_{il}\} \ \forall i, j, k, l. \tag{2.5}$$

As Corter (1996) notes, this means that there is no longer any compelling justification for the choice of any one point as "the" root for an additive tree. Arguably, additive trees make more sense in an unrooted than a rooted form. Two examples of unrooted additive tree representations are shown in Figures 2.4 and 2.5.

Additive trees can also be interpreted as featural models, but in a different manner to ultrametric trees. As observed by Corter (1996), the natural interpretation of an unrooted additive tree is as a distinctive features model according to Tversky's (1977) Contrast Model, by including distinctive features for every edge in the tree, with saliencies equal to the edge lengths. As special cases of additive trees, one could interpret ultrametric and star trees in the same manner. Alternatively, Corter notes that a rooted additive tree with m internal nodes can be treated as the sum of an ultrametric tree and a star tree. If so, the ultrametric tree can be interpreted as $m - 1$ hierarchically organised common

features, and the singular tree as n distinctive features each containing only the one stimulus.

Two extensions of the additive tree model should be mentioned. Firstly, Cunningham (1978) proposed a bidirectional additive tree model, in which some or all of the edges in the tree may be asymmetric. That is, some of the edges may be longer going from A to B than from B to A. This allows trees to accommodate asymmetric similarity judgements, when $s_{ij} \neq s_{ji}$. If this approach is adopted, it is important to recognise that asymmetric edges constitute two free parameters.

Secondly, Corter and Tversky (1986) describe an extended tree model, which derives an additive tree plus a set of common features. Some segments on the additive tree are "marked" by a common feature: these segments do not count towards the dissimilarity of stimuli that both possess the common feature. Although Corter and Tversky (1986) demonstrate that this "tree plus exceptions" model can fit data that are inherently non-hierarchical, it is not clear how the extended tree is to be interpreted psychologically. If the tree is treated as a set of features, it is implausible to represent the distinctive features by a different psychological structure (the tree) to the common features (the exceptions). It would be preferable to adopt a single representational framework (features) that simultaneously captured common and distinctive components. It is worth noting that Tversky's (1977) Contrast Model does not allow this, since it assigns a single weighting of common and distinctive features applied to all features. However, in Chapter 4 a featural model is introduced (the Modified Contrast Model) that allows for common features and distinctive features to be extracted in the one representation.

2.4.2 Growing Additive Trees

Several different algorithms have been proposed for finding tree structures in dissimilarity data. This section outlines four of these approaches.

ADDTREE

Sattath and Tversky's (1977) ADDTREE algorithm begins by linearly transforming the observed dissimilarities, adding a constant to each dissimilarity large enough to ensure that the triangle inequality is always satisfied. Then the "neighbourhood scores" are calculated for all pairs of stimuli. The neighbourhood score N for a pair of stimuli is initially zero, but is incremented according to the following rule for all i, j, k and l: if

$$d_{ij} + d_{kl} < d_{ik} + d_{jl} = d_{il} + d_{jk}$$

then both $N(i,j)$ and $N(k,l)$ are increased by two. However, if

$$d_{ij} + d_{kl} < d_{ik} + d_{jl} < d_{il} + d_{jk}$$

then $N(i,j)$ and $N(k,l)$ are increased by two, and $N(i,k)$ and $N(j,l)$ are increased by one. The simplest version of ADDTREE-style algorithm would "merge" two stimuli if and only if they had the highest neighbourhood score. These stimuli would then be attached to the same internal node on the tree, and the corresponding rows and columns in the proximity matrix averaged. This process would repeat until all stimuli have been connected to the tree. Both Sattath and Tversky's (1977) algorithm and Corter's (1982) modification employ variants on this idea, but allow for multiple mergings on each iteration. However, since Abdi, Barthelemy, and Luong (1984) and Barthelemy and Guenoche (1991) have demonstrated that this approach generally performs quite poorly, it is not discussed further.

Alternating Least Squares

The alternating least squares algorithm developed by Carroll and Pruzansky (1975, 1980) uses the observation that an additive tree can be decomposed into a star tree and an ultrametric tree. Accordingly, their algorithm initially fixes the lengths of the star tree

arcs, and uses hierarchical clustering to find an ultrametric tree. In the next step, the ultrametric tree is held constant and the star tree arc lengths are rescaled. These two steps are repeated until the star tree and the ultrametric tree stabilise. The additive tree is then recovered by adding the star and ultrametric trees.

Mathematical Programming

A mathematical programming approach is adopted by de Soete (1983), treating additive tree scaling as a continuous optimisation problem applied to the proximity matrix. The optimisation function contains a badness-of-fit error term and a penalty term designed to push the derived matrix towards satisfying the additive inequality. The result, therefore, is a matrix that satisfies the additive inequality but has as little discrepancy from the raw data as possible. Once this is complete, the additive tree that perfectly accounts for the derived matrix is recovered.

Successive Differentiation

Lee's (submitteda) successive differentiation algorithm initially finds the best-fitting star tree, and adds internal nodes to the tree one by one. The goodness of any tree is calculated using the Bayesian Information Criterion (BIC; see Section 3.1.2) which provides a trade-off between data-fit and model complexity. As a result, the process of adding nodes stops once the BIC starts to deteriorate. This deterioration occurs once the improvements to data-fit no longer justify the increase in model complexity that occurs when an extra node is added. Therefore, a tree is sought that optimises the BIC.

Nodes are introduced by a process of "splitting". Each internal node in the tree is associated with an error E given by

$$E(k) = \sum_{i \in C_k} \sum_{j \neq i} \left(d_{ij} - \hat{d}_{ij} \right)^2$$

where C_k denotes the set of stimuli (i.e., terminal nodes) that are connected by an edge to the kth internal node. When a new node is added, it is connected to the internal node with the greatest associated error (the "old node") and some of the nodes (both internal and terminal) that were connected to the old node are moved onto the new node. A stochastic hillclimbing approach to combinatorial optimisation is used to determine which nodes to move, with the aim of finding the configuration with the best BIC value.

2.5 Network Representation

Network models of similarity data represent stimuli as nodes in a graph, connected to one another by edges. Unlike additive trees however, network models may contain cycles, but do not possess non-stimulus nodes. A graph is defined as a set of connections, and thus describes a set of ways in which one can go from one node to another. It is therefore possible to interpret a graph as describing a set of "cognitive paths". Graph-theoretic models of similarity data were first canvassed in a paper by Harary (1964), which considered the embedding of basic concepts such as adjacency, equivalence and betweenness. The paper also proposed a fairly simple method for generating a representation, by including a link between two nodes if their similarity is sufficiently high. However, this method does not define any measure of fit, which makes it difficult to evaluate the representation. To the author's knowledge, this method has never been implemented.

A commonly adopted network similarity model is to assign a length to each edge in the graph, and estimate the similarity between two stimuli by the length of the shortest path that connects them (Klauer & Carroll, 1991). This "minimum path" similarity model satisfies the triangle inequality (Hakimi & Yau, 1965; Goldman, 1966). This model has been fit by a number of non-metric algorithms, such as Feger and Bien's (1982) network unfolding algorithm, Orth's (1988) monotonic network analysis, and Klauer's (1988, 1989) ordinal network representation, as well as Klauer and Carroll's

(1989, 1991) metric MAPNET algorithm and Hutchinson's (1989) NETSCAL algorithm that uses both metric and non-metric properties.

A representational assumption made by all of the network algorithms mentioned above is that the derived graph must be connected. That is, for all pairs of nodes i and j, there must be a path that connects the two. Psychologically, this amounts to the assumption that the domain under examination is in some way homogeneous. For instance, if the domain in question consisted of the concepts "red", "green", "blue", "chair", "table", "stool", it does not make sense for the derived graph to be connected. Logically, there should be two distinct connected graphs, consisting of colours on the one hand, and furniture on the other. However, given that most domains that one considers are likely to be homogeneous, this assumption will rarely cause problems.

2.6 Alignment-Based Similarity Models

Similarity is closely related to analogy, and it is from analogical reasoning, and the idea of *structure-mapping* (Gentner, 1983) that the alignment-based approach draws inspiration. It has previously been argued (Markman & Gentner, 1993) that similarity judgements involve a process of finding correspondences between stimuli that satisfy relational constraints. Therefore, an octopus' tentacle is more likely to be matched with a human arm than with a gigantic udon noodle[1], because the tentacle and the arm serve similar functions, and are physically attached to an animal in a similar manner. Indeed, Markman and Gentner (1993) and Goldstone (1998) report studies in which participants appeared to prefer making similarity judgements on the basis of relational structure rather than surface attributes, though they acknowledge the artificiality of the studies, since – like perceptual and conceptual features – relational structures tend to correlate with surface features in real-world environments.

[1]Though hopefully not if food is the topic of conversation.

The notion of matching relational elements as well as perceptual attributes recalls the distinction between perceptual and conceptual similarities, and indeed there is some correspondence between the two. As Goldstone (1998) observes, one can posit featural representations that include features for relational elements. However, he argues that this approach can rapidly lead to a proliferation of features, and fails to make the fundamental distinction between object components and their relational structure.

In the context of analogical reasoning, there are a number of models that take such an approach, including the Structure-Mapping Engine (SME; Falkenhainer, Forbus, & Gentner, 1989; Forbus & Oblinger, 1990) and the Analogical Constraint Mapping Engine (ACME; Holyoak & Thagard, 1989). In this tradition, Goldstone proposes a connectionist model called Similarity as Interactive Activation and Mapping (SIAM; Goldstone, 1994, 1998; Goldstone & Medin, 1994), which takes low-level representational objects such as features and dimensions, along with the structural relations between them, and attempts to map one stimulus onto another. The SIAM network is built up of a large number of "hypothesis nodes": the activation level of a node represents the extent to which the network "wants" to map one aspect of the ith stimulus onto another aspect of the jth stimulus. Connections between pairs of nodes are excitatory if the mappings they denote are consistent, and inhibitory if they are inconsistent (e.g., if they map a single aspect of i onto multiple aspects of j). Initial activation levels are set according to the perceptual similarity of the low-level features.

One advantage of SIAM is that it explicitly considers the process by which similarity judgements are made. However, it does specify highly elaborate representations on the basis of intuitive reasonableness, rather than basing them on evidence supplied by empirical data, which suggests some caution when evaluating the model. Furthermore, SIAM is justified by Goldstone (1994, 1998) largely in terms data fit, without considering issues such as model complexity. Overall, the alignment-based approach shows some

merit, but is as yet less fully developed than other representational theories.

2.7 Transformational Similarity Models

The transformational approach to mental representation is perhaps the least clearly articulated of the various frameworks. At its core is the notion, expressed by a number of authors, that the relationship between two stimuli can be characterised by a set of transformations that carry one stimulus into the other (e.g., Carlton & Shepard, 1990a; Freyd, 1987; Hoffman, 1966; Lewin, 1936; Leyton, 1992; Vickers, 1996, 2002). This idea can be formalised in any number of ways, and there is substantial variation amongst psychologists regarding how best to do so. Consequently, it makes little sense to refer to "the" transformational approach. Furthermore, few transformational theories are similarity models. Consequently, this discussion of transformational similarity is necessarily speculative, resembling a sketch of a future modelling framework rather than a summary of an existing one.

Empirical evidence favouring a transformational approach of some kind is provided by Imai (1977, 1992), who presented participants with simple linear patterns such as XXXOOXOO or OOOOOXXX. Their judgements demonstrated that the similarity of such patterns is sensitive to reflections (e.g., OOOOOXXX → XXXOOOOO), inversions (e.g., XXXOOOOX → OOOXXXXO), phase shifts (e.g., OOXXOOOO → OOOXXOOO) and changes of scale (e.g., OOXXOOXX → OXOXOXOX). These findings are naturally interpreted by supposing that people rate stimuli as more similar when they are related by simple, psychologically plausible transformations of this kind. This claim lies at the heart of the transformational approach. The difficulty lies in specifying transformations appropriate to the circumstances, and finding a general formalism in which to express the idea.

In an ambitious and broad-reaching endeavour, Leyton (1986a, 1986b, 1992) proposes a very general transformational theory of cognition. Central to this theory is the

proposition that a stimulus is naturally perceived (i.e., represented) as having resulted from a set of generative transformations applied to a basic object. These transformations introduce asymmetries into the basic object: the representation of a stimulus, according to Leyton, should capture these asymmetry-inducing transformations, called it's "causal history". For example, a parallelogram is perceived as a bent rectangle, a rectangle as a stretched square, a square as four rotated lines, and a line as a displaced point (Leyton, 1985). The underlying mathematics of Leyton's theory have been strongly criticised by Hendrickx and Wagemanns (1999), so this discussion is restricted to broad qualitative ideas.

Leyton (1989) generalises the notion of causal history to stimulus comparisons with his theory of shape transformations, which describes the causal history of "blobby shapes" (i.e., closed contours with no corners). This theory details a variety of transformations such as protrusion, indentation and bifurcation by which one shape can be morphed into another. The theory is intuitively plausible, although it only allows for the sequential operation of processes. For example, in order to transform a circle into a starfish-shaped blob, five protrusions must be produced. Leyton's theory requires them to be introduced one by one, even though it is equally plausible to think that they grow at the same time.

This qualification notwithstanding, the theory permits a transformational interpretation of similarity. It is possible to represent a set of blobs as nodes in a graph with edges linking those shapes that differ only by single transformation (see Leyton 1992, Figure 2.16, for instance). The dissimilarity between two blobs could be measured as the number of transformations required to connect the two, emphasising the qualitative significance of *types* of transformations. Alternatively, the length of each edge could be specified by the "extent" of the transformation required to mutate one blob into its neighbour. The dissimilarity between two blobs would then be given by the length of

the shortest path connecting them (making it a network representation in the sense described earlier). This approach emphasises "transformational distance" over the number of transformations required. Furthermore, it is now possible to include three blobs that differ only in the extent to which a transformation is applied (e.g., a circle stretches into an ellipse, which in turn stretches into a more eccentric ellipse). Interestingly, as more blobs are added to the stimulus set, the graph becomes more dense, forming a continuum in the limit. The axes of the resulting spatial representation can therefore be interpreted as transformations.

What metric should apply in this space? In Leyton's theory, transformations can only be applied one at a time, and the edges in the graph therefore represent single transformations. The result of this independence is that the city-block distance is the natural metric in the space. However, if transformations are permitted to occur simultaneously, then the graph would contain edges that directly connect stimuli that are distinguished by multiple transformations. The same co-ordinate space would result, but the Euclidean metric would prevail.

Another spatial-based transformational approach can be found in Carlton and Shepard's (1990a, 1990b) explanation of apparent motion: when a stimulus is displayed in two different positions and orientations a short time apart, people perceive it as having moved between the two locations along a path. Carlton and Shepard describe the location of an object in terms of its location in three-dimensional Euclidean space (E^3), and its orientation as co-ordinates on the surface of a three-dimensional sphere (S^3). They then observe that the path of apparent motion follows a geodesic path in the resulting six-dimensional manifold ($E^3 \times S^3$).

Transformational ideas have the potential to inform representational theory in other ways. Hoffman's (1966, 1968, 1970, 1984, see also Dodwell, 1983 and Hoffman & Dodwell, 1985) "geometric psychology" is a transformational approach based on group

theory (see Armstrong, 1988 for instance), applied mainly to visual perception. The theory is largely concerned with perceptual invariants such as size invariance (where an object is perceived as being the same size regardless of how far away it is). The theory is based on Lie groups, which are generated by infinitesimally-small transformations: for example, any translation in the x-dimension can be produced by repeated applications of the infinitesimal translation $\frac{\partial}{\partial x}$. Taken together, $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ generate a two-dimensional space. However, a two-dimensional space is also generated by the rotation $-y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}$ and the dilation $x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}$. The difference between the two emerges when considering distance metrics. The city-block metric in the first space adds displacement in x to displacement in y to calculate distance. In the second case, however, city-block distance is calculated as by summing the length of the rotational arc and the length of the dilation. As it happens, this corresponds to the distinction between using Cartesian co-ordinates and Polar co-ordinates. The transformational approach highlights the importance of considering the manner in which a spatial representation is generated, since it can substantially affect the metric. Furthermore, if stimulus dimensions are subject to an attention-weighting process (e.g., Kruschke, 1992; Nosofsky, 1986), it is important to consider whether attention shifts from the x-dimension to the y-dimension, for instance, or from the rotation component to the dilation component.

Another regard in which the choice of generating transformations is relevant is demonstrated by Feldman's (1997) theory of perceptual categories. According to this theory, a single observed stimulus is considered to be a member of a category defined by a consequential region in a psychological space in the sense referred to be Shepard (1987). This region is described as the manifold that is generated using the smallest set of transformations that can generate the stimulus. Therefore, people generalise from a square to other squares, but not to all rectangles. Importantly, different transformation sets yield different generalisations (see also Weiner-Erhlich, Best, & Millwood, 1980).

The manifold of squares that is embedded in the space of rectangles will only emerge as a natural generalisation if the dilation transformation $x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}$ is among the set of transformations available. In fact, this theory suggests that a two-dimensional space could well be characterised by many transformations, even though only two are required to generate it.

This discussion of transformational similarity has largely been concerned with considering spatial (and to a lesser extent network) representations. However, it may be possible to provide an account of featural representation in terms of discrete groups, though this is not done here. It may be that transformational ideas have the potential to provide a unified framework for modelling similarity, but at this stage such an account has not been developed.

2.8 Summary & General Discussion

Over the last 50 years or so, a wide variety of frameworks have been proposed for modelling similarity. Spatial and featural approaches have been extensively studied. So, to a lesser degree, have tree and network representations. In contrast, alignment-based representation is a new approach to similarity modelling, and transformational ideas have yet to be formalised as a modelling framework. Each of these approaches makes different assumptions about the nature of similarity judgements, and it fair to say that each has its strengths and weaknesses. Correspondingly, it makes sense to choose the representational framework best suited to the problem at hand, rather than attempting to apply a single approach on every occasion.

3. On Representational Complexity

The psychological theories discussed in the previous chapter address an important question, namely "What kinds of structures make sense as models of human mental representation?" In practice, however, a researcher who has just collected a set of similarity data is faced with a somewhat different question, "What is the specific representation that most probably gave rise to this particular data set?" To answer this, the researcher must address a psychological problem and a numerical problem. The psychological problem requires the researcher to select a representational framework (e.g., featural representation) and a specific similarity model (e.g., the common features model used by additive clustering).

Once a similarity model (say, common features) has been chosen, the numerical problem is immediately apparent. The common features similarity model does not specify which set of features $\mathbf{F}$ to use, nor what saliency weights $\mathbf{w}$ to apply. Indeed, the number of common features representations that *might* have given rise to the data is astronomical. Thus the numerical problem, simply put, is the task of finding the best representation from the set of all representations possible under the similarity model.

This problem has two aspects. The first part, which might be called the *algorithmic problem*, involves finding an efficient and effective procedure for searching through the set of possible representations. This is an optimisation problem, and each similarity model constitutes a different problem that may call for a different solution. Consequently, there is little need for a general discussion of numerical optimisation techniques in this

thesis. Instead, whenever a new optimisation problem arises, an algorithm is developed, and an evaluation of its performance is presented.

The purpose of this chapter is to address the second aspect, the *model selection problem*. Again, this problem can be succinctly stated: "What makes one representation better than another?" The model selection issue has been discussed in general terms by a number of authors (e.g., Collyer, 1985; Myung, 2000; Pitt, Myung, & Zhang, 2002; Roberts & Pashler, 2000), and although there is little consensus on the right way to choose between models, there is considerable agreement on the wrong way to do so. The practice, common in psychology, of selecting the model with the greatest data-fit (e.g., the highest VAF), is almost universally condemned (though see the discussion between Massaro, Cohen, Campbell, & Rodriguez, 2001 and Pitt, Kim, & Myung, in press).

The essence of the argument against selection-by-fit arises from consideration of the type of data prevalent in psychology. Inevitably, the data will contain noise, which means that some proportion of the variability in the data is due to measurement error, and is *not* the outcome of any psychologically relevant process. Correspondingly, a model that gives a perfect account of noisy data should be regarded with suspicion, since it is attempting to explain the error variation in terms of a psychological process. Crucially, this error variation will be different if the experiment is replicated, but the psychologically relevant variation should remain the same. Therefore, the goal of model selection is to find the model that explains only the psychologically relevant variation in the data, not the noise. A model that fits the noise is guilty of the cardinal sin of overfitting, and will generalise poorly to new data.

Replacing selection-by-fit with a more principled framework is important, but not simple. The ideal model is of course the true one, but since the true model is never known in advance, this criterion is of no practical use whatsoever. Instead, a model

should be assessed with regard to a range of considerations. In broad, qualitative terms, a good model should give a good account of the data (data-fit). It should generalise well to new data sets. It should not be unnecessarily elaborate. It should make appropriate psychological assumptions, and it should be interpretable: a computational model that lacks a sound theoretical interpretation is *not* a psychological theory. See Myung and Pitt (in press) for a related discussion.

These guidelines are not easily met, and some are not amenable to quantification. In the end, there is no substitute for scientific judgement. Nonetheless, quantitative measures can go a long way towards providing a solid foundation for model selection. At the very least, it is certainly possible to improve on data fit as a selection criterion. Therefore, this chapter discusses several quantitative approaches to model selection, and then applies these ideas to similarity modelling. The goal is to provide well-founded selection criteria that may be used by the extraction algorithms when fitting a similarity model to empirical data.

3.1 Approaches to Model Selection

The model selection ideas discussed in this section come from a number of backgrounds, including computer science, statistics, and probability theory. In each case, some of the motivation, history and applicability of the criterion to similarity modelling is considered.

3.1.1 The Akaike Information Criterion

Akaike's (1973, 1983) widely-used Information Criterion (the AIC) is an approximation to the Kullback-Leibler information (Kullback, 1968), which measures how closely two distributions resemble one another. Suppose the data set D is the outcome of a continuous random variable $\mathbf{x}$ whose probability density function (pdf) is given by $f(\mathbf{x})$. Furthermore, let $g(\mathbf{x}|\theta)$ denote the density function generated by the model when

the parameter values $\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$ are substituted. Then the Kullback-Leibler information is given by

$$\mathrm{KL} = E\left[\ln f(\mathbf{x}) - \ln g(\mathbf{x}|\theta)\right],$$

where the expectation is taken over the data variable $\mathbf{x}$. Since $\mathbf{x}$ has pdf $f(\mathbf{x})$, this becomes

$$\mathrm{KL} = \int f(\mathbf{x}) \ln f(\mathbf{x})\, d\mathbf{x} - \int f(\mathbf{x}) \ln g(\mathbf{x}|\theta)\, d\mathbf{x}.$$

Since $\int f(\mathbf{x}) \ln f(\mathbf{x})\, d\mathbf{x}$ is constant with respect to $g(\mathbf{x}|\theta)$, the first term is disregarded. Furthermore, since f and g are both pdfs, the Kullback-Leibler information is minimal when $f(\mathbf{x})$ and $g(\mathbf{x}|\theta)$ are as similar as possible. That is, since

$$-\int f(\mathbf{x}) \ln g(\mathbf{x}|\theta)\, d\mathbf{x} = -E_{\mathbf{x}}\left[\ln g(\mathbf{x}|\theta)\right],$$

the model should be chosen that maximises $E\left[\ln g(\mathbf{x}|\theta)\right]$, which is the *expected* log likelihood of $g(\mathbf{x}|\theta)$, taken over the observed data, denoted $p(\mathbf{x}|\theta)$. If the data set D is expressed as a sample of observations x_i drawn from $f(\mathbf{x})$, it can be shown (see Bozdogan, 2000) that

$$-\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} \ln p(x_i|\theta^*)\right) = -\int f(\mathbf{x}) \ln g(\mathbf{x}|\theta^*)\, d\mathbf{x},$$

where θ^* are the parameter values that yield the maximum likelihood estimate. This relationship states that, as the sample size becomes arbitrarily large, the mean of the log maximum likelihood estimate gives the expected maximum log likelihood. However, this mean log maximum likelihood expression is a biased estimate, in that,

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \ln p(x_i|\theta^*) - \int f(\mathbf{x}) \ln g(\mathbf{x}|\theta^*)\, d\mathbf{x}\right] \neq 0.$$

It is difficult to find exact expressions for this bias in specific cases (Bozdogan, 2000), much less the general case. However, assuming that the true distribution $f(\mathbf{x})$ lies within

the set of density functions described by $g(\mathbf{x}|\theta)$, and that the parameters are independent of one another, the bias reduces to $\frac{k}{n}$ (Akaike, 1974), and therefore

$$\begin{aligned} \text{AIC} &= 2n\left(-\frac{1}{n}\sum_{i=1}^{n}\ln p(x_i|\theta^*) + \frac{k}{n}\right) \\ &= -2\ln p(\mathbf{x}|\theta^*) + 2k. \end{aligned}$$

Recalling that $\mathbf{x}$ describes the data distribution, this is rewritten as

$$\text{AIC} = -2\ln p(D|\theta^*) + 2k.$$

The AIC can be considered to consist of the data fit term $-2\ln p(D|\theta^*)$ and the complexity penalty term $2k$. However, it is important to recognise that the assumptions that parameters are independent and that the true density belongs to the model family are not always met.

3.1.2 Bayesian Model Selection

Bayesian statistics can be traced back to 1763, when an essay written by Thomas Bayes was published posthumously in the *Philosophical Transactions of the Royal Society of London*. Bayes' essential insight was to state "the rules for finding the probability of an event from the number of times it actually happens and fails" (Bayes, 1763, p. 394). This rule, known as Bayes' Theorem, was applied to statistics by Jeffreys (1935, 1961), who argued that if a set of data D is observed, a model M that assigns D the probability $p(D|M)$ has the following probability (called the *posterior probability*) of being true:

$$p(M|D) = p(D|M)\frac{p(M)}{p(D)}.$$

In contrast to the frequentist approach to statistics, Bayesian statistics require the specification of the prior likelihood of the model $p(M)$ and the data $p(D)$. In the context of

model selection, however, one is rarely concerned with finding the absolute probability of a model being true (which, in all honesty, is generally zero), but rather which of a number of competing models is more likely to be true (or "closer" to the truth). For instance, when comparing two models M_1 and M_2, it is often useful to calculate the ratio between the posterior probabilities of the two models:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)p(M_1)}{p(D|M_2)p(M_2)}.$$

The multi-model variant of the posterior odds ratio

$$\frac{p(M_i|D)}{\sum_k p(M_k|D)} = \frac{p(D|M_i)p(M_i)}{\sum_k p(D|M_k)p(M_k)}$$

has also been applied in psychology (see, for example, Lee's submittedb analysis of retention functions).

The posterior odds ratio has two important advantages. Firstly, the $p(D)$ term, being common to both posterior odds, disappears from the equation. Secondly, rather than being required to specify the absolute prior probability of each model $p(M_1)$ and $p(M_2)$, one needs only to specify the *relative* probability $\frac{p(M_1)}{p(M_2)}$. The remaining term, $\frac{p(D|M_1)}{p(D|M_2)}$, is known as the Bayes factor, denoted B_{12}, and its use has been advocated by authors such as Kass and Raftery (1995) and Myung and Pitt (1997). The Bayes factor is inherently interpretable, since it compares the probability of two events (or models). Jeffreys (1961) provides a rough guide as to the standards of evidence that should be applied, suggesting that a Bayes factor between 1 and 3 is "not worth more than a bare mention", whereas $3 < B_{12} < 10$ is positive evidence, $10 < B_{12} < 100$ is strong evidence and $B_{12} > 100$ is conclusive. It is important to note that Jeffreys was referring to general standards in science, and these guidelines are not hard and fast rules. Raftery (1995), for instance, replaces 10 with 20 in the guidelines, in order to provide a parallel with the .05 significance level in null hypothesis testing.

When a uniform model prior is chosen (i.e., all $p(M_i)$ values are equal), the Bayesian comparison between models reduces to the Bayes factor, and the best model is the one that has assigns the highest marginal likelihood $p(D|M)$ to the data. It is important to distinguish the marginal likelihood used in Bayesian model selection from the maximum likelihood used in many frequentist approaches. If M is a model containing free parameters with values given by the vector θ, then the likelihood function $p(D|M,\theta)$ gives the probability of the data under the model at the parameter values given by θ. The maximum likelihood function is the maximum value of the likelihood function, occurring at the optimum parameter values θ^*. In other words, the maximum likelihood is given by $p(D|M,\theta^*)$. In contrast, the marginal likelihood assigns to each θ a prior likelihood, $p(\theta)$, and is given by

$$p(D|M) = \int p(D|M,\theta)p(\theta)\, d\theta.$$

When $p(\theta)$ is a uniform prior, the marginal likelihood reduces to the integral of the likelihood function. Figure 3.1 demonstrates the difference between the maximum likelihood and marginal likelihood under uniform priors for two hypothetical single-parameter models. Notice that, although model A has the higher maximum likelihood, model B has the greater marginal likelihood. Importantly, since the marginal likelihood measures the fit of the model to the data across the entire parameter space, it is a superior measure of the adequacy of the model than the maximum likelihood (see Pitt et al., in press, for a related discussion).

Since the marginal likelihood involves an integral, it can be a difficult measure for which to obtain analytic expressions. However, when the likelihood function is unimodal and the majority of the integral mass lies close to the mode (i.e., the maximum likelihood parameters θ^*), it is often well-approximated by a multivariate Gaussian distribution. This so-called Laplacian approximation to the marginal probability (de Bruijn, 1958;

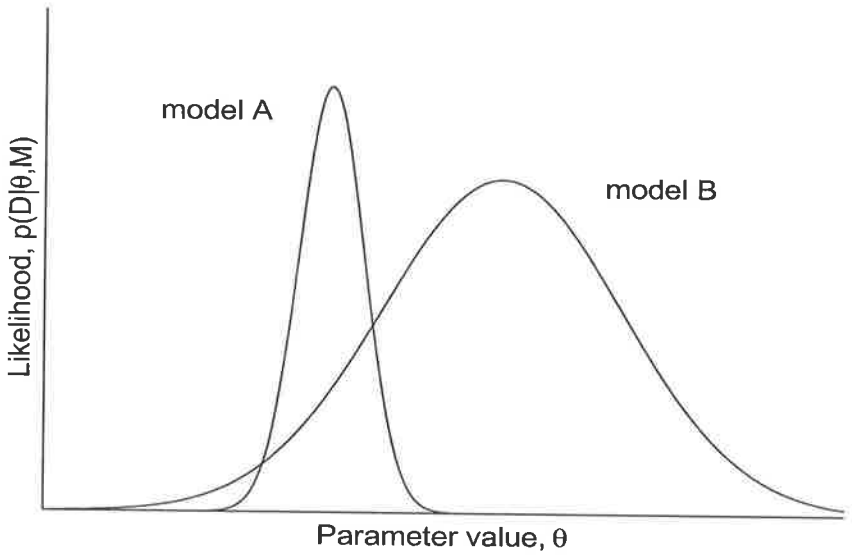

Figure 3.1: Likelihood functions for two single parameter models A and B. Note that, although model A has the greater maximum likelihood, model B has the greater marginal likelihood (under uniform priors).

Kass & Raftery, 1995; Tierney & Kadane, 1986) has been previously applied to similarity models by Lee (2001b) and Navarro and Lee (submitteda).

Another Bayesian approach is the Bayesian Information Criterion (BIC), first introduced by Schwarz (1978) as an approximation to the marginal probability. The BIC, which has been applied by Lee (2001a, 2001b) to multidimensional scaling and additive clustering, is given

$$\mathrm{BIC} = -2 \ln p(D|M, \theta^*) + k \ln N,$$

where k denotes the number of free parameters in the model, and N is the number of observations in the data (for similarity matrices, this is the number of unique and unconstrained entries in $\mathbf{S}$). The BIC has the advantage that it does not require the specification of prior densities, and is simple to calculate. The BIC approximation to the Bayes Factor is,

$$\mathrm{BIC}_1 - \mathrm{BIC}_2 \approx 2 \ln \mathrm{B}_{12}$$

and Raftery (1995) notes that the error in this approximation is of the order $O(1)$. In other words, even as $N \rightarrow \infty$ the error remains constant. However, he also observes that, in practice, the BIC is more accurate than this error term implies.

For comparative purposes, it is worth noting that Akaike (1983) provides a Bayesian derivation for the AIC. Like the BIC, the AIC can used to approximate the Bayes factor, but the AIC assumes that the amount of information conveyed by the prior is comparable to that of the data itself (Kass & Raftery, 1995). Generally speaking, this is not appropriate for data obtained in psychology. Data sets are usually collected when little information is already known, and thus the prior knowledge is substantially less than the amount of information in the data.

The difference between the two measures gives rise to different model selections. The penalty term in the BIC rises proportionally to $\ln N$ for a given number of parameters, whereas the AIC penalty remains constant. Consequently, the BIC favours more parsimonious models in the presence of large quantities of data. Kass and Raftery (1995) argue that the AIC overestimates the number of model parameters, though some criticisms of the BIC's stronger parsimony are made by Findley (1991).

3.1.3 Stochastic Complexity

Stochastic Complexity (SC: Rissanen, 1984, 1986, 1996) is based on the Minimum Description Length (MDL) principle, which originates from computer science. Under the MDL approach, a data set D is characterised as the sequence of observations $\{d_1, d_2, \ldots, d_n\}$, and the aim is to compress the sequence as much as possible. One of the early statements of the MDL principle was given by Kolmogorov (1965), who defined the MDL as the length in bits of the shortest program that prints D and then halts. However, this length depends on the programming language used to code the sequence. Fortunately, Solomonoff (1964a, 1964b) demonstrated that this concern is minor, since

the difference between two programming languages for any sequence D is independent of the length of that sequence.

Nevertheless, Kolmogorov's notion is not easily applied in practice, since it is not always possible to specify the program that gives the maximum data compression. Instead, the approach taken by Rissanen (1996, see also Grunwald, 2000) is to apply some *coding scheme* C that, although not guaranteed to give maximum compression, nevertheless provides a reasonable approximation. A coding scheme C is some rule that encodes a data sequence D as a binary sequence (note that D itself need not be binary). For instance, for a set of 5 data sequences D_1, D_2, D_3, D_4, D_5, a particular code C might yield $C(D_1) = 0$, $C(D_2) = 1$, $C(D_3) = 00$, $C(D_4) = 01$, $C(D_5) = 10$. The description length for a sequence D, as given by some code C, is the length of $C(D)$, denoted $\mathrm{L}_C(D)$. As it happens (Grunwald, 2000), if non-integer code lengths are allowed, then for any code[1] C there exists a probability density function p_C such that $p_C(D) = 2^{-\mathrm{L}_C(D)}$, and for every density function there exists a corresponding code. Functionally speaking, a code denotes a probability distribution. That is, a statistical model M with parameters θ is equivalent to what the MDL nomenclature calls a *model class* M consisting of a set of codes θ.

As Grunwald (2000) points out, this correspondence implies that

$$\min_{\theta} \mathrm{L}(D|M,\theta) = \min_{\theta} \left(-\log_2 p(D|M,\theta)\right)$$

which occurs when $p(D|M,\theta)$ is maximised: that is, at the maximum likelihood parameter values θ^*. As it stands, the MDL principle appears to state that one should use maximum likelihood to select between models. However, although the "maximum likelihood code" θ^* gives the shortest description of D, the MDL principle requires that the code should compress all of the data sequences indexed by M to the maximum extent. The aim is therefore to find a code that yields the shortest average description

[1] To be precise, only those codes for which one can uniquely recover D from $C(D)$.

length for those sequences. This code is called the *stochastic complexity code*, and the description length of D when this code is used is called the stochastic complexity of D with respect to the model (class) M. Rissanen (1996) provides the following measure of stochastic complexity, accurate up to $O(1)$ error:

$$\mathrm{SC} = -\ln p(D|M,\theta^*) + \frac{k}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int\sqrt{|\mathbf{I}(\theta)|}\,d\theta,$$

where $\mathbf{I}(\theta)$ is the Fisher information matrix and the integral is taken over the entire parameter space of the model (or the set of codes indexed by the model class).

3.1.4 Geometric Complexity

The Geometric Complexity Criterion (Myung, Balasubramanian, & Pitt, 2000; see also Balasubramanian, 1997, 1997; Pitt et al., 2002) identifies a model with the set of *distinguishable* probability distributions that it can generate under all parameterisations. The rationale for counting only distinguishable distributions is that distributions that remain indistinguishable as $N \to \infty$ should not be counted as separate distributions for model selection purposes. This set of distributions forms a surface in the space of probability distributions. Accordingly, they identify two measures of interest: the volume of the model manifold, denoted $V(M)$, and a measure of the number of distributions in the manifold that lie close to the true distribution, denoted $V_c(M)$. With this in mind, they consider the complexity of a model to be given by the natural logarithm of the ratio of $V(M)$ to $V_c(M)$. In other words, the complexity of a model is related to (the inverse of) the number of distributions indexed by the model that lie close to the truth. The Geometric Complexity of a model is given by

$$\ln\left(\frac{V(M)}{V_c(M)}\right) = \frac{k}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int\sqrt{|\mathbf{I}(\theta)|}\,d\theta + \frac{1}{2}\ln\left(\frac{|\mathbf{J}(\theta^*)|}{|\mathbf{I}(\theta^*)|}\right),$$

where once again, $\mathbf{I}(\theta)$ is the Fisher Information Matrix, $\mathbf{J}(\theta)$ is the Hessian matrix of

second-order partial derivatives, and the integral is taken over the entire parameter space of the model. Similarly, the Geometric Complexity Criterion (GCC) is given by

$$\mathrm{GCC} = -\ln p(D|\theta^*) + \frac{k}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int\sqrt{|\mathbf{I}(\theta)|}\,d\theta + \frac{1}{2}\ln\left(\frac{|\mathbf{J}(\theta^*)|}{|\mathbf{I}(\theta^*)|}\right). \qquad (3.1)$$

The GCC is invariant under reparameterisation of the model, and perhaps represents the state of the art for quantitative model selection criteria.

Since the GCC is used as a selection criterion on a number of occasions in the remainder of this thesis, it will be useful to have some guidelines for interpreting differences in GCC values, similar to those proposed by Jeffreys (1961) for interpreting Bayes factors. As it happens, Balasubramanian (1997, 1999) provides a Bayesian derivation for the GCC, by choosing uniform model priors and choosing Jeffreys' (1961) prior[2] for the parameter values θ, and shows that the GCC is an expansion of the log posterior probability. That is,

$$\mathrm{GCC}_1 - \mathrm{GCC}_2 \approx \ln\frac{p(M_1|D)}{p(M_2|D)},$$

which permits the use of Raftery's (1995) adaptation of Jeffrey's (1961) guidelines displayed in Table 3.1. Alternatively, Myung et al. (2000) use the complexity terms to compare the complexity of the two models directly. The choice of whether to use relative probabilities or to compare complexities when evaluating models will, as always, depend on the needs of the analysis.

[2]Balasubramanian provides a compelling justification for this choice, observing that Jeffreys' prior, $p(\theta) = \sqrt{|\mathbf{I}(\theta)|}\Big/\int\sqrt{|\mathbf{I}(\theta)|}\,d\theta$, corresponds to the assumption that each *distribution* indexed by the model has equal prior likelihood.

Table 3.1: Evidence provided by differences in GCC values.

GCC difference	Bayes factor	Evidence
0-2	1-3	Weak
2-6	3-20	Positive
6-10	20-150	Strong
>10	>150	Very Strong

3.2 Choosing an Additive Clustering Representation

In this section a GCC expression is derived for the common features similarity model $\hat{s}_{ij} = \sum_k w_k f_{ik} f_{jk} + c$ used in additive clustering procedures (Shepard & Arabie, 1979). Once this derivation is complete, the implications for featural representation are considered: the GCC expression allows a clear statement of what makes a featural representation more or less complex.

3.2.1 GCC Derivation

Before commencing the derivation, observe that the free parameters in an m-feature additive clustering representation are the saliency weights $\mathbf{w} = \{w_1, w_2, \ldots, w_m\}$ and the additive constant c. For the purposes of this analysis, it is convenient to treat the additive constant as a mandatory extra cluster containing all of the stimuli. Therefore an additive clustering model shall be said to contain m' features: if an additive constant is included (as is customary), then $m' = m + 1$ and one of the clusters must contain all of the stimuli; if for some reason no additive constant is included, then $m' = m$, and there are no restrictions on cluster membership. This allows the similarity model to be written $\hat{s}_{ij} = \sum_k w_k f_{ik} f_{jk}$ *without* disregarding the additive constant.

The data set D is given by the similarity matrix $\mathbf{S}$, and the number of observations

N is given by the number of unique and unconstrained entries in $\mathbf{S}$, not the number of judgements that gave rise to them. As a result, there are $\frac{n(n-1)}{2}$ observations in a symmetric matrix. Using these observations, the GCC can be written as,

$$\mathrm{GCC}_{\mathrm{adclus}} = -\ln p(\mathbf{S}|\mathbf{F},\mathbf{w}^*) + \frac{m'}{2}\ln\left(\frac{n(n-1)}{4\pi}\right) + \ln\int\sqrt{|\mathbf{I}(\mathbf{w})|}d\mathbf{w} + \frac{1}{2}\ln\left(\frac{|\mathbf{J}(\mathbf{w}^*)|}{|\mathbf{I}(\mathbf{w}^*)|}\right).$$

The first term in this expression requires the maximum likelihood estimate $p(\mathbf{S}|\mathbf{F},\mathbf{w}^*)$ for the model. Following Tenenbaum (1996), it is assumed that the observed similarities s_{ij} are drawn from Gaussian distributions with mean $\hat{s}_{ij}$ and common variance σ. The width of the distribution corresponds to the assumption made about the precision of the data, so it is best to estimate σ from the data. Under these assumptions, the likelihood of a similarity matrix $\mathbf{S}$ given a feature matrix $\mathbf{F}$ and saliency weights $\mathbf{w}$ is

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

Therefore, the negative maximum log likelihood function is given by

$$-\ln p(\mathbf{S}|\mathbf{F},\mathbf{w}^*) = \frac{1}{2\sigma^2}\sum_{i<j}(s_{ij}-\hat{s}^*_{ij})^2 + \frac{n(n-1)}{2}\ln\left(\sigma\sqrt{2\pi}\right).$$

The third and fourth terms of the GCC require expressions for the Hessian matrix $\mathbf{J}(\mathbf{w})$ and the Fisher Information Matrix $\mathbf{I}(\mathbf{w})$. Therefore, the second-order partial derivatives,

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

are required. The first-order partial derivative is given,

$$\frac{\partial \ln p(\mathbf{S}|\mathbf{F},\mathbf{w})}{\partial w_x} = -\frac{1}{2\sigma^2}\sum_{i<j}2\left(s_{ij}-\hat{s}_{ij}\right)\frac{\partial}{\partial w_x}\left(s_{ij}-\hat{s}_{ij}\right)$$

$$= -\frac{1}{\sigma^2}\sum_{i<j}(s_{ij}-\hat{s}_{ij})\frac{\partial}{\partial w_x}(-\hat{s}_{ij})$$
$$= -\frac{1}{\sigma^2}\sum_{i<j}(\hat{s}_{ij}-s_{ij})\frac{\partial}{\partial w_x}\hat{s}_{ij}.$$

According to the common features rule under consideration, the similarity between the ith and jth stimuli is given by the sum of the weights of shared features only. Therefore, if the xth feature is not a feature shared by the ith and jth stimuli, then $\hat{s}_{ij}$ will be constant with respect to w_x, and have a partial derivative of 0. Correspondingly, if the xth feature *is* common to both stimuli, then $\hat{s}_{ij} = w_x + \sum_{k\neq x} w_k f_{ik} f_{jk}$ the partial derivative is 1. Hence,

$$\frac{\partial \ln p(\mathbf{S}|\mathbf{F},\mathbf{w})}{\partial w_x} = \frac{1}{\sigma^2}\sum_{i<j|w_x\in C_{ij}}(\hat{s}_{ij}-s_{ij}),$$

where C_{ij} denotes the weights of those features shared by the ith and jth stimuli. Given this, the second-order partial derivatives are

$$\frac{\partial^2 \ln p(\mathbf{S}|\mathbf{F},\mathbf{w})}{\partial w_y w_x} = -\frac{1}{\sigma^2}\sum_{i<j|w_x\in C_{ij}}\frac{\partial}{\partial w_y}(\hat{s}_{ij}-s_{ij})$$
$$= -\frac{1}{\sigma^2}\sum_{i<j|w_x\in C_{ij}}\frac{\partial}{\partial w_y}\hat{s}_{ij}.$$

Again, the partial derivative of $\hat{s}_{ij}$ with respect to w_y is either 1 or 0. Accordingly,

$$\frac{\partial^2 \ln p(\mathbf{S}|\mathbf{F},\mathbf{w})}{\partial w_y \partial w_x} = -\frac{1}{\sigma^2}\sum_{i<j|w_x,w_y\in C_{ij}}1$$
$$= -\frac{1}{\sigma^2}\sum_{i<j} f_{ix}f_{jx}f_{iy}f_{jy}.$$

Using this result, the Hessian matrix $\mathbf{J}(\mathbf{w}) = [j_{xy}(\mathbf{w})]$ can be expressed by noting that

$$j_{xy}(\mathbf{w}) = -\frac{\partial^2 \ln p(\mathbf{S}|\mathbf{F},\mathbf{w})}{\partial w_y \partial w_x}$$
$$= \frac{1}{\sigma^2}\sum_{i<j} f_{ix}f_{jx}f_{iy}f_{jy},$$

and since this expression does not depend on $\mathbf{w}$, it is trivial to note that this value remains the same when the maximum likelihood parameter values $\mathbf{w}^*$ are substituted. Similarly, the elements of the Fisher Information Matrix $\mathbf{I}(\mathbf{w}) = [i_{xy}(\mathbf{w})]$ are given by,

$$\begin{aligned} i_{xy}(\mathbf{w}) &= -E\left[\frac{\partial^2 \ln p(\mathbf{S}|\mathbf{F},\mathbf{w})}{\partial w_y \partial w_x}\right] \\ &= -E\left[-\frac{1}{\sigma^2}\sum_{i<j} f_{ix}f_{jx}f_{iy}f_{jy}\right] \\ &= \frac{1}{\sigma^2}\sum_{i<j} f_{ix}f_{jx}f_{iy}f_{jy}. \end{aligned}$$

Again, this expression is constant with respect to $\mathbf{w}$, so the substitution $\mathbf{w} = \mathbf{w}^*$ is trivial. Consequently, $\mathbf{J}(\mathbf{w}^*) = \mathbf{I}(\mathbf{w}^*)$, and the last term in the GCC reduces to $\frac{1}{2}\ln 1 = 0$.

The third term in the GCC requires the integration of the constant expression $\sqrt{|\mathbf{I}(\mathbf{w})|}$ over the parameter space of the model. Since it is customary to normalise similarity data to lie between 0 and 1, Lee (2001b) has argued that $0 \leq w_k \leq 1$ is the natural constraint on the saliency weights for additive clustering models. If so,

$$\begin{aligned} \ln \int \sqrt{|\mathbf{I}(\mathbf{w})|}\, d\mathbf{w} &= \ln \int_0^1 \int_0^1 \cdots \int_0^1 \int_0^1 \sqrt{|\mathbf{I}(\mathbf{w})|}\, dw_1\, dw_2 \ldots dw_{m'-1}\, dw_{m'} \\ &= \frac{1}{2}\ln |\mathbf{I}(\mathbf{w})| \\ &= \frac{1}{2}\ln \left|\frac{1}{\sigma^2} \times \mathbf{G}\right|, \end{aligned}$$

where $\mathbf{G} = [g_{xy}]$ denotes the $m' \times m'$ *complexity matrix* such that $g_{xy} = \sum_{i<j} f_{ix}f_{jx}f_{iy}f_{jy}$. In other words, the xyth element of $\mathbf{G}$ counts the number of pairs of stimuli that share the xth feature and the yth feature. Main diagonal elements of $\mathbf{G}$ are given by the number of pairs of stimuli that share a single feature: that is, g_{xx} reduces to $\sum_{i<j} f_{ix}f_{jx}$. This complexity matrix is equivalent to the complexity matrix found by Lee (2001b), who used the Laplacian approximation to the Bayesian posterior to estimate the complexity

of additive clustering models. By extracting a factor of $\frac{1}{\sigma^2}$ from each of the m' rows of $\left|\frac{1}{\sigma^2} \times \mathbf{G}\right|$, the expression becomes $\left(\frac{1}{\sigma^2}\right)^{m'} \times |\mathbf{G}|$. Therefore,

$$\begin{aligned} \ln \int \sqrt{|\mathbf{I}(\mathbf{w})|}\, d\mathbf{w} &= \frac{1}{2} \ln \left(\frac{1}{\sigma^{2m'}} |\mathbf{G}| \right) \\ &= -m' \ln \sigma + \frac{1}{2} \ln |\mathbf{G}|. \end{aligned}$$

The GCC for common features representations is therefore given by

$$\begin{aligned} \mathrm{GCC}_{\mathrm{adclus}} &= \frac{1}{2\sigma^2} \sum_{i<j} (s_{ij} - \hat{s}_{ij})^2 + \frac{m'}{2} \ln \left(\frac{n(n-1)}{4\pi} \right) - m' \ln \sigma + \frac{1}{2} \ln |\mathbf{G}| \\ &\quad + \frac{n(n-1)}{2} \ln \left(\sigma \sqrt{\pi} \right) \end{aligned}$$

which, after rearrangement, becomes

$$\begin{aligned} \mathrm{GCC}_{\mathrm{adclus}} &= \frac{1}{2\sigma^2} \sum_{i<j} (s_{ij} - \hat{s}_{ij})^2 + \frac{m'}{2} \ln \left(\frac{n(n-1)}{4\pi\sigma^2} \right) + \frac{1}{2} \ln |\mathbf{G}| \\ &\quad + \frac{n(n-1)}{2} \ln \left(\sigma \sqrt{2\pi} \right). \end{aligned} \tag{3.2}$$

The last term, the *data constant* $\frac{n(n-1)}{2} \ln \left(\sigma \sqrt{2\pi} \right)$, is invariant across models, though not across data sets. However, since one generally evaluates models on the same data set, the data constant makes no contribution to model comparison, and may be dropped[3]. Therefore, it is often convenient to disregard the constant (e.g., Navarro & Lee, submittedb), and treat the GCC as

$$\mathrm{GCC}_{\mathrm{adclus}} = \frac{1}{2\sigma^2} \sum_{i<j} (s_{ij} - \hat{s}_{ij})^2 + \frac{m'}{2} \ln \left(\frac{n(n-1)}{4\pi\sigma^2} \right) + \frac{1}{2} \ln |\mathbf{G}| + \text{constant}. \tag{3.3}$$

[3]The data constant does matter if two models cannot be tested on the same data set, and must be evaluated on separate data sets that differ in σ or n. However, it would be highly unusual to find two models, purporting to account for the same phenomenon, that are so incommensurate that they cannot be applied to the same data sets. Arguably, if two models cannot be compared using the same data set, then one should not try compare them with the GCC either.

This measure is the sum of three terms, reading from left to right: the *error*, the *parametric complexity* and *structural complexity* of a featural model. The error term measures the extent to which the predictions made by the model depart from the observed data, whereas the other terms measure the complexity of the model. Specifically, the parametric complexity term penalises a representation for the number of features it possesses, whereas the structural complexity terms measures the complexity that arises from the pattern of overlap and encompassment of those features.

3.2.2 Data Precision and the GCC

Suppose a second experiment were conducted, yielding an identical similarity matrix, but with different precision. What happens to the GCC for a featural representation? The error term is proportional to $\frac{1}{\sigma^2}$, so as precision increases (i.e., σ decreases), the error term grows quadratically. The parametric complexity term is proportional to $\ln \frac{1}{\sigma^2}$, so it rises logarithmically as σ shrinks. The structural complexity term is independent of σ, and so is unaffected. Finally, the data constant, being proportional to $\ln \sigma$, shrinks logarithmically with σ.

The interpretation of these effects is straightforward. Firstly, since the data constant is independent of the model, the shrinking constant will not cause the GCC to select differently between candidate models, and may therefore be disregarded. Secondly, since the error term grows much more rapidly than any other term, the major effect of decreasing σ is that the error term becomes more heavily weighted than the two complexity terms. From a model selection standpoint, this means that as the data become more precise, it grows more important for a model to provide a good fit than to have low complexity. The trade-off between fit and complexity tilts further toward data fit as the data precision rises, in line with the intuitive notion of what data precision is intended to capture. Finally, as data precision increases, the relative importance of

the two complexity terms shifts towards parametric complexity. As σ^2 shrinks, the parametric complexity term grows, but the structural complexity term remains constant. Correspondingly, the importance of $|\mathbf{G}|$ (which captures the inherent complexity of the feature structure, $\mathbf{F}$) diminishes relative to number of free parameters. Put simply, when the data are highly precise, the number of features matters more than the manner in which they are arranged.

3.2.3 Domain Size and the GCC

Suppose the data reflect the similarity of 10 tones with frequencies within the range 100-1000Hz. In this situation, it is easy to imagine that the domain could consist of 20 tones as easily as 10, and still be the "same" domain. Therefore, it worth investigating the effect of increasing the "resolution" of the domain by raising the number of stimuli n (assuming that precision and data fit remain constant). The error term counts the squared discrepancy between data and model for all $\frac{n(n-1)}{2}$ pairs of stimuli. If each pair of stimuli is associated (on average) with some fixed amount of (squared) error, then the error term increases quadratically with n. In contrast, the parametric complexity term increases only logarithmically with n. The data constant rises quadratically with n, but since it never differs for two models, is irrelevant.

The structural complexity term is also affected by increasing n, but in a less straightforward fashion. To understand its behaviour, consider the example of the tones. Suppose the initial stimulus set were 10 tones at frequencies of 100Hz, 200Hz, 300Hz, and so on up to 1000Hz, and that the expanded stimulus set consisted of the initial tones, plus those at 105Hz, 205Hz and so on. In such a case, one could consider the relationship between two complexity matrices $\mathbf{G}_1$ and $\mathbf{G}_2$, corresponding to two feature structures $\mathbf{F}_1$ and $\mathbf{F}_2$ that have the same number of features and the same pattern of overlap. However, each feature in $\mathbf{F}_2$ contains twice as many stimuli as its counterpart in $\mathbf{F}_1$ (as

does every possible union or intersection of features). In a sense, each stimulus in $\mathbf{F}_1$ gets "split" into two stimuli in $\mathbf{F}_2$. If a given feature (or intersection of two features) contains x stimuli, the corresponding entry in $\mathbf{G}_1$ is $x(x-1)/2$. If this is doubled, then the same cell in $\mathbf{G}_2$ is $2x(2x-1)/2$, approximately four times the original entry. More generally, if the n_1 stimuli are split into n_2 stimuli in this manner,

$$\begin{aligned}
\ln|\mathbf{G}_2| &= \ln\left|\frac{n_2(n_2-1)}{n_1(n_1-1)} \times \mathbf{G}_1\right| \\
&= \ln\left[\left(\frac{n_2(n_2-1)}{n_1(n_1-1)}\right)^{m'} \times |\mathbf{G}_1|\right] \\
&= m' \ln\left(\frac{n_2(n_2-1)}{n_1(n_1-1)}\right) + \ln|\mathbf{G}_1|.
\end{aligned}$$

Therefore, the structural complexity term should be expected to increase in a logarithmic fashion as n increases.

Increasing the domain size causes all three terms to rise. Since the error term rises most quickly, the effect is to weight data-fit more heavily than model complexity. On a more general note, the resemblance between the effects of n and σ may not be entirely coincidental. In general terms, the precision value indicates how much a given similarity value s_{ij} should be taken to constrain the model: precise data provide stronger constraints. When the number of stimuli is increased, the number of empirical similarities also increases, and therefore provides a stronger constraint on the model. It is for this reason that the number of stimuli can be considered to be a "resolution parameter".

3.2.4 The Structure of F and the GCC

In this section the discussion turns to the feature structure $\mathbf{F}$ itself, and the effect it has on the complexity of the model. The only part of the GCC that is affected by $\mathbf{F}$ is

the structural complexity term. Correspondingly, this section discusses the effect of $\mathbf{F}$ on $|\mathbf{G}|$. This complexity matrix has been previously analysed by Lee (2001b), so this section briefly summarises and comments on his analysis (see also Navarro, submitted).

Notation for the Complexity Matrix

Up to this point the inclusion (or exclusion) of the additive constant has not mattered greatly. If an additive constant is included, one substitutes $m' = m + 1$ as the number of parameters, and treats the additive constant as the last "cluster", which must contain all stimuli. However, in discussing the behaviour of the complexity matrix, it is crucial to be explicit about whether the additive constant is included. Therefore, to avoid any confusion regarding these matrices, a complexity matrix that does not incorporate an additive constant is denoted by the $m \times m$ matrix $\mathbf{G}_f$ and a matrix that includes the additive constant by the $(m+1) \times (m+1)$ matrix $\mathbf{G}^+$. Accordingly, $\mathbf{G}_f$ can be written,

$$\mathbf{G}_f = \begin{bmatrix} \sum_{i<j} f_{i1}f_{j1} & \sum_{i<j} f_{i1}f_{j1}f_{i2}f_{j2} & \cdots & \sum_{i<j} f_{i1}f_{j1}f_{im}f_{jm} \\ \sum_{i<j} f_{i2}f_{j2}f_{i1}f_{j1} & \sum_{i<j} f_{i2}f_{j2} & & \sum_{i<j} f_{i2}f_{j2}f_{im}f_{jm} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i<j} f_{im}f_{jm}f_{i1}f_{j1} & \sum_{i<j} f_{im}f_{jm}f_{i2}f_{j2} & & \sum_{i<j} f_{im}f_{jm} \end{bmatrix},$$

and $\mathbf{G}^+$ is given by

$$\mathbf{G}^+ = \begin{bmatrix} \mathbf{G}_f & \mathbf{y} \\ \mathbf{y}' & z \end{bmatrix},$$

where the scalar z is given by $\frac{n(n-1)}{2}$ and the vector $\mathbf{y} = \{y_1, \ldots, y_m\}'$ contains the main diagonal elements of $\mathbf{G}_f$: that is, $y_x = g_{xx} = \sum_{i<j} f_{ix}f_{jx}$. Whenever the term $\mathbf{G}$ appears in the text, it applies to both $\mathbf{G}_f$ and $\mathbf{G}^+$.

Summary of Lee's Conclusions

Lee demonstrates that for non-degenerate feature structures, the complexity matrix is positive definite, and applies Hadamard's inequality (Bellman, 1970, pp. 129-130),

which states that the determinant is less than or equal to the product of its main diagonal,

$$|\mathbf{G}_f| \leq \prod_x \sum_{i<j} f_{ix} f_{jx}$$

with equality occurring when all off-diagonal elements of $\mathbf{G}_f$ are zero[4]. From this Lee argues that, for a fixed number of clusters, the most complex representation is a partition, in which every stimulus belongs to precisely one cluster, since these models have diagonal complexity matrices. There are two caveats to attach to this analysis: the first regards diagonal complexity matrices, and the second regards the applicability of Hadamard's inequality.

Regarding Diagonal Complexity Matrices

It should be observed that although all partitions have diagonal complexity matrices, not all diagonal complexity matrices correspond to partitions. A diagonal complexity matrix results whenever no two stimuli ever share two or more features. Therefore, so long as every pair of clusters have no more than a single stimulus in common, $\mathbf{G}_f$ remains diagonal. A concrete example of this is illustrated by Figure 3.2, in which feature structures A and B yield precisely the same (diagonal) complexity matrix. All features have 3 stimuli and hence $\binom{3}{2} = 3$ stimulus pairs, and no two features are shared by any two stimuli, even though only feature set A is a partition. Lee's analysis of diagonal matrices was restricted to partitions. He argued that transferring one stimulus from a smaller cluster to a larger one always reduces complexity, suggesting that the minimally complex partition is the one in which all clusters save one possess only two stimuli, and the rest encompasses all remaining stimuli, and furthermore that this result holds irrespective of whether an additive constant is included (i.e., applies to $\mathbf{G}^+$ as well as $\mathbf{G}_f$). However, this result holds not just for partitions, but also for *any* feature structure

[4]Although Hadamard's inequality holds for $\mathbf{G}^+$ as well as $\mathbf{G}_f$, it is impossible for $\mathbf{G}^+$ to be a diagonal matrix. In any case, since the additive constant is not considered to be a true feature, a representation that has diagonal $\mathbf{G}_f$ could still be considered a partition even though $\mathbf{G}^+$ is not diagonal.

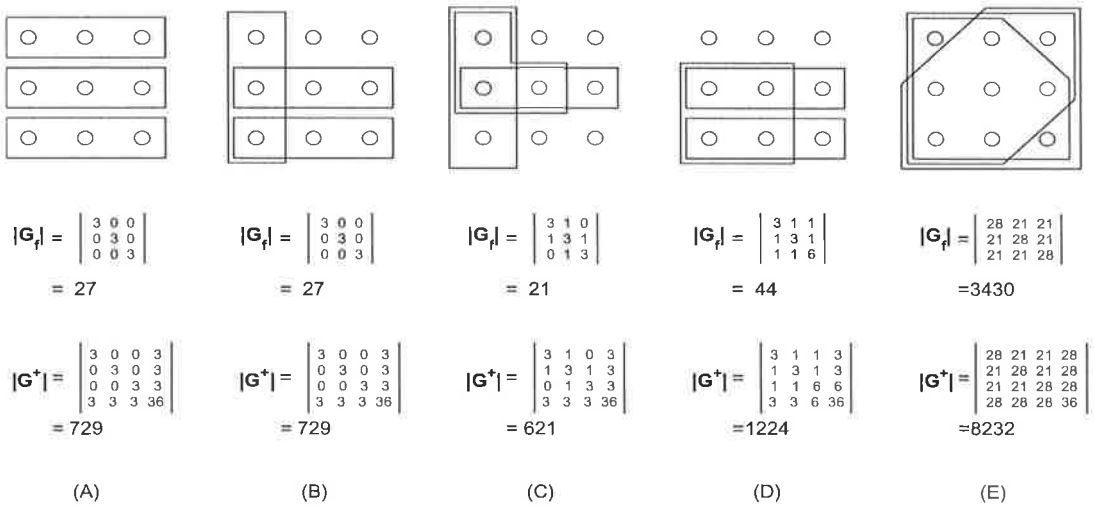

Figure 3.2: Five feature structures for a nine-stimulus domain. Feature set A has a partitioning structure, whereas B is an example of a non-partitioning structure that also has a diagonal complexity matrix. In set C, each cluster still encompasses three stimuli, but some overlap emerges. Feature structure D introduces a small amount of overlap at the expense of increasing the size of one cluster, whereas the features in E are large and overlap extensively. Two complexity matrices are given for each: $\mathbf{G}_f$ is the complexity matrix for the features shown, whereas $\mathbf{G}^+$ incorporates the additive constant.

that yields a diagonal feature matrix (the mathematics are given by Lee, 2001b, and it is a small matter to observe that the logic holds for any diagonal feature matrix).

The broader notion of what makes for a "diagonal feature structure" invites a second observation that (so long as the feature structure remains non-degenerate) complexity is always reduced by removing a stimulus from a cluster. If there is no additive constant, this is very easily observed, since the result is to reduce one of the elements in the product $\prod_x \sum_{i<j} f_{ix} f_{jx}$, and hence lowering $|\mathbf{G}_f|$. Once an additive constant is included, the story is slightly more complex. Nevertheless, consider the case when the first cluster contains $\alpha \geq 3$ stimuli. In this case $|\mathbf{G}^+|$ is given by

$$|\mathbf{G}^+| = \frac{\alpha(\alpha-1)}{2} \times \prod_{x=2}^{m} g_{xx} \times \left(z - \frac{\alpha(\alpha-1)}{2} - \sum_{x=2}^{m} g_{xx} \right). \tag{3.4}$$

By removing one stimulus from this cluster, the determinant of the complexity matrix

becomes

$$|\mathbf{G}^{+}| = \frac{(\alpha-1)(\alpha-2)}{2} \times \prod_{x=2}^{m} g_{xx} \times \left(z - \frac{(\alpha-1)(\alpha-2)}{2} - \sum_{x=2}^{m} g_{xx} \right). \tag{3.5}$$

So, the determinant $|\mathbf{G}^{+}|$ is larger in Eq. 3.4 than in Eq. 3.5 when

$$\alpha \left(z - \frac{\alpha(\alpha-1)}{2} - \sum_{x=2}^{m} g_{xx} \right) > (\alpha-2) \left(z - \frac{(\alpha-1)(\alpha-2)}{2} - \sum_{x=2}^{m} g_{xx} \right)$$

which, after rearrangement, reduces to

$$\frac{\alpha^2(\alpha-1)}{2} + \frac{(\alpha-1)(\alpha-2)^2}{2} + 2z + 2\sum_{x=2}^{m} g_{xx} > 0$$

and it is clear from inspection that since $\alpha \geq 3$, all terms on the left hand side are positive, ensuring that the inequality holds.

To summarise: for any feature structure that yields a diagonal complexity matrix, with or without an additive constant, complexity decreases whenever one (a) transfers a stimulus from a smaller cluster to a larger cluster, or (b) removes a stimulus from a cluster. This is illustrated in Figure 3.3.

Regarding the Applicability of Hadamard's Inequality

Hadamard's inequality indicates that, when clusters share pairs of stimuli without changing size, as in feature structure C shown in Figure 3.2, the determinant of the complexity matrix decreases in accordance with Hadamard's inequality: $|\mathbf{G}_f|$ for structures A and B is 27, whereas $|\mathbf{G}_f|$ for structure C equals 21. If the additive constant is included, the determinant of the expanded matrix $\mathbf{G}^{+}$ is 729 for A and B, and 621 for C. In general, the more a pair of clusters overlap (in terms of stimulus *pairs*) the less complexity is introduced, since the unique contribution each cluster makes to the overall complexity is smaller.

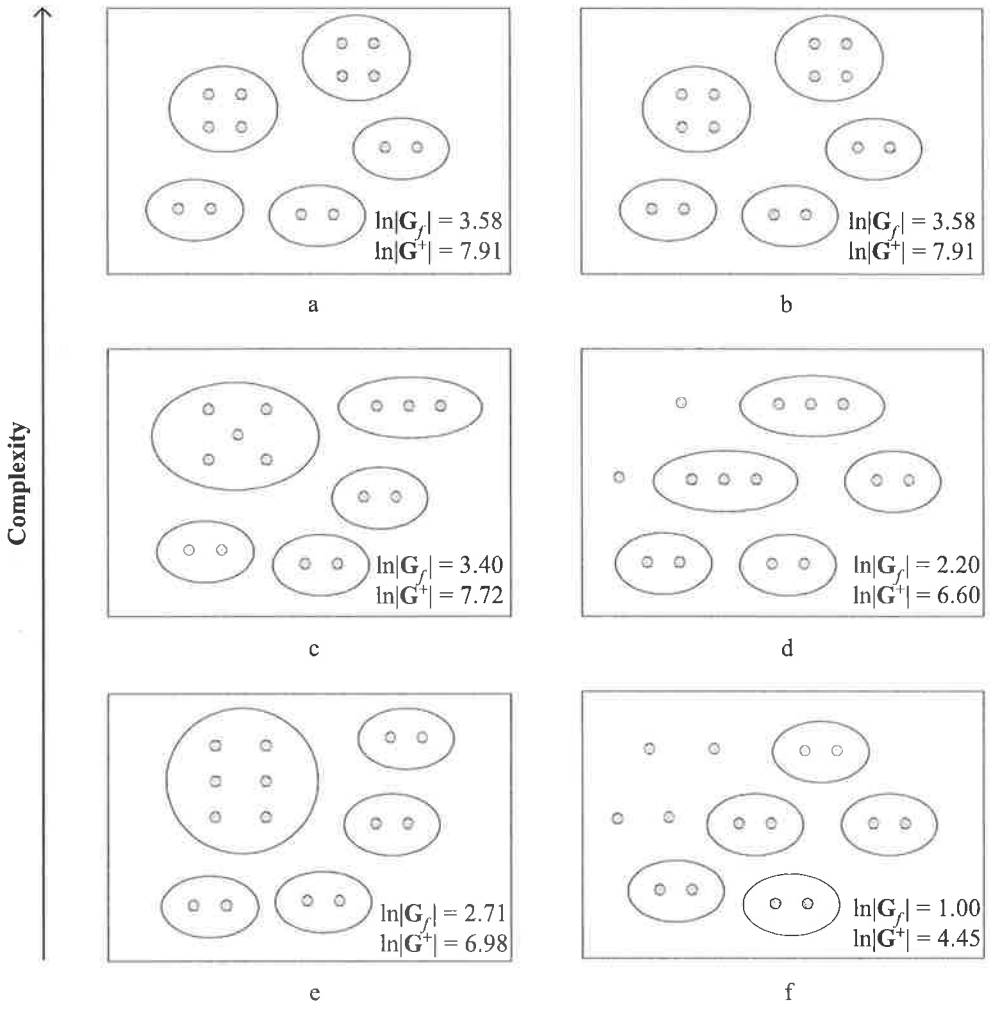

Figure 3.3: Six disjoint feature structures of varying complexity. Panels a, c and e illustrate the increase in complexity achieved by transferring stimuli from larger to smaller clusters, whereas panels b, d and f show how complexity increases by enlarging features.

The second caveat that attaches to Lee's discussion is that Hadamard's inequality applies only if the product of the main diagonal elements remains constant: that is, when the number of stimuli (and hence pairs of stimuli) in each cluster remains constant. Hadamard's inequality does not indicate what happens to the model's complexity as the number of stimuli in a cluster changes. Therefore, although Lee identifies encompassment and overlap as sources of model complexity, arguments based on Hadamard's inequality only take overlap into account. In some situations, these two factors can be varied independently: for example, a stimulus that does not belong to any cluster can be added to one of them without causing any change in the off-diagonal elements of $\mathbf{G}_f$. Similarly, the comparison between feature structures A and C in Figure 3.2 involves manipulating the overlap between clusters without changing their size. Nevertheless, such independence is not the norm, and it is not immediately obvious what happens to complexity when a feature is enlarged at the expense of introducing more overlap. Consider feature structures A, D and E in Figure 3.2. Two of the features in A and D are identical, but the third feature in D contains four stimuli rather than three, and shares one stimulus pair with each of the other two features. As it turns out, D is the more complex representation, with $|\mathbf{G}_f| = 44$ and $|\mathbf{G}^+| = 1224$ (compared to 27 and 729 for A). Feature structure E involves larger clusters and more overlap, as there are 8 stimuli in each cluster and 7 stimuli shared between all pairs of clusters, yielding $|\mathbf{G}_f| = 3430$. Once the additive constant is introduced, it is no longer possible to have larger features or more overlap without including the same feature twice (which is degenerate), and $|\mathbf{G}^+|$ for this representation is 8232. In this example at least, representations with smaller clusters are simpler than those with larger clusters, even though it comes at the expense of reduced overlap.

It is also worthwhile to note that, for a fixed number of clusters the simplest representation is one consisting only of clusters containing two stimuli. The complexity

matrix for this representation is the identity, and therefore has determinant 1. Since $\mathbf{G}$ is positive definite, its determinant must be positive, and since the elements of $\mathbf{G}$ are integers, no complexity matrix can ever have a determinant smaller than 1. This argument does not incorporate the additive constant, but it is heartening to note that a representation of nine stimuli using three two-stimulus clusters has $|\mathbf{G}^+| = 33$, making it simpler than any of those displayed in Figure 3.2.

In order to see if cluster size is the dominant contributing factor to complexity in the general case, the following brief evaluation was carried out. A random sample of 100,000 feature structures containing 6 features and 10 stimuli were generated[5], and their structural complexity, average cluster size, and average overlap were measured. The size of a cluster containing α stimuli was measured as the proportion of stimulus pairs that were encompassed by the feature, $\alpha(\alpha-1)/n(n-1)$. Similarly, the overlap between two features containing $\alpha \geq \beta$ stimuli, of which γ are encompassed by both, was measured as the number of stimulus pairs encompassed by both features, expressed as a proportion of the maximum possible (i.e., the number in the smaller feature), $\gamma(\gamma-1)/\beta(\beta-1)$.

Figure 3.4 shows the relationship between size, overlap and complexity for a representative subsample of 1000 of these feature structures. As previously suggested, size and overlap tend to covary, but it is clearly evident from the figure that increase in complexity due to size substantially outweighs the decrease due to increased overlap. In fact, the extent of the covariation between size and overlap makes it difficult to estimate the independent effect of overlap from the figure. Figure 3.5 plots the relationship between overlap and complexity for 800 feature structures out of the 100,000 with a constant average cluster size ($\approx$41.6%). Though the relationship is far from exact, it is evident that, as expected, increased overlap between features decreases complexity.

[5]The reason for sampling so extensively was to ensure that a large number of feature structures in the sample would have same average cluster size (see Figure 3.5).

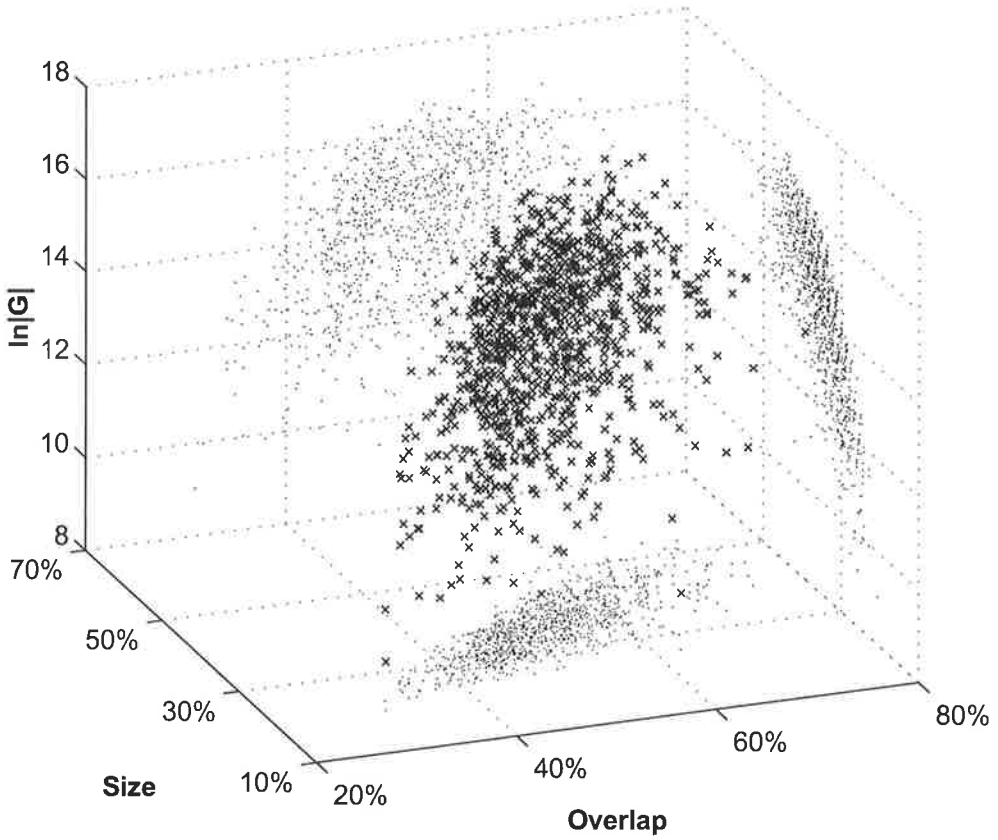

Figure 3.4: Structural complexity ($\ln |\mathbf{G}^+|$) for a random sample of 1000 feature structures ($n = 10$, $m = 6$) plotted (as crosses) as a function of average size and overlap of the features. Each pair of variables (i.e, size, overlap, and complexity) is also plotted (as dots).

3.3 Choosing an Additive Tree Representation

This section derives and interprets the GCC measure for additive tree representations. Strictly speaking, the measure can be applied to bidirectional trees as well as additive trees (see Section 2.4), though this discussion focusses on the latter. Applying the measure to extended trees is straightforward, though it is not done here.

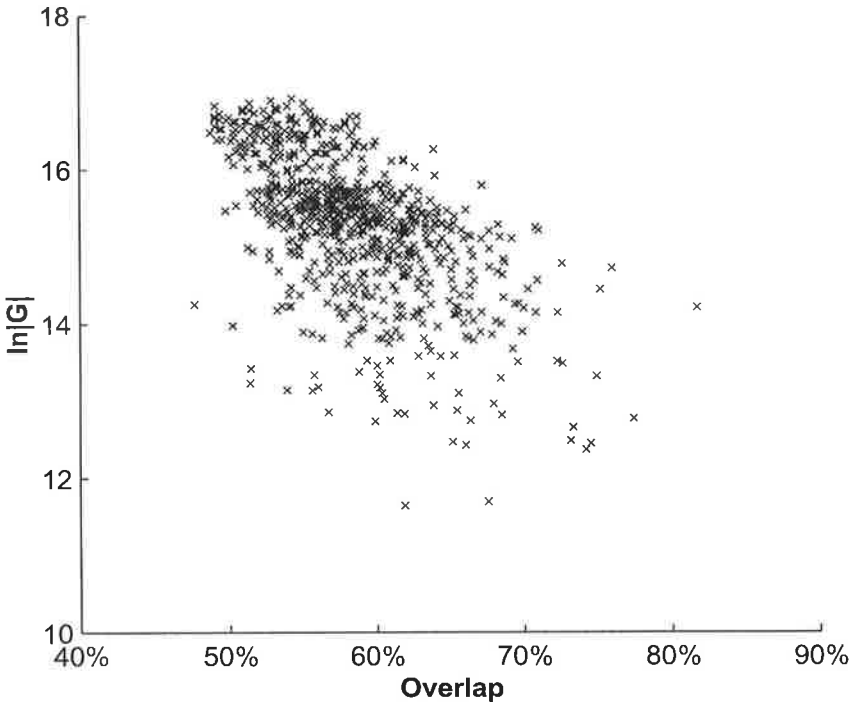

Figure 3.5: Structural complexity ($\ln |\mathbf{G}^+|$) for a sample of 823 feature strutures with constant average size (≈41.6%), and overlap varying from 47.7% to 81.7%.

3.3.1 GCC Derivation

Mathematically speaking, trees are a special case of featural representations (Corter, 1996), so it is hardly surprising that the derivation of the Geometric Complexity Criterion for additive trees parallels that for additive clustering models. The free parameters in a tree representation correspond to the lengths $\mathbf{l} = \{l_1, l_2, \ldots, l_{m+n-1}\}$ of the edges in the tree. Unlike additive clustering, no additive constant is required: in order to increase all similarity estimates by c, one can increase the lengths of all terminal edges by $\frac{c}{2}$. Additive trees have $m + n - 1$ edge lengths, though bidirectional trees may have up to $2(m + n - 1)$. The number of data points N is given by the number of independent entries in the proximity matrix $\mathbf{D}$, which for additive trees is $\frac{n(n-1)}{2}$ but for bidirectional trees (which use asymmetric proximity matrices) is $n(n-1)$. Note that for bidirectional

trees, every instance of $\sum_{i<j}$ should be replaced with $\sum_{i\neq j}$.

If $\mathbf{T}$ denotes the tree topology, then the GCC for additive trees is

$$\mathrm{GCC}_{\mathrm{tree}} = -\ln p(\mathbf{D}|\mathbf{T},\mathbf{l}^*)+\frac{m+n-1}{2}\ln\left(\frac{n(n-1)}{4\pi}\right)+\ln\int\sqrt{|\mathbf{I}(\mathbf{l})|}d\mathbf{l}+\frac{1}{2}\ln\left(\frac{|\mathbf{J}(\mathbf{l}^*)|}{|\mathbf{I}(\mathbf{l}^*)|}\right).$$

Assuming that the proximities d_{ij} are drawn from Gaussian distributions with mean $\hat{d}_{ij}$ and common variance σ, the likelihood of the data is given by

$$\begin{aligned} p(\mathbf{D}|\mathbf{T},\mathbf{l}) &= \prod_{i<j}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}\left(d_{ij}-\hat{d}_{ij}\right)^2\right) \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^{n(n-1)/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i<j}\left(d_{ij}-\hat{d}_{ij}\right)^2\right). \end{aligned}$$

Therefore, the negative log maximum likelihood is given,

$$-\ln p(\mathbf{D}|\mathbf{T},\mathbf{l}^*) = \frac{1}{2\sigma^2}\sum_{i<j}\left(d_{ij}-\hat{d}^*_{ij}\right)^2+\frac{n(n-1)}{2}\ln\left(\sigma\sqrt{\pi}\right).$$

As with the additive clustering case, the second-order partial derivatives

$$\frac{\partial^2\ln p(\mathbf{D}|\mathbf{T},\mathbf{l})}{\partial l_x\partial l_y}$$

are required. The first-order partial derivative is given by,

$$\begin{aligned} \frac{\partial\ln p(\mathbf{D}|\mathbf{T},\mathbf{l})}{\partial l_x} &= -\frac{1}{2\sigma^2}\sum_{i<j}2\left(d_{ij}-\hat{d}_{ij}\right)\frac{\partial}{\partial l_x}\left(d_{ij}-\hat{d}_{ij}\right) \\ &= -\frac{1}{\sigma^2}\sum_{i<j}\left(d_{ij}-\hat{d}_{ij}\right)\frac{\partial}{\partial l_x}\left(-\hat{d}_{ij}\right) \\ &= -\frac{1}{\sigma^2}\sum_{i<j}\left(\hat{d}_{ij}-d_{ij}\right)\frac{\partial}{\partial l_x}\hat{d}_{ij}. \end{aligned}$$

It is useful at this point to observe that the proximity model for stimuli in an additive tree can be written as

$$\hat{d}_{ij} = \sum_{k \in P_{ij}} l_k,$$

where P_{ij} denotes the set of edges that make up the unique path between the ith and jth stimuli. Correspondingly, the partial derivative of $\hat{d}_{ij}$ with respect to l_x is 1 if the xth edge belongs to P_{ij}, and 0 if it does not. Therefore,

$$\frac{\partial \ln p(\mathbf{D}|\mathbf{T}, \mathbf{l})}{\partial l_x} = -\frac{1}{\sigma^2} \sum_{i<j|l_x \in P_{ij}} \left(\hat{d}_{ij} - d_{ij}\right).$$

The second-order partial derivatives are thus given by,

$$\begin{aligned} \frac{\partial^2 \ln p(\mathbf{D}|\mathbf{T}, \mathbf{l})}{\partial l_y l_x} &= -\frac{1}{\sigma^2} \sum_{i<j|l_x \in P_{ij}} \frac{\partial}{\partial l_y} \left(\hat{d}_{ij} - d_{ij}\right) \\ &= -\frac{1}{\sigma^2} \sum_{i<j|l_x \in P_{ij}} \frac{\partial}{\partial l_y} \hat{d}_{ij}, \end{aligned}$$

and by applying the same argument to $\frac{\partial}{\partial l_y}\hat{d}_{ij}$,

$$\frac{\partial^2 \ln p(\mathbf{D}|\mathbf{T}, \mathbf{l})}{\partial l_y \partial l_x} = -\frac{1}{\sigma^2} \sum_{i<j|l_x, l_y \in P_{ij}} 1.$$

It is therefore apparent that the elements of the Hessian matrix $\mathbf{J}(\mathbf{l}) = [j_{xy}(\mathbf{l})]$ are given by,

$$\begin{aligned} j_{xy}(\mathbf{l}) &= -\frac{\partial^2 \ln p(\mathbf{D}|\mathbf{T}, \mathbf{l})}{\partial l_y \partial l_x} \\ &= \frac{1}{\sigma^2} \sum_{i<j|l_x, l_y \in P_{ij}} 1. \end{aligned}$$

Likewise, the Fisher Information Matrix $\mathbf{I}(\mathbf{l}) = [i_{xy}(\mathbf{l})]$ becomes

$$\begin{aligned} i_{xy}(\mathbf{l}) &= -E\left[\frac{\partial^2 \ln p(\mathbf{D}|\mathbf{T}, \mathbf{l})}{\partial l_y \partial l_x}\right] \\ &= \frac{1}{\sigma^2} \sum_{i<j|l_x, l_y \in P_{ij}} 1. \end{aligned}$$

As with additive clustering representations, both matrices are independent of $\mathbf{l}$, making the substitution $\mathbf{l} = \mathbf{l}^*$ trivial. Since $\mathbf{I}(\mathbf{l}^*) = \mathbf{J}(\mathbf{l}^*)$ the fourth term of the GCC is again 0.

Again, the third term of the GCC requires the integration of $\sqrt{|\mathbf{I}(\mathbf{l})|}$ over all parameterisations of the tree. Since dissimilarities are assumed to lie between 0 and 1, the natural parameter constraint is $0 \leq l_k \leq 1$. Furthermore, since $\sqrt{|\mathbf{I}(\mathbf{l})|}$ is constant with respect to $\mathbf{l}$, the integral is given,

$$\begin{aligned}
\ln \int \sqrt{|\mathbf{I}(\mathbf{l})|}\, d\mathbf{l} &= \ln \int_0^1 \int_0^1 \cdots \int_0^1 \int_0^1 \sqrt{|\mathbf{I}(\mathbf{l})|}\, dl_1\, dl_2 \ldots dl_{m+n-2}\, dl_{m+n-1} \\
&= \frac{1}{2} \ln |\mathbf{I}(\mathbf{l})| \\
&= \frac{1}{2} \ln \left| \frac{1}{\sigma^2} \times \mathbf{G} \right| \\
&= \frac{1}{2} \ln \left(\frac{1}{\sigma^{2(m+n-1)}} |\mathbf{G}| \right) \\
&= -(m+n-1) \ln \sigma + \frac{1}{2} \ln |\mathbf{G}|. \qquad (3.6)
\end{aligned}$$

In this case, $\mathbf{G} = [g_{xy}]$ denotes the $(m+n-1) \times (m+n-1)$ *complexity matrix* for additive trees, where g_{xy} counts the number of pairs of stimuli whose paths go through both the xth and yth edge in the tree. This matrix can be written as,

$$\mathbf{G} = \begin{bmatrix}
\sum_{i<j} t_{ij1} & \sum_{i<j} t_{ij1} t_{ij2} & \cdots & \sum_{i<j} t_{ij1} t_{ijm} \\
\sum_{i<j} t_{ij2} t_{ij1} & \sum_{i<j} t_{ij2} & \cdots & \sum_{i<j} t_{ij2} t_{ijm} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{i<j} t_{ijm} t_{ij1} & \sum_{i<j} t_{ijm} t_{ij2} & \cdots & \sum_{i<j} t_{ijm}
\end{bmatrix},$$

where t_{ijk} is 1 if the kth arc is on the unique path linking objects i and j, and 0 if it is not. Thus, the GCC for additive trees is given by:

$$\mathrm{GCC}_{\mathrm{tree}} = -\frac{1}{2\sigma^2} \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2 + \frac{m+n-1}{2} \ln \left(\frac{n(n-1)}{4\pi} \right) - (m+n-1) \ln \sigma$$

$$
\begin{aligned}
& + \frac{1}{2} \ln |\mathbf{G}| + \frac{n(n-1)}{2} \ln \left(\sigma \sqrt{2\pi} \right) \\
= & -\frac{1}{2\sigma^2} \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2 + \frac{m+n-1}{2} \ln \left(\frac{n(n-1)}{4\pi\sigma^2} \right) + \frac{1}{2} \ln |\mathbf{G}| \\
& + \frac{n(n-1)}{2} \ln \left(\sigma \sqrt{2\pi} \right).
\end{aligned}
$$

Reading left to right, it is again possible to view the GCC as the sum of an error term, a parametric complexity term, a structural complexity term, and the data constant. As in additive clustering, the data constant makes no contribution to model selection, and may be ignored. The GCC may therefore be written as,

$$
\mathrm{GCC}_{\mathrm{tree}} = -\frac{1}{2\sigma^2} \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2 + \frac{m+n-1}{2} \ln \left(\frac{n(n-1)}{4\pi\sigma^2} \right) + \frac{1}{2} \ln |\mathbf{G}| + \text{constant},
$$

thus completing the derivation.

3.3.2 Precision and Domain Size

The resemblances between $\mathrm{GCC}_{\mathrm{tree}}$ and $\mathrm{GCC}_{\mathrm{adclus}}$ mean that the effect of precision and domain size on $\mathrm{GCC}_{\mathrm{tree}}$ are very similar to their effects on $\mathrm{GCC}_{\mathrm{adclus}}$. That is, as data precision increases, the importance of a good fit increases relative to the importance of a parsimonious model. Additionally, greater precision means that the complexity of the model is evaluated more in terms of the number of edges in the tree, and less in terms of the tree topology. An n increases, the error term rises quadratically, and the structural complexity term rises logarithmically (as with additive clustering), but since there is an n at the front of the parametric term, that term rises linearly rather than logarithmically. Nevertheless, the net effect is that as the domain size (or "resolution") increases, data-fit becomes more important than parsimony.

3.3.3 The Complexity of Tree Topologies

A third source of variation in tree complexity is the topology of the tree itself $\mathbf{T}$, reflected in differences in $|\mathbf{G}|$. This section discusses the effect of $\mathbf{T}$ on complexity, providing

analytic results for one and two node trees, and a numerical analysis of more general tree structures.

Star Trees

It is instructive to first consider the single-node ($m = 1$) star tree (see Figure 3.6, left panel) in the first instance, since it has the property that all $n - 1$ edges in the tree are terminal arcs. The terminal edge that connects the ith stimulus to the tree must belong to each of the $n - 1$ paths that connects the ith stimulus to another stimulus, but will not belong to any other path. Hence every element of the main diagonal of $\mathbf{G}$ is $n - 1$. Recall that an off-diagonal element of $\mathbf{G}$ counts the number of pairs of stimuli whose unique connecting path passes though both the xth and yth edges. Suppose that the path between stimuli i and j meets this criterion. This implies that the path must terminate in both the ith and jth stimuli, and hence that the criterion will only hold once. Therefore, all off-diagonal elements are 1, and the $n \times n$ complexity matrix for a star tree is

$$\mathbf{G}_{\text{star}} = \begin{bmatrix} n-1 & 1 & 1 & & 1 \\ 1 & n-1 & 1 & & 1 \\ 1 & 1 & n-1 & & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & & n-1 \end{bmatrix},$$

which has inverse

$$\mathbf{G}_{\text{star}}^{-1} = \frac{1}{2(n-1)(n-2)} \begin{bmatrix} 2n-3 & -1 & -1 & \cdots & -1 \\ -1 & 2n-3 & -1 & \cdots & -1 \\ -1 & -1 & 2n-3 & \cdots & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & -1 & \cdots & 2n-3 \end{bmatrix}. \tag{3.7}$$

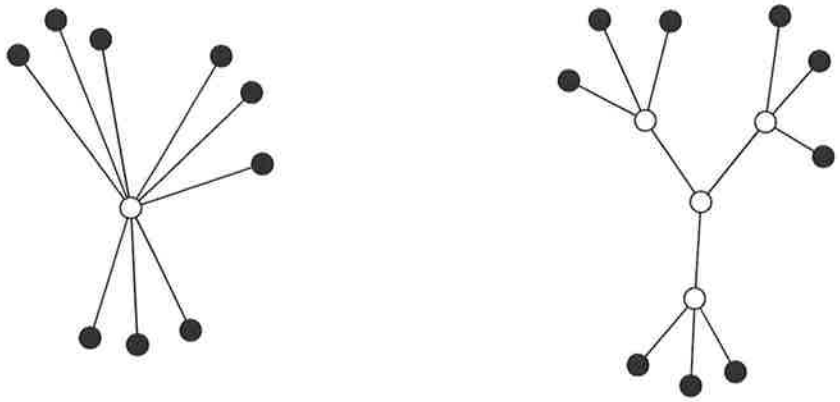

Figure 3.6: A star tree (left) and an additive tree containing non-terminal edges (right). There can only ever be one pair of stimuli whose path contains any given pair of terminal nodes.

Although it is of little use from a model selection standpoint to have an analytic expression for the complexity of star trees, since for any n there is only a single one, this result will be of use shortly in considering other trees.

2-Node Trees

Before considering more general tree structures, it is helpful to consider trees containing a single internal edge ($m = 2$). Once again, there are $(n-1)$ paths that pass through a terminal edge, and there is only ever a single path passing through two different terminal edges. Indeed, this will be true of any tree structure, because the number of paths that pass through a terminal edge (or pair of such edges) is not affected by any internal edges that might lie between them, which is visually apparent from inspection of Figure 3.6. Therefore, $\mathbf{G}$ can be written as

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{\text{star}} & \mathbf{y} \\ \mathbf{y}' & z \end{bmatrix},$$

where z counts the number of paths that pass through an internal edge, and $\mathbf{y}$ is the column vector $\{y_1\ y_2 \ldots\ y_n\}'$ such that y_i counts the number of paths that pass through

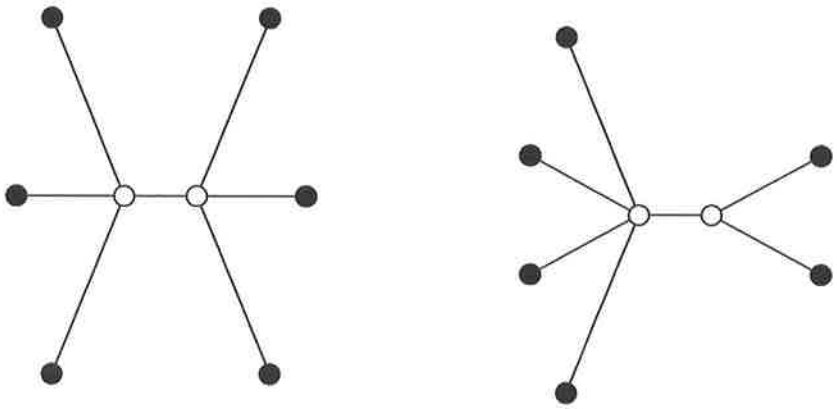

Figure 3.7: The two possible trees containing six stimuli and one internal edge.

both the single internal edge and the ith terminal edge. Therefore, the determinant of the complexity matrix is given by

$$|\mathbf{G}| = |\mathbf{G}_{\text{star}}|(z - \mathbf{y}'\mathbf{G}_{\text{star}}^{-1}\mathbf{y}).$$

For the moment, consider the simplest case, in which one non-terminal edge is added to a star tree containing 6 stimuli. There are exactly two non-equivalent non-degenerate possibilities, shown in Figure 3.7. The tree on the left divides the stimuli into two groups of three. Each of the $3 \times 3 = 9$ paths that cross from one group to the other involve the internal edge, so in this case $z = 9$. Each terminal edge shares with the internal edge all paths starting at that terminal edge and cross from one side of the tree to the other, so in this case $\mathbf{y} = [3\ 3\ 3\ 3\ 3\ 3]'$. Substituting these numbers gives $z - \mathbf{y}'\mathbf{G}_{\text{star}}^{-1}\mathbf{y} = 9 - 5.4 = 3.6$. Similarly, when considering the tree on the right, $z = 2 \times 4 = 8$, $\mathbf{y} = [2\ 2\ 2\ 2\ 4\ 4]'$, and $z - \mathbf{y}'\mathbf{G}_{\text{star}}^{-1}\mathbf{y} = 8 - 5.6 = 2.2$. The tree on the right is less complex.

More generally, when splitting the star tree to include an internal edge, one divides the n stimuli into two groups, one containing r and the other containing $n - r$ stimuli. The number of paths that pass through the new edge is $z = r(n - r)$. Furthermore, the

value of the first r elements of $\mathbf{y}$ is $n-r$ and value of the remaining $n-r$ elements is r. Using the formula for $\mathbf{G}_{\text{star}}^{-1}$ given by Eq. 3.7 the expression for $z - \mathbf{y}'\mathbf{G}_{\text{star}}^{-1}\mathbf{y}$ for two-node trees is

$$\begin{aligned} z - \mathbf{y}'\mathbf{G}_{\text{star}}^{-1}\mathbf{y} &= r(n-r) - \frac{r(n-r)\left(n(n-1) - 2r(n-r)\right)}{(n-1)(n-2)} \\ &= r(n-r)\left(1 + \frac{2r(n-r)}{(n-1)(n-2)} - \frac{n}{n-2}\right). \end{aligned}$$

Inspection of this expression reveals that $z - \mathbf{y}'\mathbf{G}_{\text{star}}^{-1}\mathbf{y}$ increases with $r(n-r)$. This is shown visually in Figure 3.8, which plots the value of $z - \mathbf{y}'\mathbf{G}_{\text{star}}^{-1}\mathbf{y}$ as a function of n and $\frac{r}{n}$. It is important to recognise that, although the function is defined for all r and n, and that for reasons of clarity the graph has been plotted as a continuous function, only integer values of r and n have a meaningful interpretation as tree complexities. That said, the figure confirms the claim that the simplest configuration for a tree with a single internal edge is achieved by minimising $r(n-r)$. This is accomplished by producing the maximum disparity in the size of the two groups without introducing a degenerate structure. This minimum corresponds to a tree in which two stimuli are grouped together on one side of the new edge, and the rest of the stimuli form the other group.

General Tree Structures

In order to discuss tree structures with multiple internal edges, it is again useful to partition $\mathbf{G}$. The $(n+m-1) \times (n+m-1)$ complexity matrix of any tree structure can always be partitioned into

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{\text{star}} & \mathbf{Y} \\ \mathbf{Y}' & \mathbf{Z} \end{bmatrix}.$$

The lower right submatrix $\mathbf{Z}$ is the $(m-1) \times (m-1)$ matrix whose rows and columns correspond to internal edges, and whose elements represent the interaction between pairs

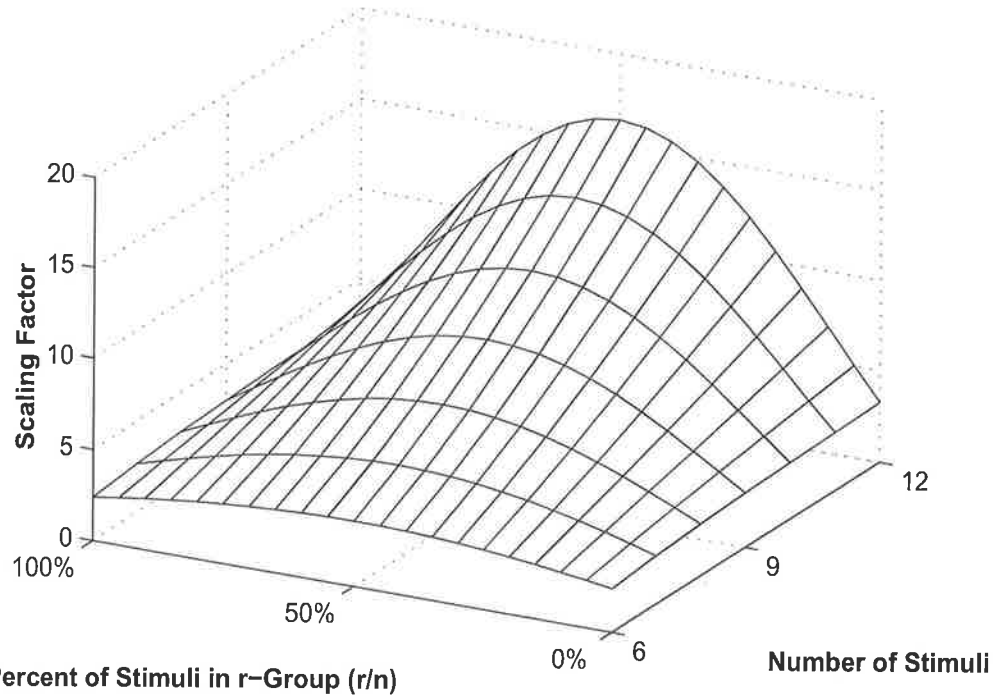

Figure 3.8: The value of the scaling factor $z - \mathbf{y}'\mathbf{G}_{\text{star}}^{-1}\mathbf{y}$ for two-internal node trees as a function of the number of stimuli n and the proportion of those stimuli that belong to one of the two subgroups $\frac{r}{n}$.

of internal edges. Similarly, the off-diagonal matrix $\mathbf{Y}$ denotes the $n \times (m-1)$ matrix with rows corresponding to terminal edges and columns corresponding to internal edges, and whose elements represent interactions between terminal and internal edges. Because $\mathbf{G}_{\text{star}}$ is invariant across trees with the same number of stimuli n, the determinant

$$|\mathbf{G}| = |\mathbf{G}_{\text{star}}|.|\mathbf{Z} - \mathbf{Y}'\mathbf{G}_{\text{star}}^{-1}\mathbf{Y}|$$

is dependent only on $\mathbf{Y}$ and $\mathbf{Z}$.

The following evaluation was carried out to demonstrate the effect of tree topology on complexity: all possible tree structures with 5 to 10 internal nodes were generated, such that all internal nodes were connected to either 2, 3 or 4 stimuli. The value of $\ln|\mathbf{G}|$ was then calculated for each tree. For a given number of internal nodes, it was found

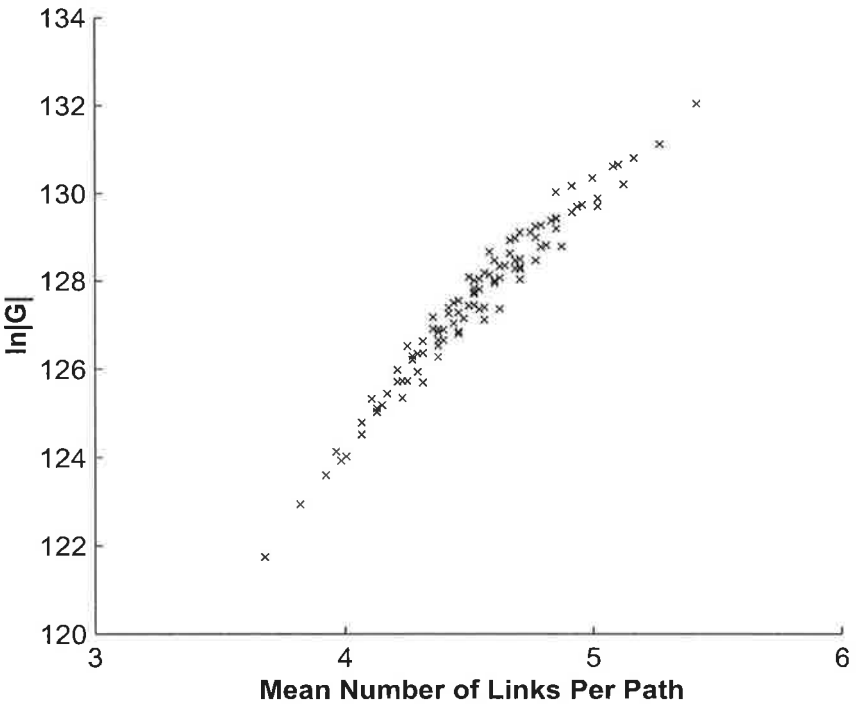

Figure 3.9: Structural complexity ($\ln |\mathbf{G}|$) for all possible 10-node additive trees with 3 stimuli per node, plotted against mean links per path. Trees with short paths traversing them are simpler than trees with long paths: the correlation between complexity and average path length is 0.98.

that $\ln |\mathbf{G}|$ increased linearly with the average number of edges in the paths connecting stimuli, though the relationship was not exact. The shape of the relationship was not affected either by the number of internal nodes, or the number of stimuli per node. Figure 3.9 shows the relationship for 10-node trees with three stimuli per node. The correlation between path length and $\ln |\mathbf{G}|$ was never less than 0.97 in this evaluation. It appears that trees with longer path lengths are more complex: the most and least complex 10-node trees with 3 stimuli per node are shown in Figure 3.10.

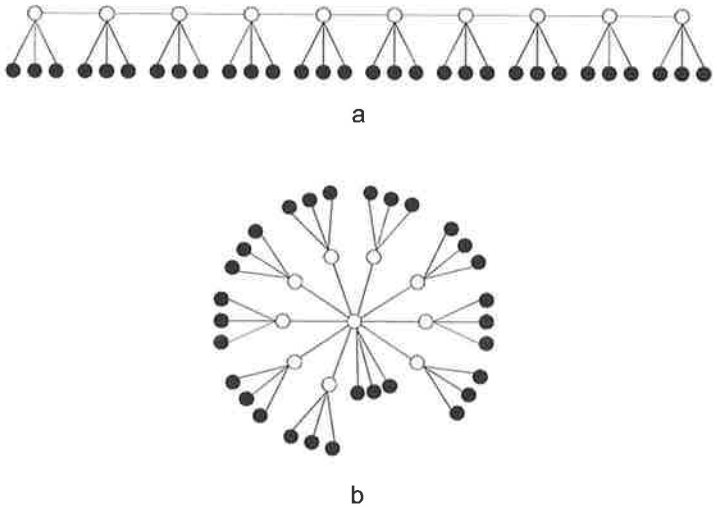

Figure 3.10: Topologies for the most complex (a) and the least complex (b) of the trees displayed in Figure 3.9.

3.4 Choosing a Spatial Representation

As discussed in Section 2.2 the similarity model for metric spatial representations typically assumes that the dissimilarity between two stimuli is given by the distance between their co-ordinates in the appropriate psychological space. It is usual to measure this distance using one of the Minkowski metrics,

$$\ddot{d}_{ij} = \left(\sum_k |p_{ik} - p_{jk}|^r \right)^{\frac{1}{r}} + c,$$

where p_{ik} denotes the co-ordinate value of the ith stimulus on the kth dimension, r denotes the choice of metric (with $r = 2$ corresponding to the Euclidean metric and $r = 1$ corresponding to the City Block distance), and c is an additive constant. The free parameters of a spatial representation are the co-ordinate values $\mathbf{p} = \{p_{11}, p_{21}, \ldots, p_{nm}\}$ and the additive constant c. However, since distances under the Minkowski metrics are translation-invariant, the first point can be fixed at the origin. An m-dimensional spatial representation of n stimuli therefore contains $m(n-1)+1$ parameters, and the GCC

measure takes the form,

$$\begin{aligned} \mathrm{GCC}_{\mathrm{mds}} &= -\ln p(\mathbf{D}|\mathbf{p}^*) + \frac{m(n-1)+1}{2}\ln\left(\frac{n(n-1)}{4\pi}\right) \\ &\quad + \ln\int\sqrt{|\mathbf{I}(\mathbf{p})|}d\mathbf{p} + \frac{1}{2}\ln\left(\frac{|\mathbf{J}(\mathbf{p}^*)|}{|\mathbf{I}(\mathbf{p}^*)|}\right) \end{aligned}$$

Finding expressions for $\mathbf{I}(\mathbf{p})$ and $\mathbf{J}(\mathbf{p})$, as well as the integral $\int\sqrt{|\mathbf{I}(\mathbf{p})|}d\mathbf{p}$ proved to be difficult. Numerical approximation to the GCC for spatial representations yielded problematic results. Further investigations in this area may be warranted.

3.5 Summary & General Discussion

The aim in this chapter has been to apply statistically well-founded model selection ideas to the field of similarity modelling. The purpose of this endeavour is to allow representational modelling to proceed with an understanding of what makes one representation more or less complex than another, and to extract representations that account for the data in a parsimonious manner. In doing this, expressions for the Geometric Complexity Criterion have been derived for additive clustering and additive tree representations. Analyses of these measures has shed light on the nature of representational complexity within these frameworks, though further analysis is certainly possible. The difficulties experienced in accounting for spatial complexity are troubling, and this remains an area for future work. Finally, although this point is hopefully obvious, it should be kept in mind that although the GCC is a superior measure to data-fit or simple criteria such as the BIC, it should be treated as an aid to scientific judgement, rather than a substitute for it. The GCC does not incorporate all of the criteria by which a model should be judged, and should therefore be used as a guide rather than an inflexible rule. Nevertheless, the GCC is a highly effective quantitative measure, and has the potential to guide similarity modelling in a principled manner.

4. Featural Representation

As discussed in Section 2.3, featural representation involves describing a stimulus in terms of a set of discrete characteristics. If these characteristic features are perceptual in nature, then the representation can be thought of as a description of the stimulus in terms of its constituent parts. If they are more conceptual in nature, the representation may look more like a list of categories to which the stimulus belongs. Alternatively, some features may be perceptual and others conceptual, a state of affairs that probably reflects the norm, and need not present a theoretical difficulty.

This chapter considers this notion of featural representation on a number of fronts[1]. The most fully developed featural model, as far as deriving representations from similarity data is concerned, is the additive clustering model. The majority of this chapter is taken up developing three other featural models to a comparable level, and providing a detailed evaluation of all four candidate models. While two of the "new" models have been considered by other authors, they have not previously been used as clustering models, whereas the third model is completely novel. The evaluation of these models involves fitting pre-existing data, two Monte Carlo studies and two experimental studies. The remainder of this chapter is devoted to the discussion of other classes of featural models besides the four models evaluated here.

[1]Much of the work in this chapter appears in Navarro and Lee (2001, in press, submittedb).

4.1 A Menagerie of Featural Models

While discussing the topic of featural representation, it is useful to distinguish between the psychological problem of featural similarity and the numerical problem of finding features. The psychological problem is a similarity modelling problem: given a set of features, how should similarities be estimated? In contrast, the numerical problem is a data fitting problem: given a set of data, and assuming a particular psychological model, what set of features most probably gave rise to the data? For the moment, the numerical problem is disregarded, and the discussion focusses on the psychological issues. Therefore, this section presents four candidate models that express similarity as a function of a set of binary features.

The appropriate place to begin is with the common features model and the distinctive features model (Sattath & Tversky, 1987), which complement one another. Under a *common features model*, the similarity between two stimuli is a monotonically increasing function Λ of the features that they share. That is,

$$\hat{s}_{ij} = \Lambda(\mathbf{f}_i \cap \mathbf{f}_j) + c. \tag{4.1}$$

where $\mathbf{f}_i$ denotes the set of features possessed by stimulus i, and c is a non-negative constant added to all similarity estimates. The *distinctive features model* assumes that two stimuli become more dissimilar as a function of the number of features possessed by only one of them. This model takes the form,

$$\hat{s}_{ij} = c - \Upsilon(\mathbf{f}_i - \mathbf{f}_j) - \Upsilon(\mathbf{f}_j - \mathbf{f}_i). \tag{4.2}$$

where Υ, like Λ, is monotonically increasing. The distinctive features model was originally proposed by Restle (1959) who referred to it as the symmetric distance metric, and used it as a psychologically plausible distance metric for sets. It is closely related

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

Sattath and Tversky (1987) present a proof that any set of data generated by a common features model can also be generated by some distinctive features model, and vice versa. Their discussion focussed on dissimilarity rather than similarity, but the argument does not rely on this point. The essence of the proof is to show that if a set of features $\mathbf{F}_1$ produces a similarity matrix $\hat{\mathbf{S}}$ under one model, then there exists a second set of features $\mathbf{F}_2$ that produces $\hat{\mathbf{S}}$ under the other model. Consequently, they argue that there is nothing inherent in the data to distinguish between common features models and distinctive features models.

Nevertheless, there are two elements of the proof that deserve close examination and suggest caution in interpreting the result. Firstly, the proof requires that for each stimulus i, there exists a complementary feature $\bar{\mathbf{f}}_i$[2] that is possessed by all stimuli except i. This can be achieved by adding "dummy" features to $\mathbf{F}_1$, obtained by taking the intersection, union or complement of existing features. Thus, although the extended feature set $\mathbf{F}_2$ consists of the original features $\mathbf{F}_1$ plus the new complementary features, no free parameters are introduced (that is, $\mathbf{F}_2$ has the same rank as $\mathbf{F}_1$).

The second important feature of the proof is that it relies on being able to define different functional forms for Λ and Υ (if $\Lambda \equiv \Upsilon$ the proof fails). The relationship between Λ and Υ is trivial for all similarity ratings that do not involve the complementary features in $\mathbf{F}_2$, but involves a substantial change for ratings that do involve these new features. The consequence of these properties is that although $\mathbf{F}_1$ and $\mathbf{F}_2$ have the same number of free parameters, the difference in functional form implies that complexity may nevertheless differ, providing a quantitative means to distinguish between a common features model and distinctive features model, even given their equivalent data fit, as

[2]A brief note on nomenclature: $\mathbf{f}_i$ has been used to refer to the set of features possessed by the ith stimulus (which is a column in $\mathbf{F}$). However, $\bar{\mathbf{f}}_i$ denotes a feature possessed by all stimuli except the ith (and is hence a row from $\mathbf{F}$).

discussed in Chapter 3. Therefore, there remain quantitative differences between the common features model and the distinctive features model.

On a separate but equally important note, the proof is difficult to interpret in psychological terms. Consider a building that possesses the characteristics "dome shaped" ($\mathbf{f}_{\text{dome}}$) and "made of ice" ($\mathbf{f}_{\text{ice}}$), both of which are denoted by elementary features in $\mathbf{F}$. This building's complementary feature would be something along the lines of "not made of ice and not dome shaped" ($\bar{\mathbf{f}}_{\text{dome}} \cap \bar{\mathbf{f}}_{\text{ice}}$). The problematic aspect is that in order to derive equivalent common features and distinctive features models, decisions made using "not made of ice and not dome shaped" must use a different rule than decisions made using "dome shaped" or "made of ice". There is no compelling reason to believe that composite features are evaluated in a manner that is fundamentally different to those features of which they are composed. Furthermore, since most features can be viewed as a composite of lower level features, it is unclear which features should be evaluated according to which rule.

Looking beyond the two models considered so far, there is some evidence to suggest that some combination of the common features approach and the distinctive features approach is warranted. A series of studies (Gati & Tversky, 1987, 1987; Ritov, Gati, & Tversky, 1990; Tversky, 1977; Tversky & Gati, 1978) investigated the contributions of common and distinctive features in a number of ways. One such method was to take a pair of easily manipulated stimuli (such as schematic drawings of faces) and add a feature (such as a pair of glasses) as either a common feature or as a distinctive feature, and measured the effect on similarity ratings. Unsurprisingly, they found that both common features and distinctive features affected the similarity judgements. Letting $C(\mathbf{f})$ denote the impact of some feature $\mathbf{f}$ when introduced as a common feature, and $D(\mathbf{f})$ denote its effect as a distinctive feature, they measured the relative impact of $\mathbf{f}$ as $W(\mathbf{f}) = \frac{C(\mathbf{f})}{C(\mathbf{f})+D(\mathbf{f})}$ and were able to measure $W(\mathbf{f})$ for a range of different features. It

is important to recognise that such studies are required to use contrived stimuli so that the researchers are able to know (or assume) in advance what the underlying featural representation is. Although this is not unproblematic (e.g., Brooks, 1991; Goodman, 1972; Komatsu, 1992), measures of independence provide some protection against unwarranted assumptions, and the sheer number of experiments (30 in the list of papers cited above) is reassuring.

Such concerns notwithstanding, the results suggest $W(\mathbf{f})$ is quite variable, although this variation is to some extent orderly (see Shannon, 1988). For instance, Gati and Tversky (1984) found reliable differences in $W(\mathbf{f})$ for stimuli presented in written form and stimuli presented pictorially. In all written-stimulus experiments the median value for $W(\mathbf{f})$ was greater than $\frac{1}{2}$, ranging from .56 to .87. Experiments using pictorial stimuli elicited a very different pattern, with the median $W(\mathbf{f})$ ranging from .06 to .35. It appears that people judge the similarity of verbal stimuli using a combination of common and distinctive features that draws more heavily on the common features model, whereas the pictorial stimuli were judged using a model biased more towards distinctive features. Although it could be argued that this effect results from other causes, Ritov et al. (1990) present experiments that provide some evidence against such explanations (but see Keren, 1990).

There is some evidence that when shown two highly similar objects – that is, objects with many common features and few distinctive features – people "tend to take the shared features for granted and to focus on the distinctive features. On the other hand, in the comparison of dissimilar objects, ... people tend to take the differences for granted and to focus on the common features" (Ritov et al., 1990, pp 30-31). This was directly measured by Gati and Tversky (1984), by adding multiple common features or distinctive features, and comparing the results to those obtained by adding a single feature. The results suggested that a feature had less impact as a second addition than as a sole

addition. This could be explained as a shift of attention from common features to distinctive features (and vice versa), or alternatively as evidence that Λ is subadditive (i.e. $\Lambda(a+b) < \Lambda(a) + \Lambda(b)$). It is not the objective here to resolve the issue, merely to observe the possibility that increasing or decreasing similarity *may* itself influence the weighting of common and distinctive features.

One general framework for assessing featural similarity that accounts for common and distinctive features is *Tversky's Contrast Model* (Tversky, 1977; see also Gati & Tversky, 1984). The Contrast Model consists of three terms in a weighted sum: the common features term, $\mathbf{f}_i \cap \mathbf{f}_j$, and the two distinctive features terms $\mathbf{f}_i - \mathbf{f}_j$ and $\mathbf{f}_j - \mathbf{f}_i$. Thus the similarity estimate is given,

$$\hat{s}_{ij} = \theta\Lambda(\mathbf{f}_i \cap \mathbf{f}_j) - \alpha\Upsilon(\mathbf{f}_i - \mathbf{f}_j) - \beta\Upsilon(\mathbf{f}_j - \mathbf{f}_i) + c,$$

where Λ and Υ are monotonically increasing functions and θ, α, and β are non-negative hyper-parameters that assign weights to each of the terms. This version of the Contrast Model, used by Gati and Tversky (1984), does not require the common features component to have the same functional form as the distinctive features component. However, Tversky's (1977) original formulation did impose this restriction, and the model becomes

$$\hat{s}_{ij} = \theta\Lambda(\mathbf{f}_i \cap \mathbf{f}_j) - \alpha\Lambda(\mathbf{f}_i - \mathbf{f}_j) - \beta\Lambda(\mathbf{f}_j - \mathbf{f}_i) + c. \tag{4.3}$$

It has already been argued that setting $\Lambda \equiv \Upsilon$ provides an important psychological and quantitative constraint on featural representations. If a particular feature can be used as both a common and a distinctive feature, it is intuitive to assume that the same function would be used for both. In short, it makes sense to evaluate common features and distinctive features using the same functional form. If nothing else, Sattath and Tversky's (1987) proof demonstrates that without this axiom, the Contrast Model is underspecified, with each of its components able to substitute for the other. It is for this

reason that Tversky's (1977) original model is the one used in this chapter.

Tversky's Contrast Model is committed to the assumption that the balance between common features and distinctive features is invariant across features, since the weighting hyper-parameters θ, α and β are applied equally to all. This presents an intuitive difficulty with the theory, in that it seems unlikely that all features are evaluated in the same manner. Fortunately, this model is not the only plausible way of striking a balance between common and distinctive features. One method by which to address this concern might be to assign parameter values for each feature. The problem with such an approach is that it would lead to a proliferation of free parameters, which is undesirable. Instead, a more parsimonious similarity model is proposed, according to which a feature is declared to be either a common feature (which increases the similarity of pairs of stimuli that both possess 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* takes the form,

$$\hat{s}_{ij} = \Lambda(\mathbf{f}_i^{\mathbf{c}} \cap \mathbf{f}_j^{\mathbf{c}}) - \Lambda(\mathbf{f}_i^{\mathbf{d}} - \mathbf{f}_j^{\mathbf{d}}) - \Lambda(\mathbf{f}_j^{\mathbf{d}} - \mathbf{f}_i^{\mathbf{d}}) + c. \tag{4.4}$$

The $\mathbf{f}^{\mathbf{c}}$ and $\mathbf{f}^{\mathbf{d}}$ terms refer to the set of common features and the set of distinctive features respectively.

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". Gender is a good example of a distinctive feature: two people are not necessarily more alike because they are of the same gender, but they are less similar if they are of different genders. In contrast, hair colour makes sense as a common feature: two people with the same hair colour look

more alike (e.g., two people with blonde hair look more similar), but differences in hair colour are less important (e.g., someone with brown hair need not look dissimilar to someone with blonde hair). 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 Modified Contrast Model, 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 Tversky's model. That is, where Tversky's Contrast Model assumes that common and distinctive features are weighted during the decision process, the Modified Contrast Model 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 Modified Contrast Model 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 Modified Contrast Model treats these two aspects as two distinct featural regularities.

4.2 Clustering Models

The psychological models discussed so far are quite general in scope. Returning to the matter of finding plausible models of similarity that can be used by a clustering algorithm, it is easy to see that the additive clustering model

$$\hat{s}_{ij} = \sum_k w_k f_{ik} f_{jk} + c \tag{4.5}$$

is an example of a *common features clustering model*, in which each feature is assigned a saliency weight, and Λ denotes the sum-of-saliencies (plus an additive constant) function. Given that there is good reason to assume $\Upsilon \equiv \Lambda$, the corresponding *distinctive features clustering model* is

$$\hat{s}_{ij} = c - \frac{1}{2}\sum_k w_k f_{ik}\left(1 - f_{jk}\right) - \frac{1}{2}\sum_k w_k \left(1 - f_{ik}\right) f_{jk}. \tag{4.6}$$

Regarding Tversky's (1977) Contrast Model, the same functional form is applied, but there is the additional matter of the three hyper-parameters α, β and θ. Since this discussion of similarity is restricted to symmetric data (i.e. $s_{ij} = s_{ji}$), it is safe to assume that $\alpha = \beta$. Moreover, it may be assumed without loss of generality that $\theta + \alpha + \beta = 1$, since the saliency weights derived during clustering are automatically scaled to maximise data-fit. Allowing the hyper-parameters the freedom to sum to an arbitrary number merely leaves the clustering model underdetermined. Therefore, by setting $\theta = \rho$ and $\alpha = \beta = \frac{1-\rho}{2}$ such that $0 \leq \rho \leq 1$, an appropriate clustering model based on Tversky's Contrast Model is given by

$$\hat{s}_{ij} = \rho\sum_k w_k f_{ik} f_{jk} - \frac{1-\rho}{2}\sum_k w_k f_{ik}\left(1 - f_{jk}\right) - \frac{1-\rho}{2}\sum_k w_k \left(1 - f_{ik}\right) f_{jk} + c. \tag{4.7}$$

Since it is based on the Tversky's Contrast Model, it is referred to as the *Tversky Clustering Model* (TCM). Note that the additive clustering model results when $\rho = 1$, and the distinctive features clustering model results when $\rho = 0$. In the TCM, ρ denotes the overall balance between common and distinctive features, and is a direct analogue of the empirical W measure used by Gati and Tversky (1984, 1987).

Finally, the same functional form is applied to the Modified Contrast Model. The resulting *Modified Clustering Model* (MCM) is thus given by,

$$\hat{s}_{ij} = \sum_{k\in \mathrm{CF}} w_k f_{ik} f_{jk} - \frac{1}{2}\sum_{k\in \mathrm{DF}} w_k f_{ik}\left(1 - f_{jk}\right) - \frac{1}{2}\sum_{k\in \mathrm{DF}} w_k \left(1 - f_{ik}\right) f_{jk} + c, \tag{4.8}$$

where $k \in \mathrm{CF}$ implies that the sum is taken over the common features, and $k \in \mathrm{DF}$ means that only distinctive features are considered. It is important to note that the status

of a feature as a common feature or a distinctive feature is not a free parameter in the MCM: it is a fixed structural property of the model, like the elements of the feature matrix $\mathbf{F}$. The additive clustering model results when all the features are common and the distinctive features model results when all the features are distinctive.

4.3 Geometric Complexity Criteria

The derivation of GCC expressions for the four featural similarity models discussed in this chapter is identical in form to the derivation for additive clustering models in Section 3.2, except that the value of $\frac{\partial \hat{s}_{ij}}{\partial w_x}$ is different for each of the four models, though it always constant with respect to $\mathbf{w}$. Since every featural representation considered in this chapter has an additive constant, the GCC formula is

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

irrespective of whether the clustering model is the common features model, the distinctive features model, the TCM or the MCM. The difference between the four models lies in the expression for the $(m+1)\times(m+1)$ *complexity matrix* $\mathbf{G}=[g_{xy}]$, where

$$g_{xy} = \sum_{i<j}\left(\frac{\partial \hat{s}_{ij}}{\partial w_x}\times\frac{\partial \hat{s}_{ij}}{\partial w_y}\right).$$

Ignoring the additive constant c for the moment, consider the common features model, which assumes that similarity is the sum of shared saliencies. Therefore, the only term that is not constant with respect to w_x is the xth term of the sum, $w_x f_{ix} f_{jx}$, so the partial derivative is

$$\frac{\partial \hat{s}_{ij}}{\partial w_x} = f_{ix} f_{jx}.$$

Turning to the distinctive features model, the same logic eliminates all but the xth terms of the two sums, and therefore

$$\frac{\partial \hat{s}_{ij}}{\partial w_x} = -\frac{1}{2} f_{ix}(1 - f_{jx}) - \frac{1}{2}(1 - f_{ix}) f_{jx}.$$

The partial derivative for the TCM is given by

$$\frac{\partial \hat{s}_{ij}}{\partial w_x} = \rho f_{ix} f_{jx} - \frac{1-\rho}{2} f_{ix}(1 - f_{jx}) - \frac{1-\rho}{2}(1 - f_{ix}) f_{jx},$$

whereas for the MCM it is

$$\frac{\partial \hat{s}_{ij}}{\partial w_x} = \begin{cases} f_{ix} f_{jx} & \text{if } x \text{ is common} \\ -\frac{1}{2} f_{ix}(1 - f_{jx}) - \frac{1}{2}(1 - f_{ix}) f_{jx} & \text{if } x \text{ is distinctive} \end{cases}$$

For the TCM, $\frac{\partial \hat{s}_{ij}}{\partial w_x}$ is a weighted sum of the additive clustering and distinctive features clustering expressions for $\frac{\partial \hat{s}_{ij}}{\partial w_x}$, and for the MCM, $\frac{\partial \hat{s}_{ij}}{\partial w_x}$ always reduces to either the additive clustering expression or the distinctive features expression. It should be noted that if the parameter happens to be the additive constant c (rather than one of the saliency weights), then the expression $\frac{\partial \hat{s}_{ij}}{\partial c}$ is always 1, irrespective of which clustering model is used.

4.4 Algorithms for Fitting Featural Models

The algorithms used to derive representations in this chapter are based on the additive clustering algorithm proposed by Lee (in press, see Section 2.3.4). These algorithms examine potential representations one at a time, and search through the space of possible representations for the one that minimises the GCC. Therefore, they maintain a feature matrix $\mathbf{F}$ for the representation currently being considered. In the case of the MCM, a

binary-valued vector of "feature types" is also maintained, denoting which features are common and which are distinctive.

The algorithms initially specify a single-cluster representation, which is optimised (in the sense of finding the representation with minimum GCC) by employing the stochastic hillclimbing procedure described in Section 2.3.4. In the case of MCM representations, the elements of the feature types vector as well as the feature matrix are optimised by the stochastic hillclimbing. Once this process terminates, a new (randomly generated) cluster is added, and this solution is used as the starting point for a new optimisation procedure. As features are added, the representations become increasingly more complex. Therefore, at some point the increased data-fit achieved by adding features will no longer justify the increased complexity, and the GCC will start to deteriorate. The algorithms terminate once the GCC of the best representation with x features is sufficiently (e.g., 10 points; see Table 3.1) worse than the GCC of the best representation encountered during the entire search, where x is the number of features that have been added so far. The representation returned is the one with the best GCC.

Note that the TCM algorithm requires ρ to be specified in advance, because it not considered to be a parameter of the *representation*. The ρ-values for the TCM were found by running the algorithm across the full range of possible ρ-values (i.e., by gridsearch). This is consistent with the view that the underlying representation consists of the feature structure $\mathbf{F}$ and the saliency weights $\mathbf{w}$, and that ρ relates to the decision process. It would be a small matter to modify the algorithm to automatically derive the most appropriate ρ-value, though if ρ is considered to be a model parameter, then the GCC derivation should accommodate this.

4.5 Monte Carlo Study I: Do the Algorithms Work?

Prior to the use of any algorithm for the analysis of empirical data, it is general practice to demonstrate its ability to recover known structures from artificial data containing some level of noise. Therefore a small Monte Carlo study was undertaken to demonstrate that the GCC-driven stochastic hillclimbing algorithms developed in the previous sections do recover known representations from artificial data. However, such an investigation deals only with the numerical "representation recovery" problem, not the psychological "similarity modelling" problem. It is easily possible (perhaps commonplace) to have good algorithms fitting an inappropriate psychological model, or to fit a good model using poor algorithms. The purpose of this section is to check that the algorithms used in this chapter are good ones, in order to discuss the validity of the psychological models in subsequent sections. In order to differentiate between psychological models, this sort of Monte Carlo evaluation is next to useless. The true test of a model is how it deals with real data: the psychological evidence provided by Monte Carlo studies is minimal.

4.5.1 Method

Eight similarity matrices were used for this evaluation, by adding Gaussian noise ($\mu = 0$, $\sigma = .05$) to the similarity values generated by an underlying "true representation", subject to the constraint that $0 \leq s_{ij} \leq 1$. The saliency weights $\mathbf{w}^*$ and additive constant c were non-negligible in magnitude ($> .2$) but otherwise random. Two feature structures were used for each of the four similarity models, one with four features and the other with six. The representations incorporate features with variable patterns of encompassment and overlap, as shown in Figure 4.1. None of the representations is degenerate, since the complexity matrix $\mathbf{G}$ for each representation has full rank. However, the value of $\ln|\mathbf{G}|$ varies slightly across the four similarity models, as would be expected given their different functional forms. It is important to recognise that this presents no difficulties

for this study, as the purpose is *not* to compare similarity models. The aim is to verify that the algorithms serve their intended purpose of extracting appropriate representations, no more.

Each of the eight similarity matrices was analysed by using the appropriate algorithm to extract a representation (with 10 restarts). Since the intent was not to compare similarity models, an algorithm that uses one model was never to fit to data generated under another. Each procedure was repeated 10 times.

4.5.2 Results

Figures 4.2 through 4.5 display the VAF and GCC values for solutions obtained by the additive clustering, distinctive features clustering, TCM ($\rho = 0.5$), and MCM algorithms respectively. The dotted lines represent error bars, showing one standard error above and below the mean. No dotted lines are visible when all 10 runs yielded the same result. Note that, though the error bars are symmetric, the samples in question are not. When the runs yielded different results, the sample of GCC or VAF values typically consisted of several instances of the optimal representation, and a few suboptimal results. Nevertheless, inspection of the figures reveals that the GCC selected the representation with the appropriate number of features in 7 out of 8 cases. The exception is the six-feature TCM data, where a four-feature representation is preferred. Table 4.1 displays the number of "hits" for each data set, defined as the number of occasions when the true representation was recovered (if not preferred).

The apparent "failure" of the GCC to prefer a six-feature representation for the six-feature TCM data is easily explained. The five-feature representation (GCC=34.8) recovered 9 times out of 10 from the six cluster data, as well as the four cluster representation (GCC=35.1) recovered on all 10 occasions both had slightly lower GCC-values than the "true" representation (GCC=36.6). These representations are both identical to

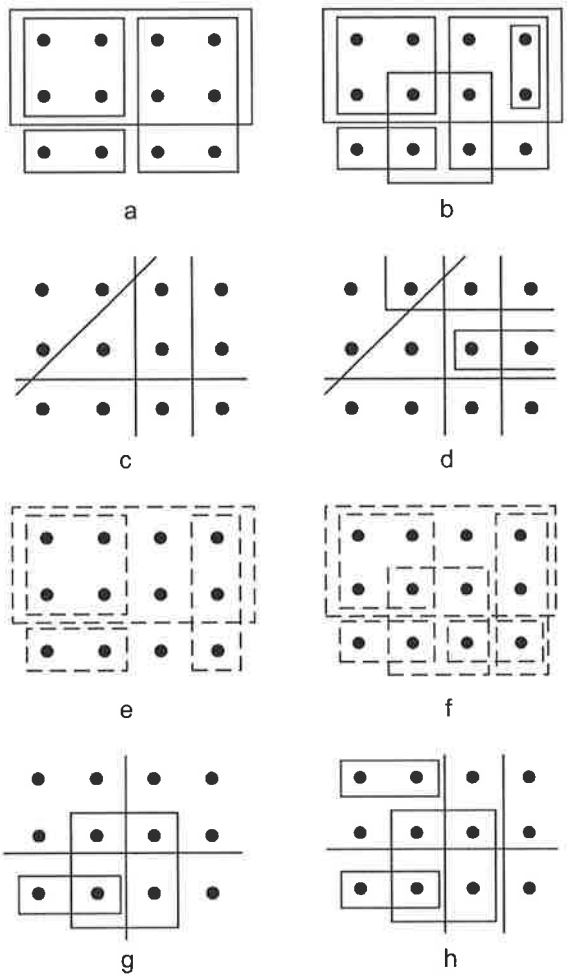

Figure 4.1: Features used to generate data for the Monte Carlo evaluation, using the common features model (a and b), distinctive features model (c and d), TCM ($\rho = .5$, e and f), and MCM (g and h). Common features are depicted as rectangles, whereas distinctive features are displayed by lines dividing the stimulus set. Since the ρ-value adopted for the TCM was 0.5, features in those representations were neither common nor distinctive, and are drawn with dashed lines.

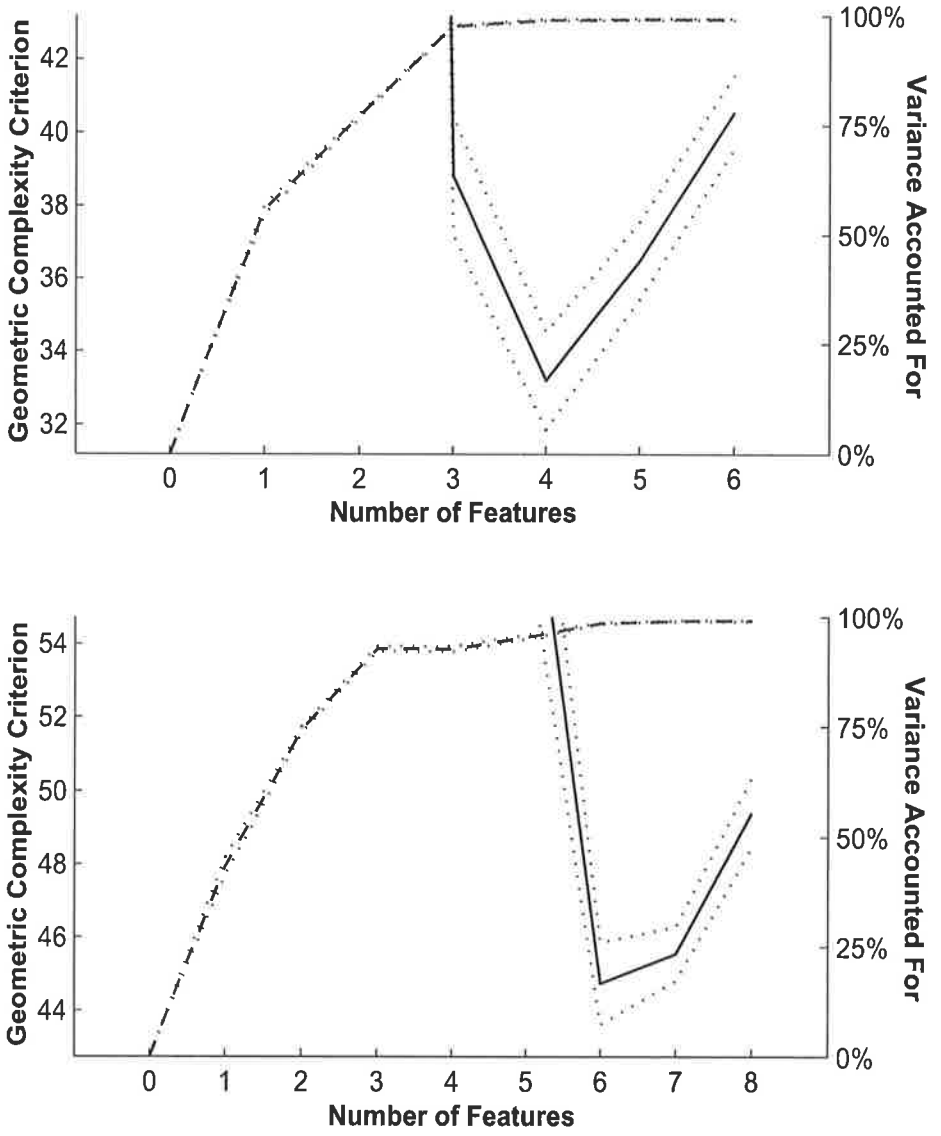

Figure 4.2: GCC and VAF values for representations derived from the four-feature (top) and six-feature (bottom) common features model data. Dotted lines are drawn one standard deviation above and below the mean.

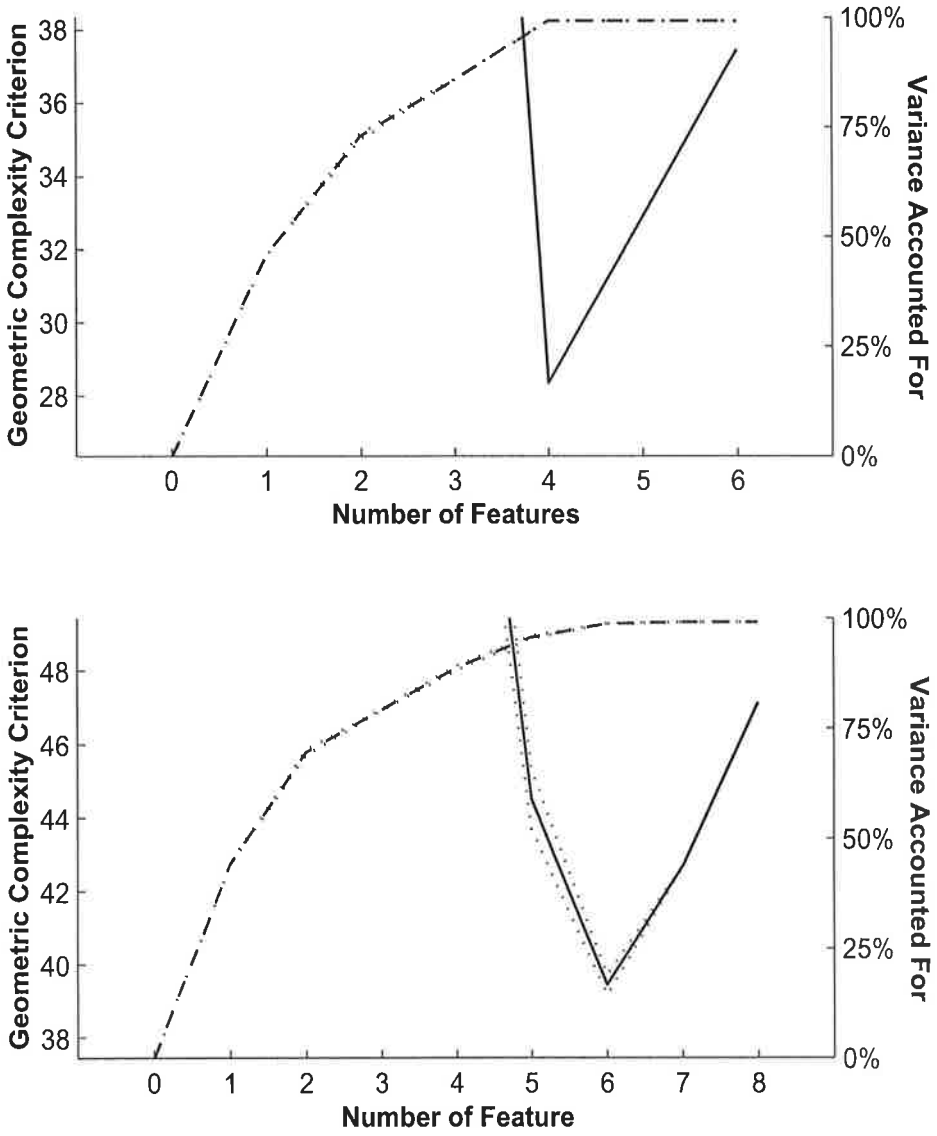

Figure 4.3: GCC and VAF values for representations derived from the four-feature (top) and six-feature (bottom) distinctive features model data. Dotted lines are drawn one standard deviation above and below the mean.

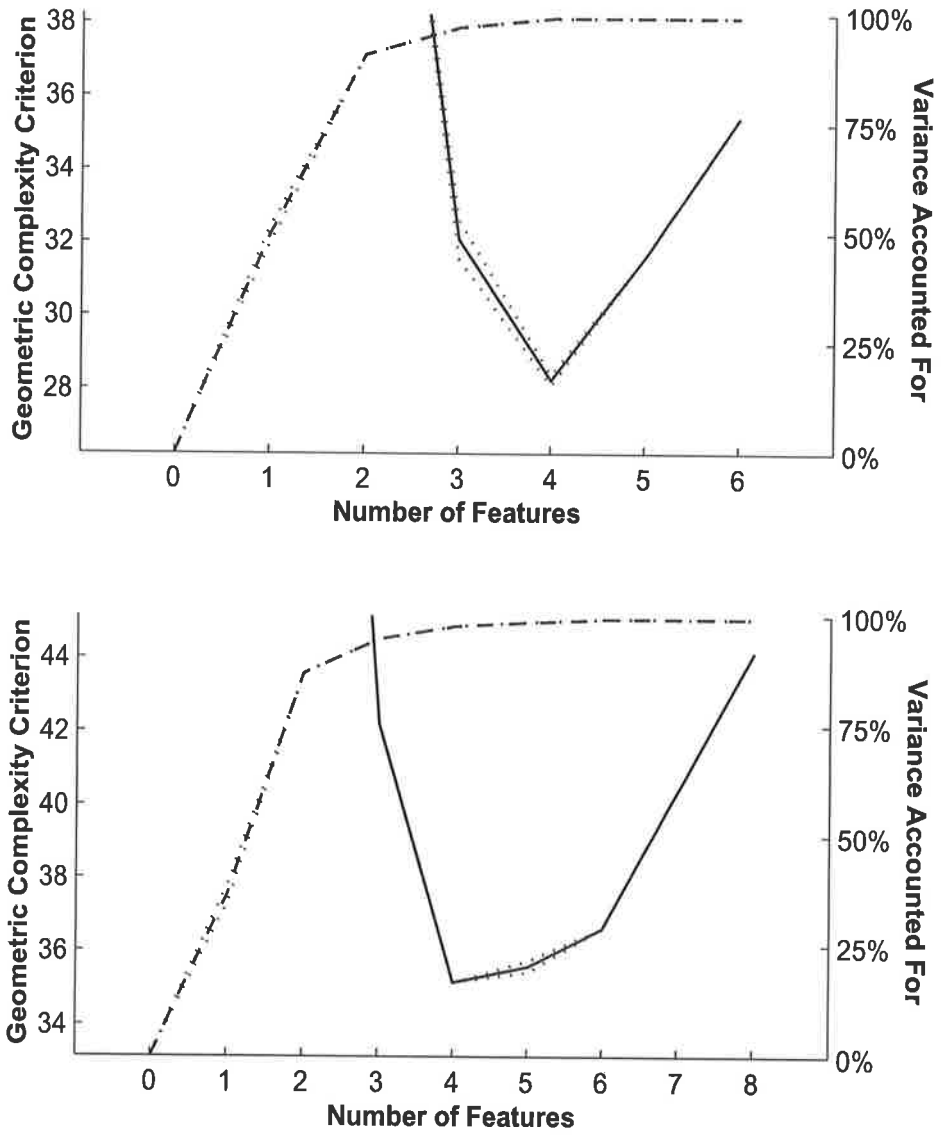

Figure 4.4: GCC and VAF values for representations derived from the four-feature (top) and six-feature (bottom) TCM data. Dotted lines are drawn one standard deviation above and below the mean.

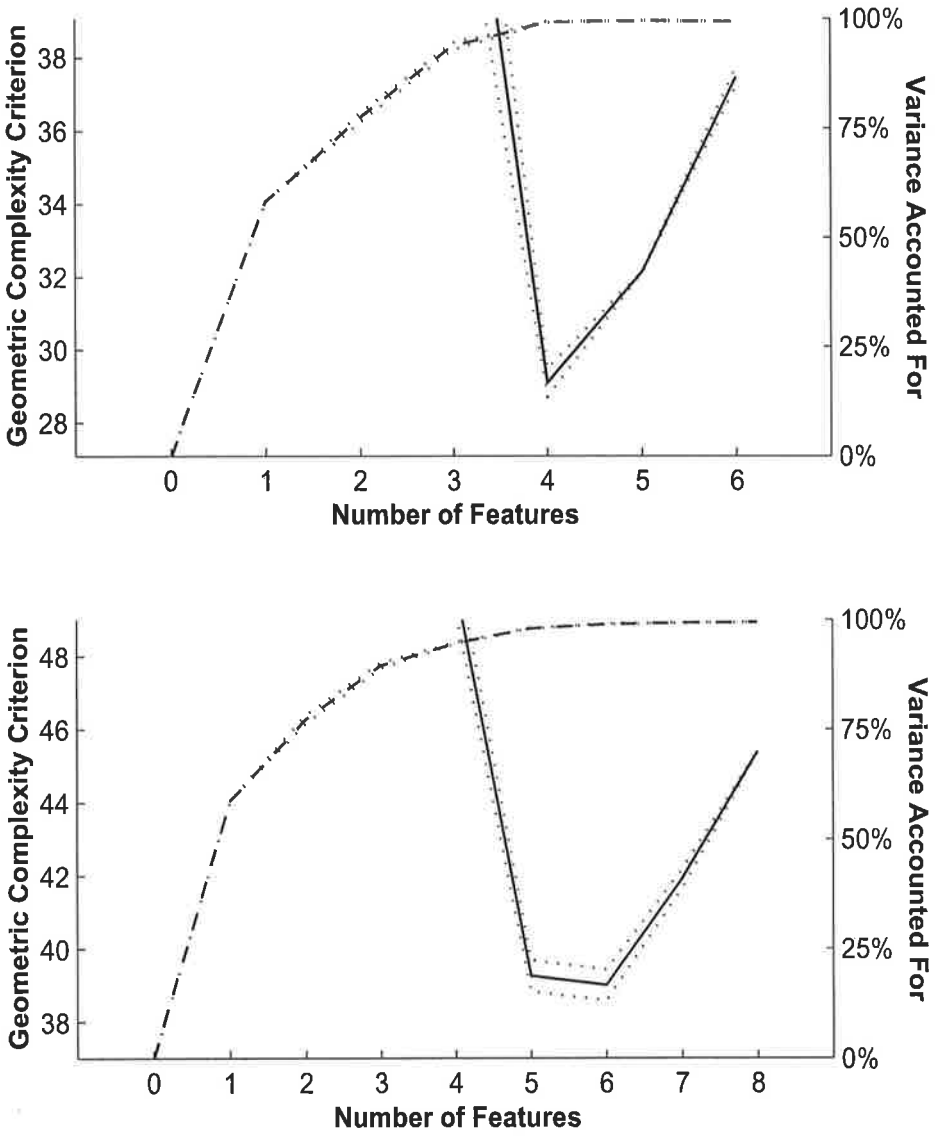

Figure 4.5: GCC and VAF values for representations derived from the four-feature (top) and six-feature (bottom) MCM data. Dotted lines are drawn one standard deviation above and below the mean.

Table 4.1: Number of times out of 10 that the algorithms recovered the "true" representation.

	Four Features	Six Features
Common Features Model	9	6
Distinctive Features Model	10	8
TCM	9	10
MCM	9	7

the true one except for the omission of one or both of the two smallest clusters (the ones containing only two stimuli each). It may be that these features do not make enough of a unique contribution to the data to warrant modelling. However, the difference in GCC between these representations is less than two, which is classified as "weak" evidence in Table 3.1 (or, "not worth more than a bare mention" to use Jeffreys' 1961 terminology). In short, this data set does not discriminate between these three representations.

4.5.3 Discussion

Not only do the algorithms recover the right number of clusters in most cases, but the correct feature structure is the one generally returned. The fact that the lowest-GCC representation for the six-cluster data for the TCM omitted one of features is not a cause for concern. As previously noted, the difference in GCC is very slight. Furthermore, there is an argument that if a feature does not make a substantial contribution to the model, then it ought not be included, and it may be that the omitted feature(s) did not make a sufficient contribution (see Kass & Raftery, 1995 for a similar argument regarding model recovery). In any case, this study indicates that the algorithms are capable of recovering known feature structures from noisy data. This does not imply that they will always recover the right representation from real data: empirical data

tends to present a much more difficult recovery problem. Nor does the study provide any support for the psychological models presented earlier. It does, however, demonstrate that the tools meet a minimum necessary level of effectiveness to justify their use in the following sections.

4.6 Representations of Kinship Terms

The GCC derivations and algorithm validations in the previous sections provide the tools required to evaluate the four featural similarity models. In this section each of these four models are used to extract representations from Rosenberg and Kim's (1975, see Section 2.1) data on the similarity of English kinship terms. Of the 15 terms included in the data set, 14 denote specific, gendered relationships (the exception being the word 'cousin'). In order to examine the manner in which each of the four similarity models represents the concept of gender, 'cousin' is excluded.

The first model applied to the data was the common features model. The GCC-driven algorithms extracted the ten-feature representation displayed in Table 4.2. The representation is by and large a sensible one, containing simple, interpretable features such as sibling (brother, sister) and parent (father, mother), and accounting for 93.1% of the variance. The only substantial shortcoming in this representation is that it requires two features to represent gender: the sixth feature contains the female terms and the seventh feature contains the male terms. The fact that these two features are the complement of one another, and have virtually identical saliency weights suggests that a more compact representation of gender is possible. Indeed, the eight-feature distinctive features representation shown in Table 4.3 contains a single feature that distinguishes the male terms from the female terms. This single distinctive feature accounts for as much variance as the two common features, and is just as interpretable. However, although this representation accounts for 94.7% of the variance and has a lower GCC than the

Table 4.2: Common features representation of the kinship data.

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
Additive Constant	0.226
Variance Accounted For	93.1%
Geometric Complexity Criterion	59.6

common features representation, it fails to capture the simple, interpretable features that appear in Table 4.2.

If the common features model cannot represent gender by a single feature, and the distinctive features model does not capture the concepts of sibling or parent, then some combination of common and distinctive features is required. The preferred TCM representation contains seven features with $\rho = 0.2$, and is shown in Table 4.4. Inspection of this representation reveals substantial problems with interpretation. The simple common features such as sibling do not emerge, and gender still requires two features. The common features do not emerge because of the low ρ, but since $\rho > 0$ a single feature cannot account for gender. If there were a single feature containing all the female terms, then the distinctive features component would make the male and female terms less

Table 4.3: Distinctive features representation of the kinship data.

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
Additive Constant	0.902
Variance Accounted For	94.7%
Geometric Complexity Criterion	50.5

similar to each other, but the common features component would make the female terms more similar to one another without having a corresponding effect on the male terms. This asymmetry means that a second feature is required, containing all the male terms. Therefore, although the TCM allows a compromise between common features concerns and distinctive features concerns, the trade-off may not be to the advantage of either.

The ten-feature MCM representation shown in Table 4.5 adopts an interpretable compromise between common and distinctive features. The four distinctive features distinguish the male terms from the female terms, the once removed terms (aunt, nephew, niece, uncle) from those not once removed, the extreme generations (granddaughter, grandfather, grandmother, grandson) from middle generations, and the nuclear family (brother, daughter, father, mother, sister, son) from the extended family. The six common features represents simple, interpretable concepts such as sibling, parent, grandparent

Table 4.4: TCM representation of the kinship data with $\rho = .2$.

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
Additive Constant	0.679
Variance Accounted For	91.4%
Geometric Complexity Criterion	50.4

and grandchild. Importantly, these concepts are appropriately declared to be 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.

The four analyses of the kinship data raise three interesting observations. Firstly, the similarity data appear to reflect the operation of common features and distinctive features. Nevertheless, the TCM's reliance on the decision variable ρ prevents it from accommodating this. However, by declaring features to be either common features or distinctive features, the MCM is capable of doing so. Secondly, the common features and MCM representations suggest that there are many commonalities in the domain, yet the balance between common and distinctive features in the TCM is highly biased toward distinctive features. It may be that the common features component introduces

Table 4.5: MCM representation of the kinship data.

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
Additive Constant	0.660
Variance Accounted For	93.5%
Geometric Complexity Criterion	56.1

more complexity than the distinctive features component, and therefore the GCC favours a lower ρ. Evidence in favour of this suggestion is provided by observing that the distinctive features component is constrained by the triangle inequality, but the common features component is not. Furthermore, Navarro and Lee (2001) have demonstrated that common features representations can accommodate distinctive features data substantially better than vice versa. Thirdly, the TCM and distinctive features representations have the lowest GCC values (50.5 and 50.5 respectively), followed by the MCM (56.1) and then common features (59.6) representations. Using the standards shown in Table 3.1, there is no evidence to choose between the TCM and distinctive features representations, though there is "positive" evidence to suggest that the MCM performs less well than these two, and positive evidence that the common features representation is worse again.

Despite these differences, the MCM representation is arguably the best: although the GCC provides a well-founded trade-off between fit and complexity, it does not account for interpretability. The arguments made in this section regarding the interpretability of the four models favour the MCM over the other three, and the difference in GCC does not seem sufficiently extreme to justify choosing an uninterpretable representation over a meaningful one.

4.7 Monte Carlo Study II: Complexity

The previous section indicated that common features representations may be more complex than distinctive features representations. If so, it is worth examining the manner in which the TCM and MCM interpolate between the two. This section presents a Monte Carlo study investigating the complexity of the four featural similarity models.

4.7.1 Method

The study involved 100 representations of 15 stimuli, each containing 10 features. The features were assigned to stimuli at random, subject to the constraint that the representations avoid degeneracy ($\mathbf{F}$ had full rank in all cases). In this way, the feature structures were guaranteed not to be biased towards unusually simple or unusually complex structures. The complexity matrix $\mathbf{G}$ was then calculated for every feature structure, using a range of featural similarity models. For the TCM, ρ was set to 0, 0.1, 0.2, ..., 1, thus moving from a distinctive features model to a common features model. A corresponding transition was achieved for the MCM by randomly declaring 0, 1, 2, ..., 10 of the features to be common.

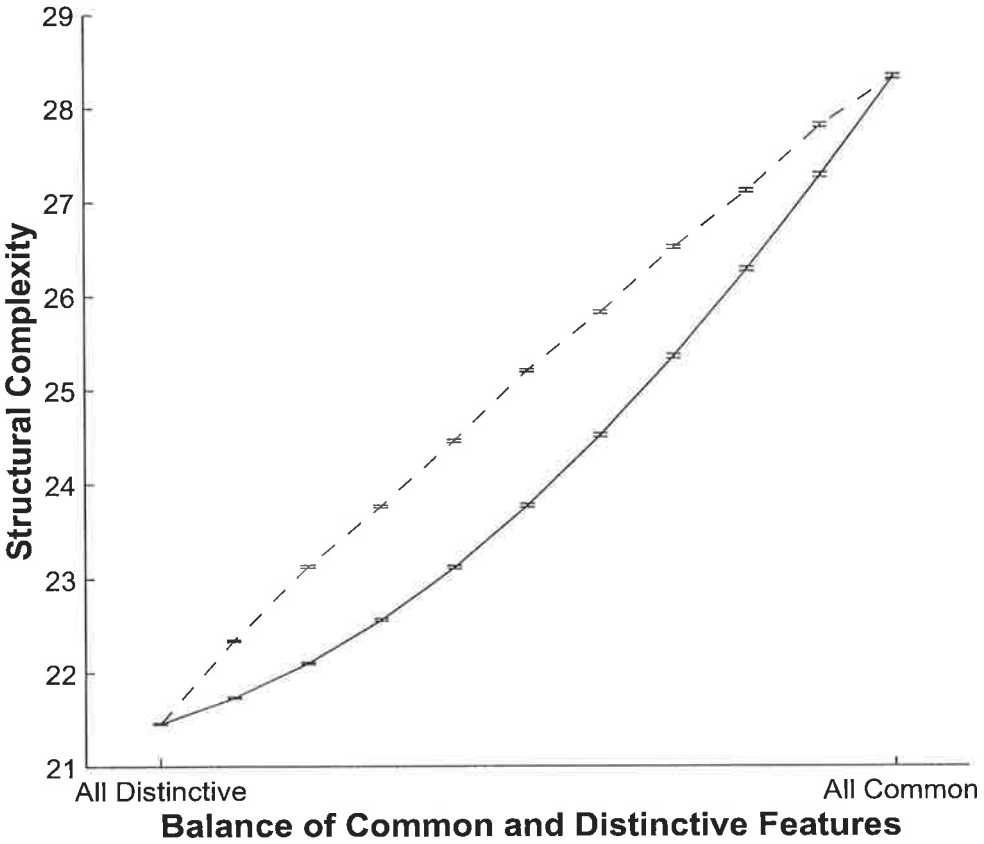

Figure 4.6: Average structural complexity ($\ln |\mathbf{G}|$) for featural representations employing a balance between common and distinctive features. The leftmost point corresponds to a distinctive features model, and the rightmost to a common features model. The solid line plots the complexity of the representations under the TCM (where "balance" equals ρ), and the dashed line represents the MCM (where "balance" refers to the proportion of common features). Error bars one standard error above and below the mean are shown.

4.7.2 Results

Figure 4.6 plots the value of the structural complexity component of the GCC, $\ln|\mathbf{G}|$, as a function of the balance between common and distinctive features employed by the similarity model. The distinctive features model (left side) is clearly the simplest, and the common features model (right side) the most complex. The solid line plots the change in complexity as the TCM moves from the distinctive features model ($\rho = 0$) to the common features model ($\rho = 1$), whereas the dashed line shows the same transition for the MCM as the proportion of common features in the model goes from 0 to 1. Apart from the endpoints, where the TCM and MCM both reduce to the common and distinctive features models, the TCM is always the simpler of the two.

4.7.3 Discussion

The speculation in the previous section appears to be borne out: the common features model is more complex than the distinctive features model, and the MCM is more complex than the TCM. Specifically, the complexity of MCM representations appears to increase linearly with the proportion of common features, whereas the complexity function for the TCM is clearly convex. In general, it is likely to be the case that any plausible featural model can provide a good account of most data sets in a parsimonious manner. However, the results here suggest that representations derived using the TCM will very likely be less complex, and if so, they will perform better on measures such as the GCC. Consequently, the onus is on the MCM to provide a more interpretable account of similarity judgements. The analyses of the kinship data suggest this to be the case, but further investigations are certainly warranted.

4.8 Experiment I: Faces

The first experiment involved an artificial domain consisting of a several cartoon faces. The faces were constructed in such a way as to present participants with a variety of features with different real-world base rates and social connotations. For these reasons, although the stimulus domain is simply designed, it need not be trivial.

4.8.1 Method

Participants

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

Stimulus Domain

A set of 10 cartoon faces were designed, varying in hairstyle (male or female), hair colour (brown, black, burgundy, grey or bright blue), glasses shape (square or round) and glasses colour (dark blue or pink). These faces are shown in Figure 4.7, and described briefly in Table 4.6.

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).

4.8.2 Results

The similarities shown in Table 4.7 were calculated by averaging across participants and normalising the data to lie between 0 and 1. Following Lee (2001a), a precision estimate $\hat{\sigma}_{ij}$ was calculated for each stimulus pair, by taking the standard deviation of

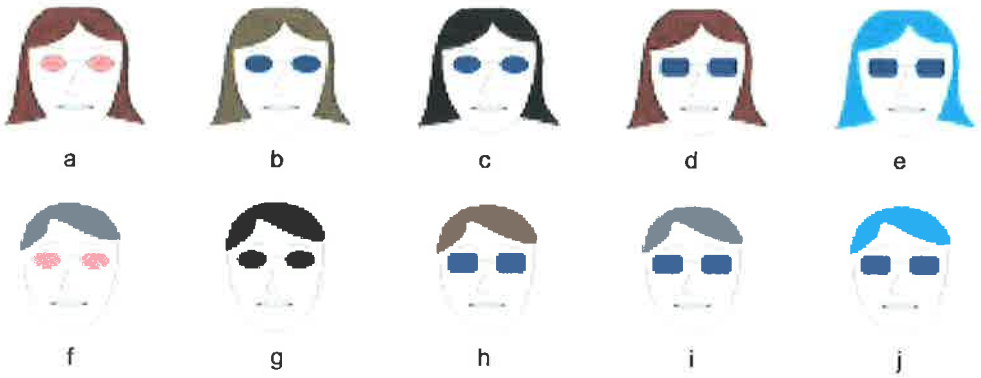

Figure 4.7: The ten cartoon faces used as stimuli in Experiment I. Faces vary in hairstyle and sunglasses only.

Table 4.6: Verbal description of the faces 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

the participants' ratings for that pair. These precision estimates are shown in Table 4.8, and range from 0.11 to 0.22 with standard deviation 0.03. The consistency of these estimates allows the median value to be used as the overall precision estimate for the data, yielding $\hat{\sigma} = 0.16$.

Each of the four similarity models was used to extract representations from the data. The stochastic hillclimbing algorithms were applied five times (with 10 restarts each) for all four similarity models, and representations were evaluated using the GCC. All VAF or GCC plots shown in this section display the best results from the five runs. Figure 4.8 displays the Variance Accounted For by the common features representations, as well as the trade-off between data-fit and model complexity as measured using the GCC. The GCC strictly preferred a representation containing two features: however, because the GCC deteriorates by only 2.2 (see Table 3.1) when a third feature is added, and that the three feature representation allows a far richer interpretation of the data, the subsequent discussion considers this three feature model. The GCC for the distinctive features model (see Figure 4.9) also preferred a two feature model. Once more the rise in GCC is small (1.6) when a third feature is included, and for similar reasons of interpretability this representation is used shortly. The Variance Accounted For by TCM representations is displayed in Figure 4.10, and the GCC in Figure 4.11. It is clear that the GCC favours $\rho = 0$, making the TCM equivalent to the distinctive feature model. Accordingly, the $\rho = 0$ TCM representations are identical to the distinctive features representations, and the three feature model is subsequently discussed. Finally, the MCM shows the same pattern (see Figure 4.12), with the deterioration being 1.8 when a third feature is included.

The common features representation displayed in Figure 4.13 contains features consisting of the male faces and the female faces. The third feature captures the 'ordinary' faces, as it consists of those faces with more conservative hair colours and sunglasses.

Table 4.7: Similarities between all pairs of faces.

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

Table 4.8: Precision estimates for the similarities between all pairs of faces.

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
a	-									
b	0.15	-								
c	0.19	0.16	-							
d	0.18	0.17	0.15	-						
e	0.12	0.18	0.20	0.16	-					
f	0.18	0.17	0.17	0.12	0.12	-				
g	0.13	0.13	0.17	0.19	0.19	0.17	-			
h	0.20	0.16	0.18	0.16	0.16	0.17	0.22	-		
i	0.15	0.17	0.16	0.20	0.14	0.14	0.20	0.14	-	
j	0.11	0.14	0.12	0.12	0.17	0.16	0.14	0.14	0.12	-

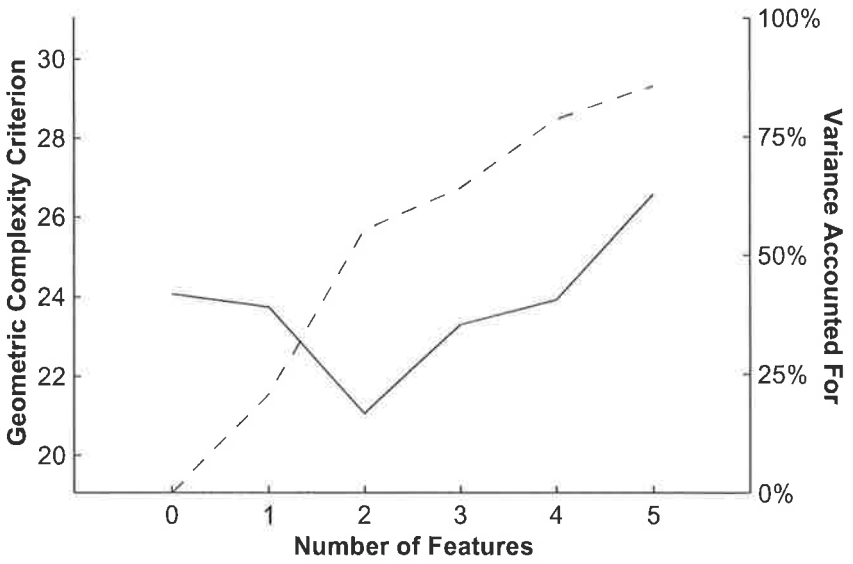

Figure 4.8: The GCC and VAF values for common features representations of the faces data. The GCC strictly prefers two features, though the three feature representation is subsequently discussed.

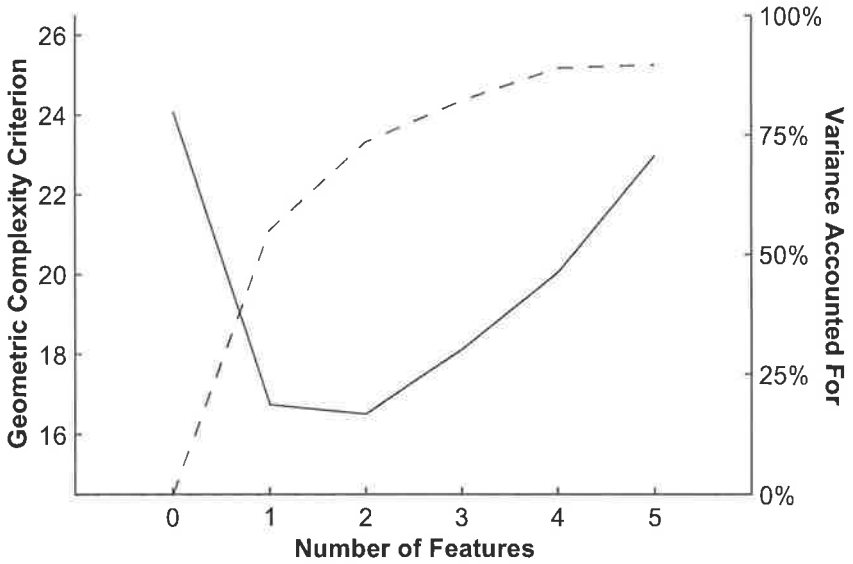

Figure 4.9: The GCC and VAF values for distinctive features representations of the faces data. The GCC strictly prefers two features, though the three feature representation is subsequently discussed.

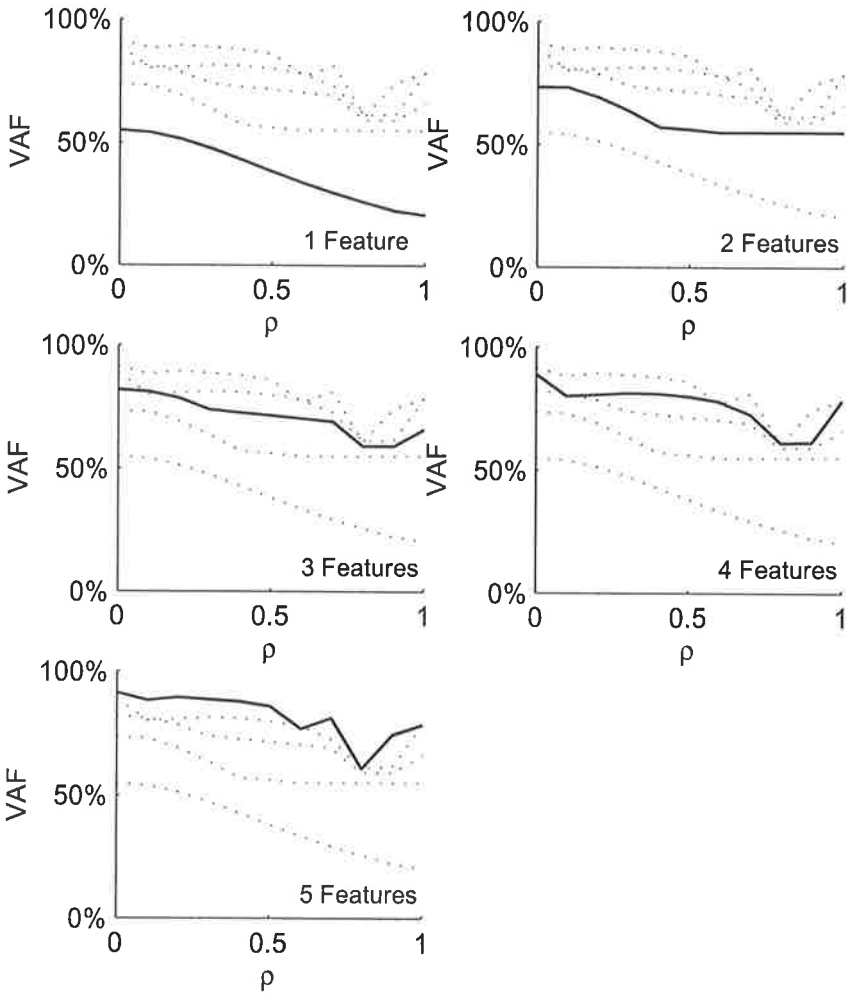

Figure 4.10: The VAF values for TCM representations of the faces data. Each panel contains the same five plots of VAF values as a function of ρ: each plot corresponds to representations with a particular number of features. Each panel highlights one of the plots: the highlighted plot is identified by the number of features indicated by the text in the bottom right corner of each panel.

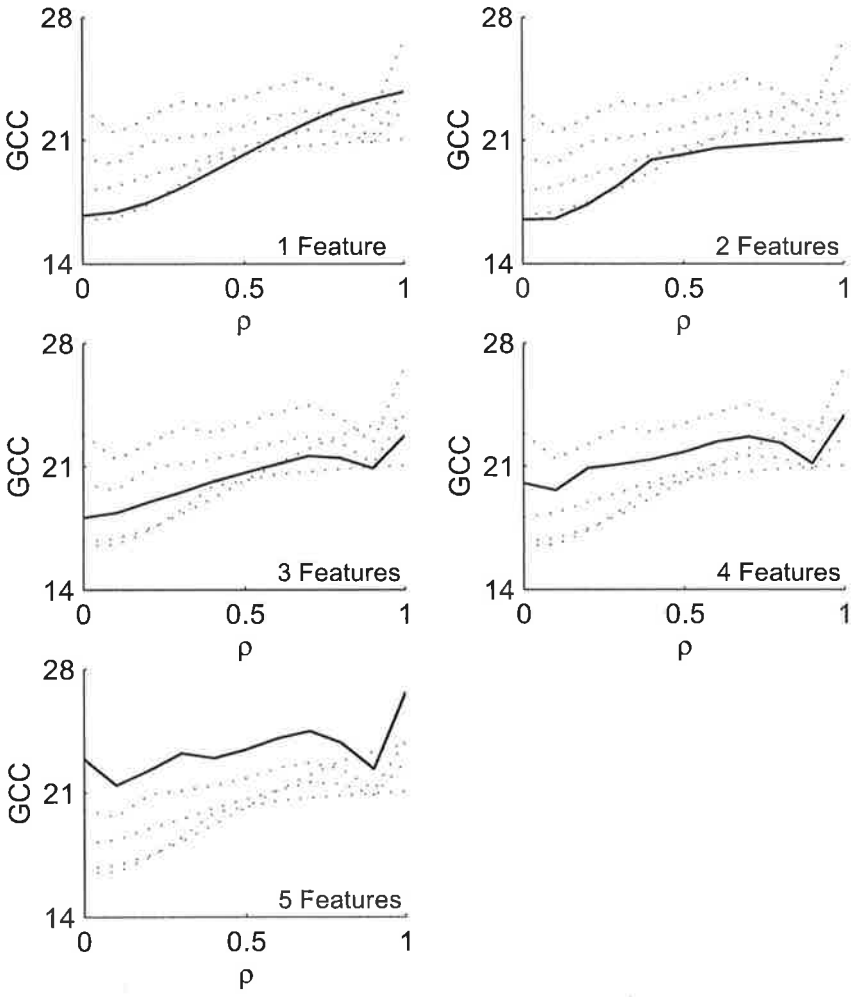

Figure 4.11: The GCC values for TCM representations of the faces data. Each panel contains the same five plots of GCC values as a function of ρ: each plot corresponds to representations with a particular number of features. Each panel highlights one of the plots: the highlighted plot is identified by the number of features indicated by the text in the bottom right corner of each panel. The GCC strictly prefers two features and $\rho = 0$, though the three feature representation is subsequently discussed.

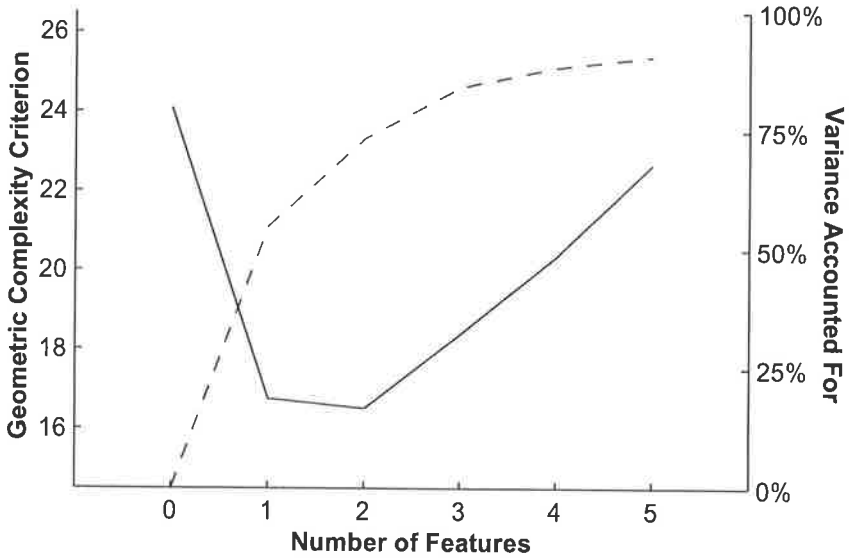

Figure 4.12: The GCC and VAF values for MCM representations of the faces data. The GCC strictly prefers two features, though the three feature representation is subsequently discussed.

The distinctive features and TCM representation displayed in Figure 4.14 contains a gender distinction, a distinction between the square and round sunglasses, and a distinction between pink and blue sunglasses. Finally, the MCM representation displayed in Figure 4.15 captures the distinctive features corresponding to gender and glasses shape, and the common feature consisting of the 'ordinary' faces.

4.8.3 Discussion

Despite the simplicity of this experiment, the data set turns out to be useful in evaluating the four featural models under consideration. The gender and glasses-shape aspects of the domain make sense only as distinctive features, and a common features representation can capture the regularities only by having two common features with the same weights. This data set also demonstrates why allocating two common features to model a distinctive feature is a poor representational strategy. The use of a principled model selection criterion demonstrates that the complexity of fitting models is strongly constrained by

Figure 4.13: Additive clustering representation of the faces data, explaining 63.9% of the variance (GCC=23.3).

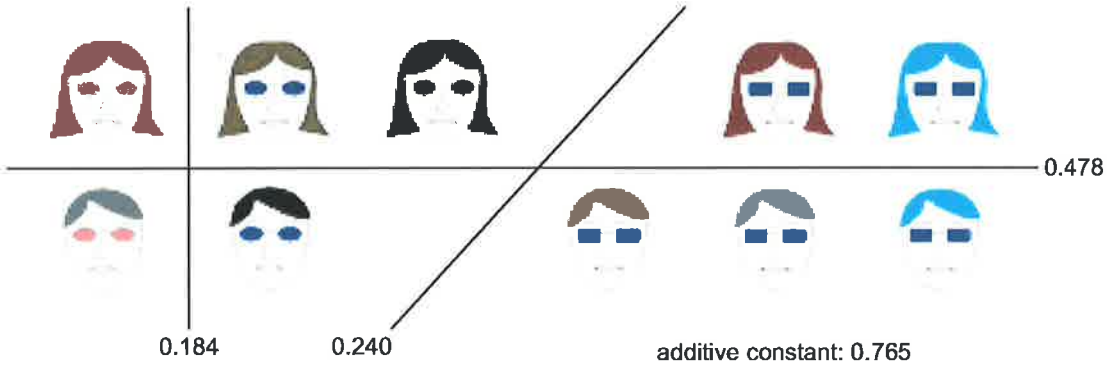

Figure 4.14: Distinctive features representation of the faces data, explaining 82.1% of the variance (GCC=18.1). This representation was also found using the TCM.

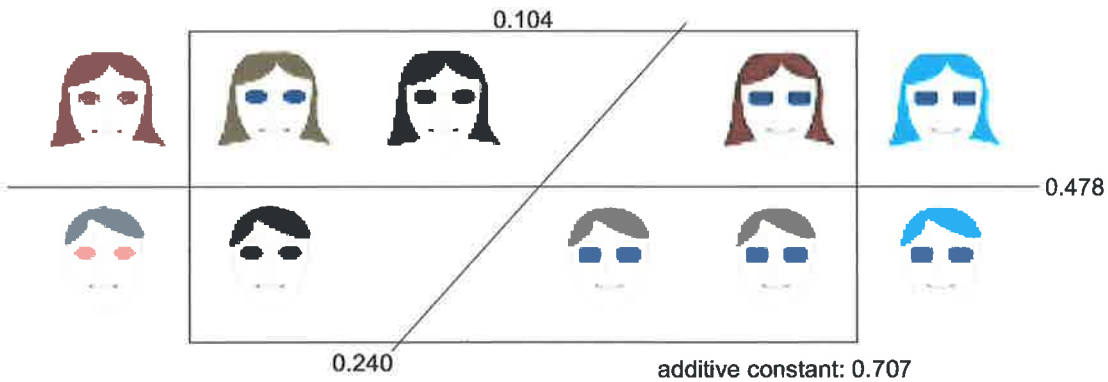

Figure 4.15: MCM representation of the faces data, explaining 84.2% of the variance (GCC=18.4).

the precision of the data. In this domain, it appears that only two or three features are justified by the data. Therefore, the common features model, which requires four features to account for gender and glasses-shape, performs poorly.

The distinctive features model, the TCM, and the MCM all account for the domain well, capturing the two most prominent regularities (gender and glasses-shape) with only two features. Since the VAF and GCC do not distinguish between these representations, the models can be judged only by the interpretability of the third feature. However, the glasses-colour distinction made by the distinctive features and TCM representations is a reasonable feature, as is the 'ordinary faces' common feature found in the MCM representation. It is perhaps sufficient to note that only the MCM is capable of simultaneously accommodating the common feature 'ordinary' and the two distinctive features. This is important because, while two ordinary things are likely to be similar to each other, two unusual things need not be. For example, people may judge two 'ordinary' Western faces (e.g., Al Gore and Tom Cruise) to be more similar to one another than two 'unusual' Western faces (e.g., Sid Vicious and John Malkovich). Correspondingly, the concept of an ordinary face is necessarily a common feature rather than a distinctive one. Capturing commonalities as well as distinctions is typical behaviour for the MCM: the TCM cannot do this under any (hyper-)parameterisation.

4.9 Experiment II: Countries

Although artificial stimuli may be useful in demonstrating theoretical points, it is often more informative to examine the behaviour of models when applied to more natural domains. Therefore, the second experimental test of the four similarity models employed natural stimuli. Inspired by Tversky and Gati's (1978) work on common and distinctive features, the domain consisted of a set of 16 nations. Tversky and Gati presented their participants with a pair of well-known (prominent) nations and a pair of less well-known

(non-prominent) nations. In one condition participants were asked to choose which of the two pairs of nation were more similar, and in the other condition, they were asked to select the more dissimilar pair. They recorded the proportion of participants choosing the prominent pair as more similar, as well as the proportion choosing the prominent pair as more dissimilar, and found that the sum of the two was greater than 1. From this they concluded that the balance between common and distinctive features shifts from common to distinctive as the task is switched from a similarity to dissimilarity.

This suggested a simple experimental design, in which both similarity and dissimilarity judgements were collected for a set of stimuli 16 nations identified by name. The nature of the task, however, made it less than satisfactory to present people with a pair of countries and ask them to provide a rating of similarity. It seems likely that this would be ambiguous, in that the initial impulse of participants may be to ask, "Compared to what?". Even if the level of similarity (or dissimilarity) between a pair of nations is obvious to a participant, they are unlikely to bring to this task a preexisting numerical scale of nation-similarity upon which to rate it. An alternative approach is to provide the participants with a context in which to make judgements, and so avoid this difficulty[3]. Consequently, the task involved presenting people with a list of four countries, and asking them to select from that list the pair of nations most similar to (or most different from) one another. For instance, if presented with "Italy, Jamaica, Nigeria and

[3]As an aside, one might ask whether this difficulty arises with regard to *any* similarity task involving a rating scale. As Goldstone et al. (1997) have argued, a similarity judgement takes the form "A is like B (in some respect, R)", where R is determined by the context. A rating scale provides very little context for the decision, so participants are required to discover (or provide) the context themselves, by finding a number of respects $\mathbf{R} = \{R_1, R_2, \ldots, R_k\}$ with which a judgement could be made. By asking people to assess similarity without providing a suitable context, the researcher is implicitly requiring them to average across a wide range of "potential contexts" in which the stimuli might be encountered. In a simple perceptual task such as the faces experiment (Section 4.8), this is unlikely to be difficult. Even for a domain such as O'Doherty and Lee's (2002; see Section 2.1) animals data, it may not be too strenuous to provide a sufficiently diverse set of contexts to make a reasonable decision. However, this may not be the case for the nations domain, so it seemed prudent to provide a more constrained "local context". Consequently, participants only needed to find a set of respects appropriate to the local context, and not to every possible context in which one might need to compare two nations.

Zimbabwe" as a list, a participant might select Zimbabwe and Nigeria as the two most similar nations.

The TCM assumes that the balance between common features similarity and distinctive features similarity is the same for every feature, although it may vary according to the nature of the task. In the (unlikely) event that dissimilarity judgements are straightforward reversals of similarity judgements, one would expect similar feature structures to emerge in the two conditions, but with different values for ρ. Alternatively, since the MCM declares features to be either common features or distinctive features, one would expect the saliency of common features to go down and the saliency of distinctive features to go up as the task shifts from similarity to dissimilarity. One aim of this study was to see if this phenomenon is observable in the derived representations. However, in a broader sense, the aim was to examine the structures that emerge from the two data sets under the different models.

4.9.1 Method

Participants

Participants in the study were 30 university students (11 male, 19 female) aged 17 to 49, with a median age of 24, who took part in the experiment for course credit. Sixteen of the participants provided similarity judgements and 14 provided dissimilarity judgements.

Stimulus Domain

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

Procedure

The participants were randomly assigned to one of two conditions. On any given trial a list of four countries was displayed (via computer) to the participant, who was asked to select the two countries most similar to or most different from each other, depending on the condition to which they were assigned. 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 similarity ratings were sensitive to all four presented stimuli, it was important to exhaust the set of 1820 quadruples exactly. To that end, the 1820 items were partitioned into 20 subsets of 91 quadruples in both conditions. Most participants provided responses to one such subset, though a few provided responses to multiple subsets. No participant provided both similarity and dissimilarity judgements. Since each quadruple involves the presentation of 6 of the 120 pairs of nations, each pair appeared a total of $\frac{1820 \times 6}{120} = 91$ times across the entire data set.

4.9.2 Results

Calculating Similarity and Dissimilarity

The natural measure of stimulus similarity for choice data is the probability with which two stimuli are chosen. If ϕ_{ij} represents the "true" similarity between the ith and jth stimulus which were chosen k times out of n (n being 91), then the similarity value indicated by the choice data is given by the expected value of ϕ_{ij}. That is,

$$\begin{aligned} s_{ij} &= E\left[\phi_{ij}|k,n\right] \\ &= \int_0^1 \phi_{ij} p(\phi_{ij}|k,n)\ d\phi_{ij}. \end{aligned}$$

Using Bayes theorem, the posterior probability $p(\phi_{ij}|k,n)$ is given by

$$p(\phi_{ij}|k,n) = \frac{p(k,n|\phi_{ij})p(\phi_{ij})}{p(k,n)} = \frac{p(k,n|\phi_{ij})p(\phi_{ij})}{\int_0^1 p(k,n|\phi_{ij})p(\phi_{ij})\ d\phi_{ij}}.$$

If a uniform prior distribution $p(\phi_{ij})$ is chosen for ϕ_{ij}, this reduces to

$$p(\phi_{ij}|k,n) = \frac{p(k,n|\phi_{ij})}{\int_0^1 p(k,n|\phi_{ij})\ d\phi_{ij}}. \tag{4.9}$$

Using a standard result in Bayesian statistics (Gelman, Carlin, Stern, & Rubin, 1995, p. 31), the expectation of this probability yields the similarity value $s_{ij} = \frac{k+1}{n+2}$. An equivalent argument yields the dissimilarity value $d_{ij} = \frac{k+1}{n+2}$ in the other condition. The similarity and dissimilarity values calculated according to this rule are displayed in Tables 4.9 and 4.10 respectively.

Estimating Precision

Estimating the precision of this data set requires a different approach to that used when analysing the faces data. The nature of this experiment required 1820 unique trials: ideally, each participant would have responded to every item, yielding a balanced similarity matrix for each participant, allowing an estimate to be made of the between-subject variability in responses. Since it was not feasible to have each participant provide 1820 judgements, this is not a viable approach. Nevertheless, it is possible to use the individual participants' data to make a best estimate of what the individual similarity matrices would have looked like, again using Bayesian posterior probability to calculate similarity. However, since the pairs appeared with different frequencies n for each participant, the confidence in these estimates is quite variable. It may sometimes be assumed that individual similarity estimates are arbitrarily precise (Lee, 2001a), but in this case it is inappropriate to do so. This is evident in Figure 4.16, which displays the posterior prob-

Table 4.9: Pairwise similarity estimates for the similarity-condition data set.

	Chi	Cub	Ger	Ind	Ira	Ita	Jam	Jap	Lib	Nig	Phi	Rus	Spa	USA	Vie	Zim
China	-															
Cuba	0.11	-														
Germany	0.06	0.04	-													
Indonesia	0.43	0.06	0.03	-												
Iraq	0.06	0.32	0.04	0.14	-											
Italy	0.02	0.09	0.70	0.02	0.03	-										
Jamaica	0.02	0.59	0.02	0.14	0.04	0.10	-									
Japan	0.69	0.01	0.26	0.35	0.03	0.06	0.03	-								
Libya	0.03	0.32	0.01	0.04	0.70	0.04	0.11	0.01	-							
Nigeria	0.01	0.12	0.01	0.04	0.20	0.03	0.31	0.01	0.45	-						
Philippines	0.42	0.12	0.01	0.87	0.09	0.02	0.17	0.31	0.05	0.04	-					
Russia	0.51	0.35	0.55	0.01	0.13	0.22	0.02	0.17	0.05	0.02	0.03	-				
Spain	0.02	0.37	0.58	0.03	0.04	0.90	0.20	0.04	0.04	0.03	0.04	0.15	-			
United States	0.30	0.11	0.42	0.03	0.06	0.20	0.12	0.46	0.02	0.04	0.01	0.43	0.20	-		
Vietnam	0.60	0.12	0.03	0.55	0.12	0.01	0.05	0.45	0.10	0.03	0.57	0.08	0.02	0.12	-	
Zimbabwe	0.01	0.08	0.01	0.11	0.15	0.02	0.29	0.01	0.31	0.83	0.08	0.01	0.02	0.01	0.03	-

Table 4.10: Pairwise dissimilarity estimates for the dissimilarity-condition data set.

	Chi	Cub	Ger	Ind	Ira	Ita	Jam	Jap	Lib	Nig	Phi	Rus	Spa	USA	Vie	Zim
China	-															
Cuba	0.12	-														
Germany	0.16	0.28	-													
Indonesia	0.03	0.14	0.33	-												
Iraq	0.12	0.06	0.32	0.04	-											
Italy	0.29	0.06	0.04	0.27	0.24	-										
Jamaica	0.46	0.05	0.45	0.12	0.40	0.18	-									
Japan	0.05	0.37	0.18	0.02	0.38	0.23	0.38	-								
Libya	0.10	0.04	0.43	0.06	0.01	0.13	0.25	0.40	-							
Nigeria	0.31	0.09	0.28	0.09	0.15	0.20	0.05	0.42	0.01	-						
Philippines	0.05	0.04	0.34	0.02	0.15	0.19	0.06	0.05	0.10	0.09	-					
Russia	0.02	0.03	0.02	0.20	0.10	0.06	0.49	0.10	0.20	0.20	0.28	-				
Spain	0.25	0.02	0.02	0.17	0.27	0.01	0.09	0.23	0.22	0.14	0.13	0.11	-			
United States	0.14	0.12	0.08	0.19	0.34	0.10	0.15	0.04	0.38	0.46	0.09	0.09	0.09	-		
Vietnam	0.04	0.12	0.34	0.01	0.08	0.33	0.12	0.04	0.11	0.18	0.02	0.12	0.30	0.33	-	
Zimbabwe	0.29	0.12	0.39	0.09	0.17	0.24	0.05	0.44	0.06	0.01	0.09	0.41	0.22	0.48	0.12	-

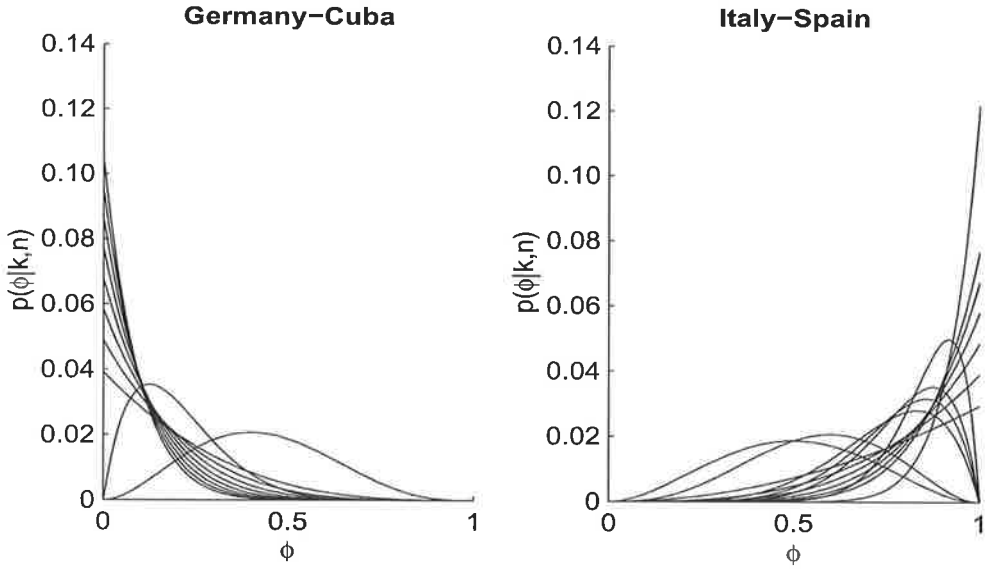

Figure 4.16: Probability distributions for ϕ for all participants for the Germany-Cuba and Italy-Spain response options in the similarity condition. These two pairs have similarity values of 0.04 and 0.90 respectively, and precision estimates of 0.19 and 0.21 respectively.

ability distributions for all participants for the Germany-Cuba and Italy-Spain response options in the similarity condition. The variation in the individual density functions highlights the importance of distributional information for these data. It is worth noting that the disagreement is in part an artifact of the experimental design: that is, the underlying variation in n. Nevertheless, experimental design *is* a source of potential imprecision, and must be taken into account when estimating precision.

The appropriate approach is to use the choice data to find, not just the expected probability $E[\phi_{ij}|k,n]$ with which each participant chose the ith and jth stimuli, but the full probability distribution for the response probability ϕ_{ij}. A given participant may have been presented with n trials that contained a particular pair of nations, and chosen that pair k times, yielding the posterior probability distribution for ϕ_{ij} described by Eq. 4.9. A simple way to estimate precision is to calculate the average squared deviation from s_{ij} of scores drawn from the set of all individual distributions. This

can be done numerically by drawing a sample from each individual distribution, and calculating

$$\hat{\sigma}_{ij} = \sqrt{\frac{\sum_{r=1}^{N} (x_r - \hat{s}_{ij})^2}{N-1}},$$

where x_r is the rth observation and N is the total number of observations overall. The advantage of this measure is that since $\hat{\sigma}_{ij}$ is a standard deviation estimate[4] of the variation around $\hat{s}_{ij}$, it preserves the underlying probabilistic similarity model (Tenenbaum, 1996), and accounts for the noise arising from differences between participants as well as the variability in n due to the experimental design. It is reassuring to observe that the precision estimates for the similarity data (Table 4.11) and the dissimilarity data (Table 4.12) do not display much variability, suggesting that the assumption of common variance is not violated. Given this, the median of these estimates is taken to be the overall precision estimate, yielding $\hat{\sigma} = 0.22$ for the similarity condition and $\hat{\sigma} = 0.24$ for the dissimilarity condition.

Similarity Condition Representations

Featural stimulus representations were extracted from the similarity-condition countries data for each of the four similarity models. The stochastic hillclimbing algorithms were applied five times (with 10 restarts each) for all four similarity models, and representations were evaluated using the GCC. All VAF or GCC plots shown in this section display the best results from the five runs. For the common features model, Figure 4.17 displays VAF and GCC values as a function of the number of features. Keeping in mind the standards of evidence advocated in Table 3.1, this figure suggests that some-

[4]Note that normalising by $N-1$ gives an estimate of the population standard deviation, whereas normalising by N gives the sample standard deviation. Ordinarily, a Bayesian approach does not use population estimators in the same way that a frequentist approach does, relying instead on the properties of the observed data. However, in this case, the x values are *not* the empirical data, but are (in theory) drawn from the same distribution. Thus the population estimate is applied, even though it is a frequentist measure rather than a Bayesian one. In practice, the point is largely irrelevant, since N is large.

Table 4.11: Precision values for each pairwise comparison in the similarity-condition data set, estimated using posterior probability distributions for each participant.

	Chi	Cub	Ger	Ind	Ira	Ita	Jam	Jap	Lib	Nig	Phi	Rus	Spa	USA	Vie	Zim
China	-															
Cuba	0.19	-														
Germany	0.19	0.19	-													
Indonesia	0.23	0.20	0.25	-												
Iraq	0.21	0.26	0.22	0.20	-											
Italy	0.20	0.22	0.26	0.23	0.26	-										
Jamaica	0.17	0.26	0.21	0.27	0.22	0.20	-									
Japan	0.27	0.21	0.24	0.28	0.21	0.23	0.22	-								
Libya	0.21	0.30	0.21	0.23	0.34	0.21	0.27	0.21	-							
Nigeria	0.19	0.24	0.21	0.26	0.27	0.21	0.28	0.20	0.24	-						
Philippines	0.23	0.23	0.19	0.23	0.23	0.20	0.21	0.23	0.22	0.24	-					
Russia	0.26	0.24	0.28	0.19	0.27	0.26	0.20	0.20	0.19	0.22	0.21	-				
Spain	0.20	0.29	0.26	0.26	0.20	0.21	0.24	0.20	0.22	0.20	0.19	0.20	-			
United States	0.24	0.19	0.24	0.19	0.21	0.24	0.22	0.28	0.21	0.20	0.20	0.27	0.23	-		
Vietnam	0.26	0.22	0.20	0.24	0.24	0.20	0.21	0.28	0.23	0.19	0.23	0.20	0.22	0.24	-	
Zimbabwe	0.22	0.26	0.19	0.24	0.29	0.21	0.27	0.20	0.30	0.28	0.19	0.22	0.19	0.20	0.18	-

Table 4.12: Precision values for each pairwise comparison in the dissimilarity-condition data set, estimated using posterior probability distributions for each participant.

	Chi	Cub	Ger	Ind	Ira	Ita	Jam	Jap	Lib	Nig	Phi	Rus	Spa	USA	Vie	Zim
China	-															
Cuba	0.25	-														
Germany	0.24	0.23	-													
Indonesia	0.24	0.20	0.26	-												
Iraq	0.26	0.24	0.23	0.20	-											
Italy	0.23	0.21	0.25	0.26	0.28	-										
Jamaica	0.21	0.24	0.29	0.21	0.23	0.24	-									
Japan	0.23	0.25	0.30	0.19	0.22	0.28	0.25	-								
Libya	0.22	0.26	0.25	0.23	0.21	0.20	0.22	0.26	-							
Nigeria	0.26	0.20	0.23	0.23	0.23	0.24	0.17	0.24	0.26	-						
Philippines	0.27	0.23	0.29	0.20	0.26	0.26	0.20	0.25	0.18	0.23	-					
Russia	0.23	0.20	0.23	0.25	0.24	0.20	0.26	0.24	0.20	0.23	0.26	-				
Spain	0.22	0.19	0.21	0.24	0.21	0.19	0.27	0.27	0.22	0.25	0.24	0.20	-			
United States	0.28	0.21	0.24	0.24	0.26	0.19	0.25	0.20	0.25	0.27	0.22	0.25	0.24	-		
Vietnam	0.23	0.18	0.27	0.20	0.25	0.25	0.23	0.20	0.21	0.26	0.22	0.24	0.21	0.25	-	
Zimbabwe	0.29	0.22	0.24	0.23	0.26	0.26	0.22	0.23	0.24	0.22	0.24	0.24	0.22	0.28	0.30	-

where between four and seven features are justified. In contrast, the GCC displays a clear preference for a five feature representation when the distinctive features model is applied (see Figure 4.18). When the TCM is applied, as shown in Figures 4.19 and 4.20, the GCC favours a four to six feature model with ρ between 0.5 and 0.7. Finally, though the GCC for the MCM reaches a minimum at four features (see Figure 4.21), a representation containing up to seven features is acceptable.

The representations displayed in Figures 4.23 through 4.24 are those with the most features within the acceptable range for each similarity model. In quantitative terms, the TCM representation ($\rho = 0.7$) performs somewhat better than the other three ("positive" evidence according to Table 3.1). Otherwise the GCC does not discriminate between the representations. However, the qualitative characteristics of the four representations are illuminating, and allow a substantial comparison of the four similarity models. Each of these representations is discussed in turn.

The common features representation shown in Figure 4.22 contains seven features that explain 78.1% of the variance in the data. The features are highly interpretable, identifying geographical features containing western European nations (Italy, Spain, Germany), Caribbean nations (Cuba, Jamaica), southern African nations (Nigeria, Zimbabwe), Asian nations (China, Japan, Vietnam, Philippines, Indonesia), and Middle Eastern nations (Iraq, Libya). The feature shared by the Philippines and Indonesia identifies a salient subset within the Asian nations, while the feature shared by Germany, Russia, the United States, China and Japan has a 'world powers' interpretation. It is also possible to give an alternative interpretation of the feature shared by Iraq and Libya in political terms, corresponding to 'rogue states'.

The distinctive features representation shown in Figure 4.23 contains five features that explain 71.0% of the variance. The top-weighted feature distinguishes between the developed nations (Italy, Spain, Germany, Russia, United States, China, Japan) and the

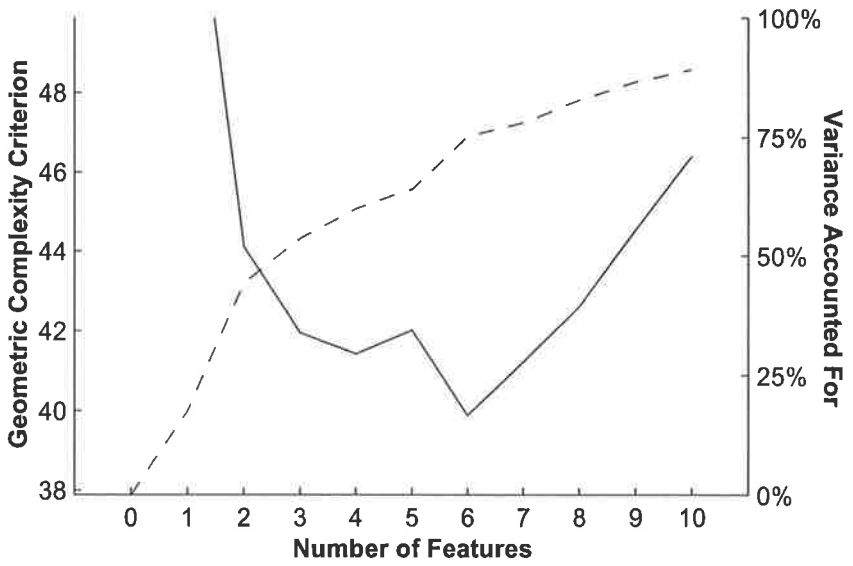

Figure 4.17: GCC (solid line) and VAF (dashed line) for common features representations of the similarity-condition countries data. Four to seven features are preferred.

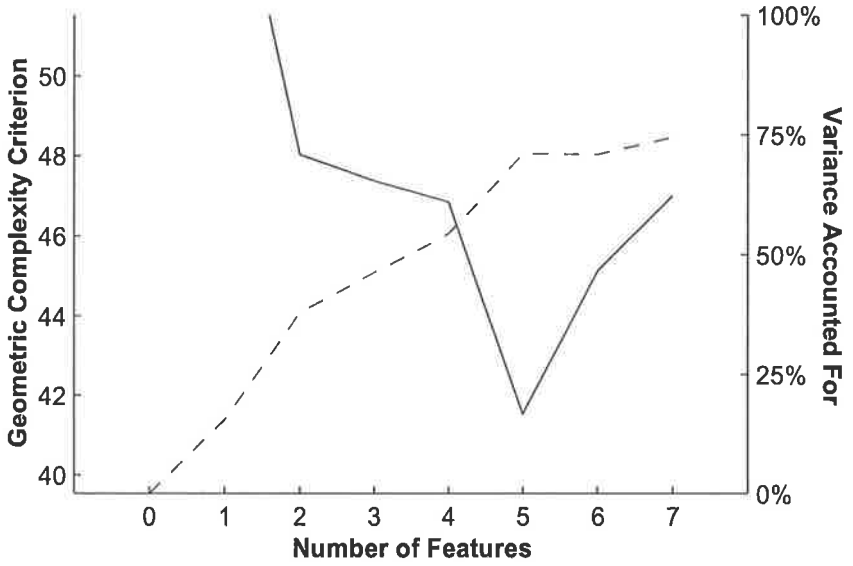

Figure 4.18: GCC (solid line) and VAF (dashed line) for distinctive features representations of the similarity-condition countries data. Five features are preferred.

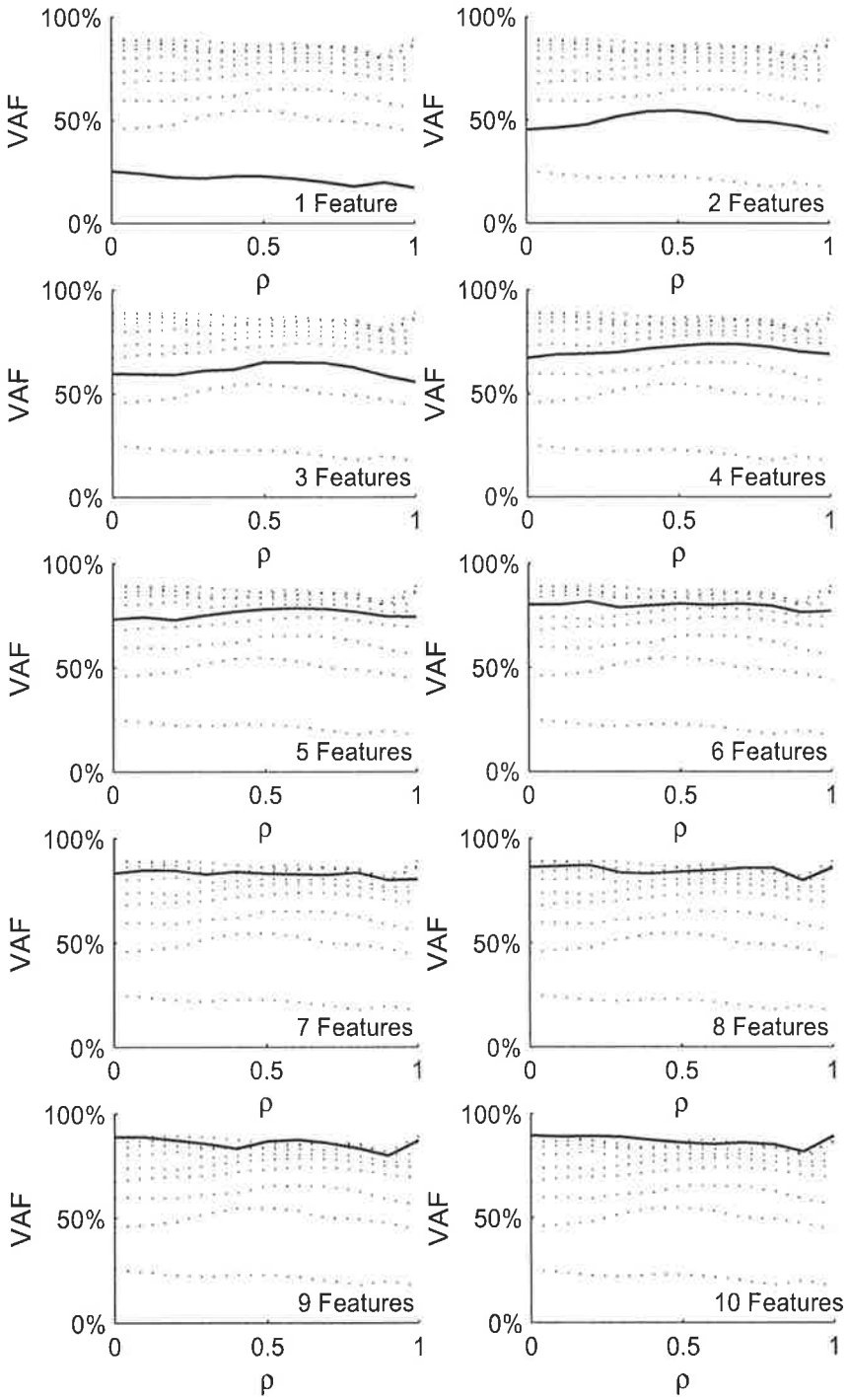

Figure 4.19: The VAF values for TCM representations of the similarity-condition countries data. Each panel contains the same ten plots of VAF values as a function of ρ: each plot corresponds to representations with a particular number of features. Each panel highlights one of the plots: the highlighted plot is identified by the number of features indicated by the text in the bottom right corner of each panel.

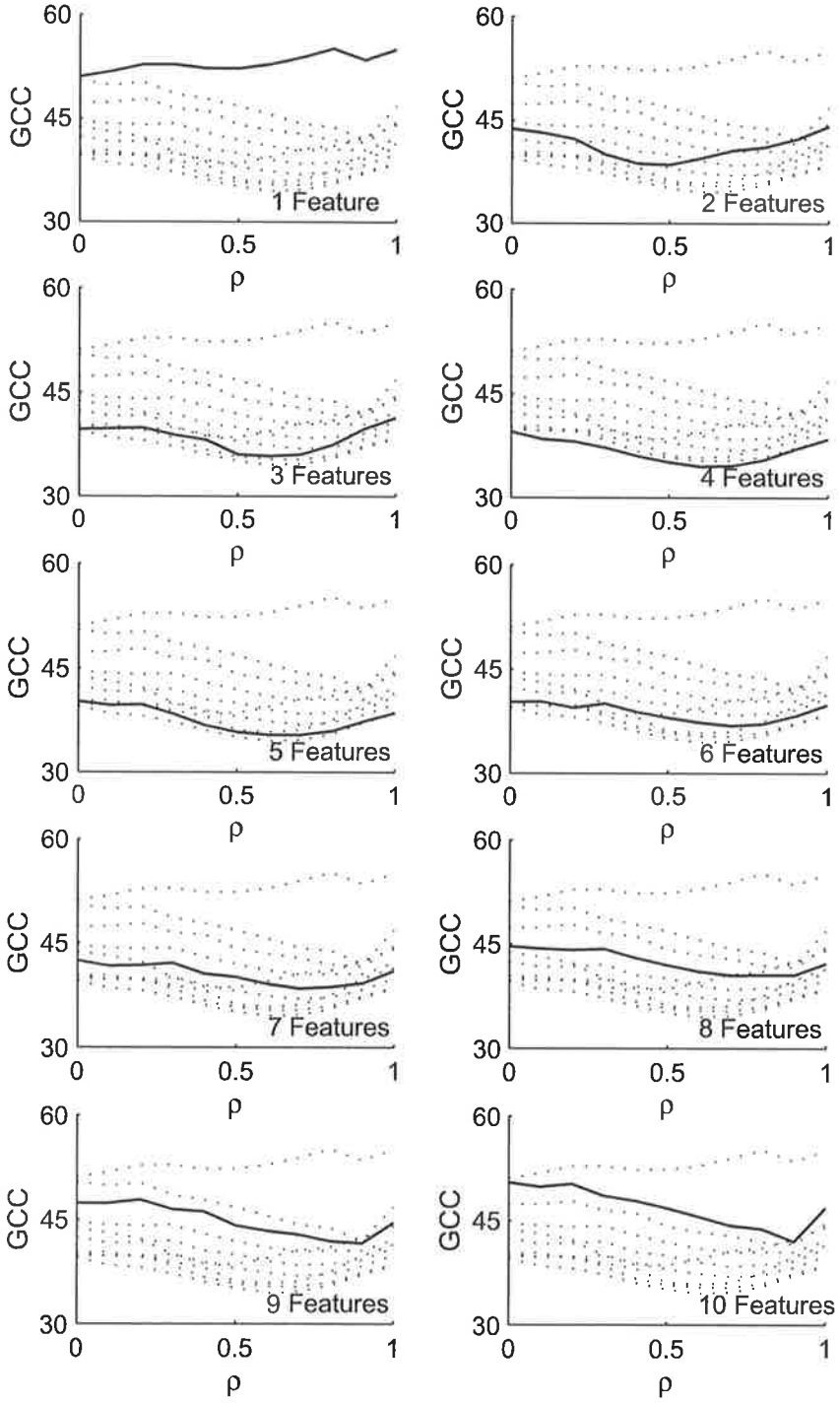

Figure 4.20: The GCC values for TCM representations of the similarity-condition countries data. Each panel contains the same ten plots of GCC values as a function of ρ: each plot corresponds to representations with a particular number of features. Each panel highlights one of the plots: the highlighted plot is identified by the number of features indicated by the text in the bottom right corner of each panel. The GCC favours a four to six feature model with ρ between 0.5 and 0.7.

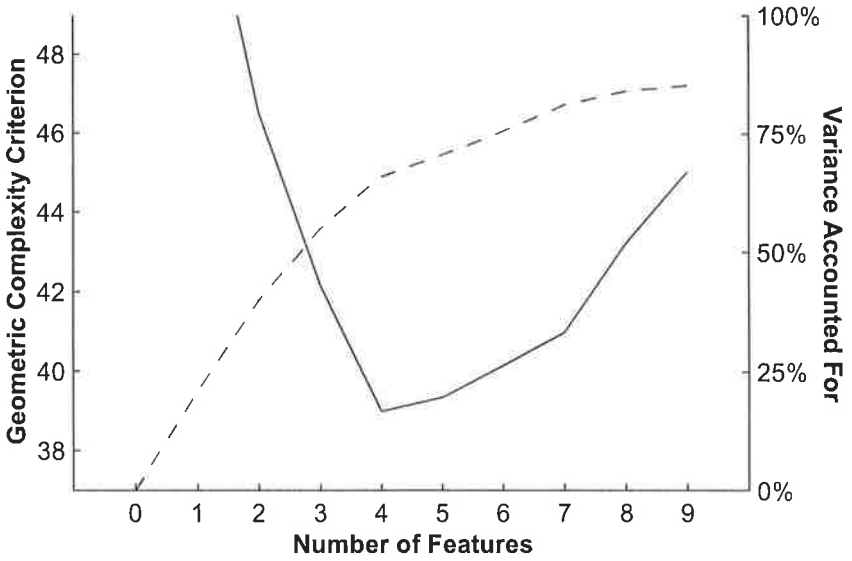

Figure 4.21: GCC (solid line) and VAF (dashed line) for MCM representations of the similarity-condition countries data. Four to seven features are preferred.

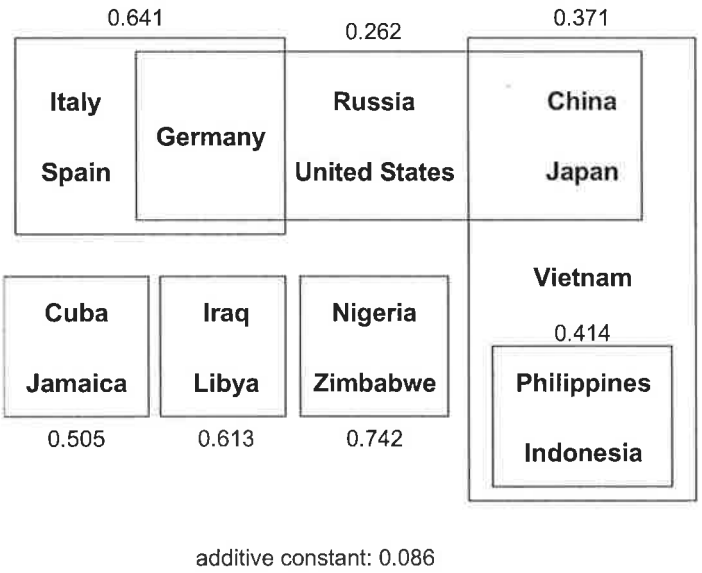

Figure 4.22: Common features representation of the similarity-condition countries data, accounting for 78.1% of the variance (GCC=41.2).

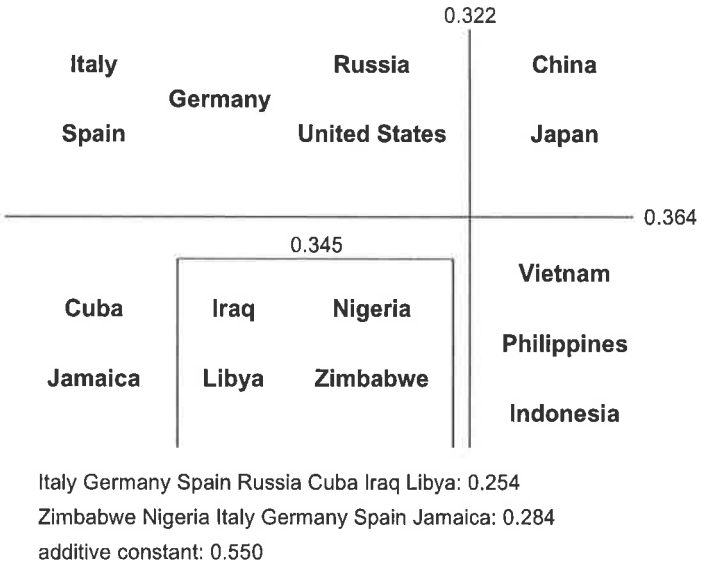

Figure 4.23: Distinctive features representation of the similarity-condition countries data, accounting for 71.0% of the variance (GCC=41.5).

Table 4.13: TCM representation of the similarity-condition countries data, employing a moderate common features bias ($\rho = 0.7$).

Feature	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
Variance Accounted For	80.8%
Geometric Complexity Criterion	36.9%

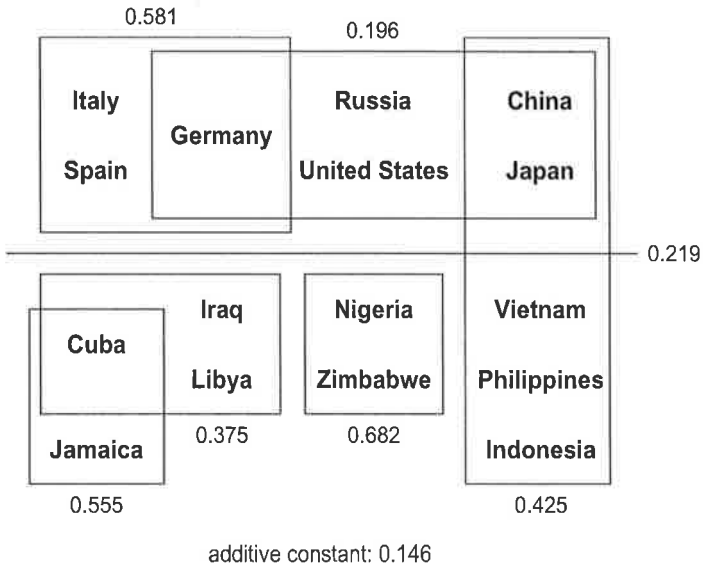

Figure 4.24: MCM representation of the similarity-condition countries data, accounting for 81.2% of the variance (GCC=41.0).

developing nations (Zimbabwe, Nigeria, Cuba, Jamaica, Vietnam, Iraq, Libya, Philippines, Indonesia). Interestingly, the GCC deteriorates only marginally (by 1.8) when China is placed among the developing nations rather than the developed nations, which may be appropriate given China's status as a rapidly developing nation. The next two features capture geographical distinctions: one distinguishes the African and Middle-Eastern nations (Iraq, Libya, Nigeria and Zimbabwe) from the rest of the world, and the other divides the nations into Asian (China, Japan, Vietnam, Philippines, Indonesia) and non-Asian nations. Unfortunately, the remaining two features appear not to reflect any interpretable structure.

Table 4.13 displays the six feature TCM representation with $\rho = 0.7$, explaining 80.8% of the variance. The high ρ value indicates that commonalities are weighted more heavily than differences, as might be expected in this experimental condition. All of the features in Table 4.13 appear in either the common features or distinctive features representations, which is not surprising. Each of the identified subsets is interpretable,

but it is difficult to see what it means for a feature to be $\rho = 0.7$ common and $(1-\rho) = 0.3$ distinctive. Crucially, the 'developed vs developing' distinction does not appear in this representation.

The MCM representation shown in Figure 4.24 explains 81.2% of the variance, and contains six common features and a single distinctive feature. The geographical common features corresponding to western Europe, the Caribbean, southern Africa, and Asia are all present, as is the political 'world powers' feature. The remaining common feature consists of Cuba, Iraq, and Libya, and has a political interpretation as 'rogue states'. Finally, the model also includes the 'developed vs developing' regularity from the distinctive features representation.

Dissimilarity Condition Representations

The dissimilarity data shown in Table 4.10 were converted to "similarity" scores[5] by the simple linear transformation $s_{ij} = 1 - d_{ij}$. The stochastic hillclimbing algorithms were employed to extract representations for each of the four similarity theories, in the same manner as for the similarity-condition data. Despite having roughly the same precision as the similarity-condition data, none of the models were able to extract much structure from the data set. The distinctive model and both contrast models favoured a representation containing a single feature, as shown by Figures 4.25, 4.27 and 4.28, whereas additive clustering did not find a representation superior to a "null" model (containing only the additive constant, explaining 0% of the variance). As it happened, the distinctive features model, the TCM and the MCM all extracted the same representation, shown in Figure 4.29. The sole feature in this model is the same 'developed vs developing'

[5]As highlighted on page 136, it is unlikely that similarity and dissimilarity are simple inverses of one another. So the "similarities" produced in this manner might be more accurately described as "antidissimilarities". Accordingly, the representations derived in this section are representations of antidissimilarity data, not similarity data. Among other things, this experiment demonstrates that antidissimilarity is not the same thing as similarity.

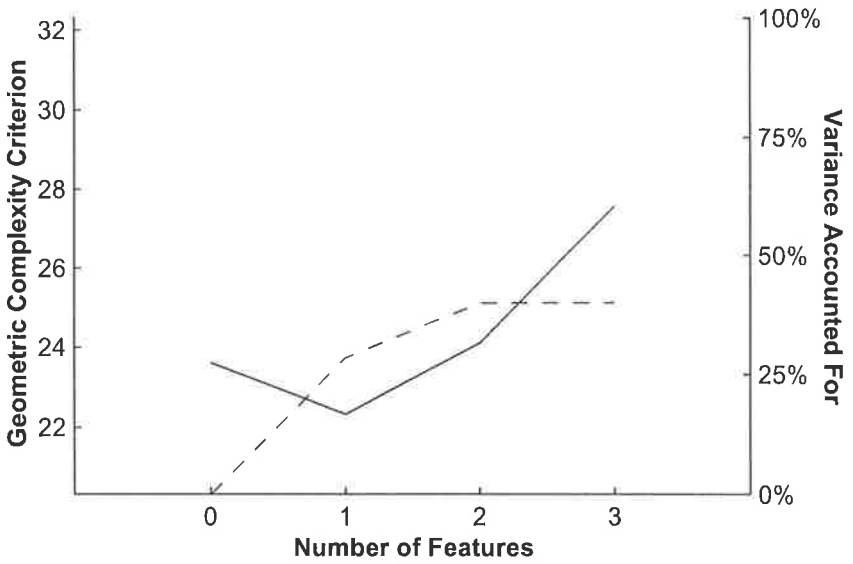

Figure 4.25: GCC (solid line) and VAF (dashed line) for distinctive features representations of the dissimilarity-condition countries data. A single feature is preferred.

feature that appears in the distinctive features (Figure 4.23) and MCM (Figure 4.24) representations of the similarity-condition data.

4.9.3 Discussion

Regarding the Similarity Models

The similarity-condition data provides a rich source of information about the four similarity models. The common features model and the MCM both extract several interpretable commonalities from the data. Similarly, the distinctive features and MCM representations demonstrate at least one important distinction in the domain, namely the 'developed vs developing nations' feature. Given that the TCM preferred a common features bias ($\rho = 0.7$), and that six of the seven MCM features were declared to be common, it appears that the domain is more heavily influenced by common features than distinctive features.

Regarding the distinctive features model, it is a problem that two of the five features

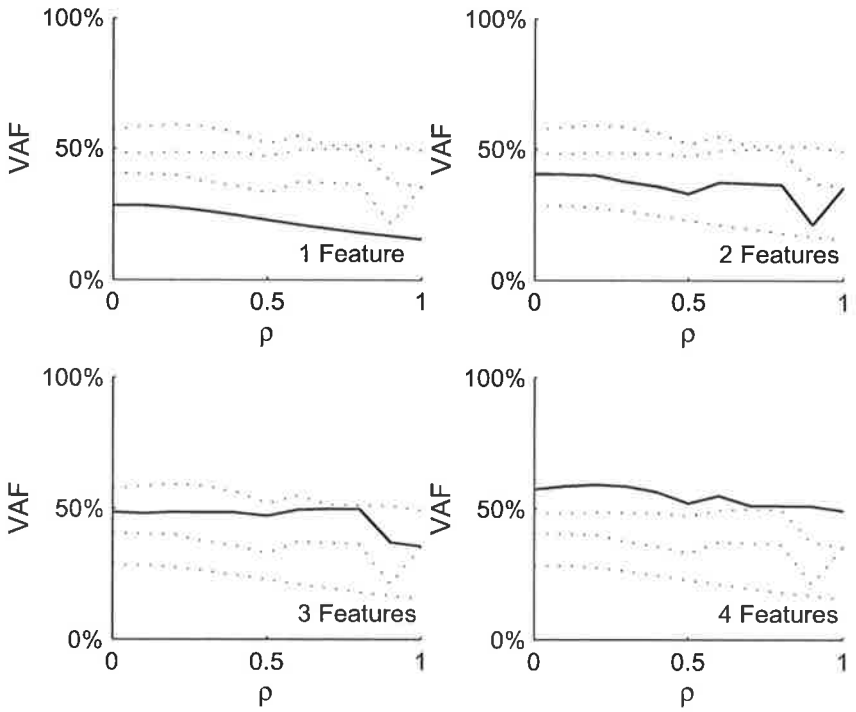

Figure 4.26: The VAF values for TCM representations of the dissimilarity-condition countries data. Each panel contains the same four plots of VAF values as a function of ρ: each plot corresponds to representations with a particular number of features. Each panel highlights one of the plots: the highlighted plot is identified by the number of features indicated by the text in the bottom right corner of each panel

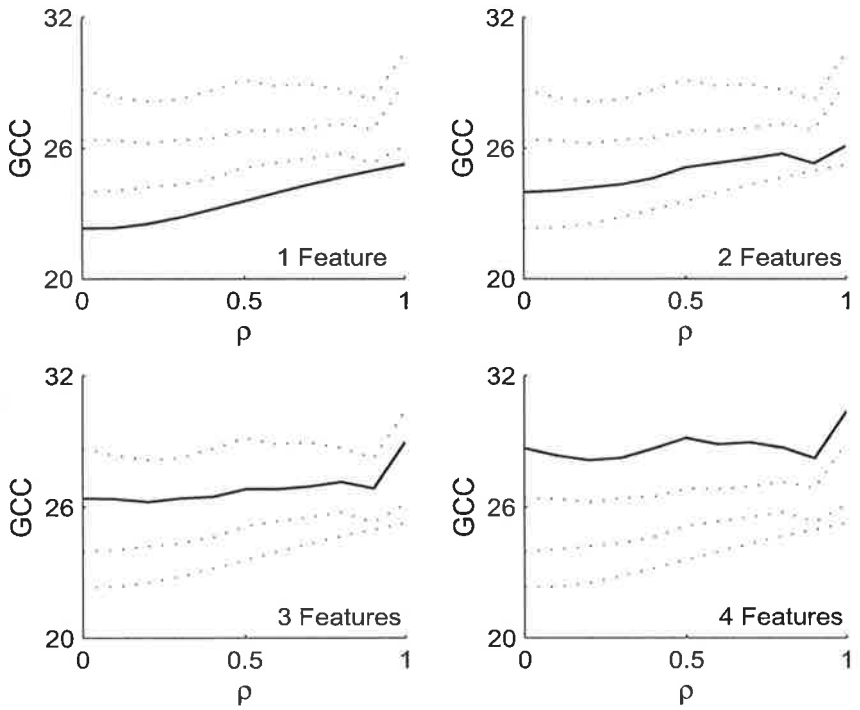

Figure 4.27: The GCC values for TCM representations of the dissimilarity-condition countries data. Each panel contains the same four plots of GCC values as a function of ρ: each plot corresponds to representations with a particular number of features. Each panel highlights one of the plots: the highlighted plot is identified by the number of features indicated by the text in the bottom right corner of each panel. The GCC favours a single feature model with $\rho = 0$.

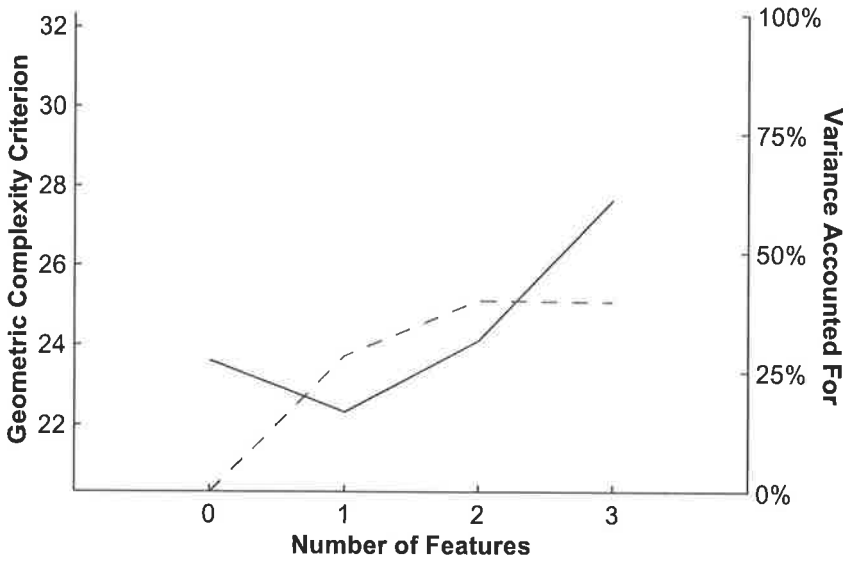

Figure 4.28: GCC (solid line) and VAF (dashed line) for MCM representations of the dissimilarity-condition countries data. A single feature is preferred.

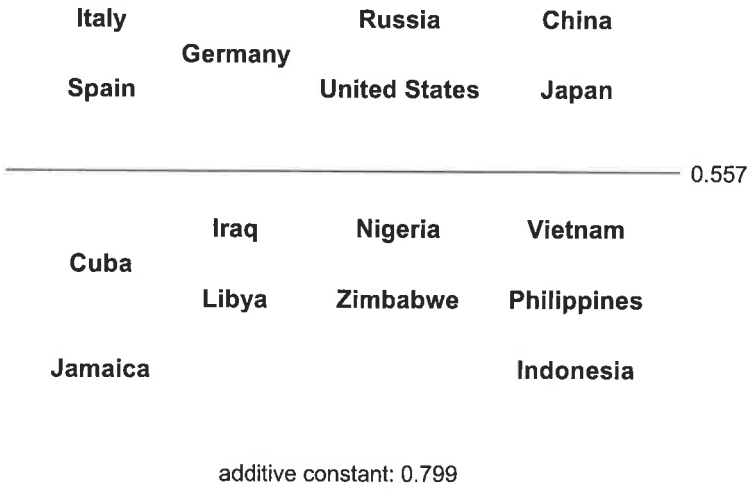

Figure 4.29: The representation of the difference-condition countries data favoured by the distinctive features model, the TCM and the MCM, accounting for 28.4% of the variance (GCC=20.9).

lack an obvious interpretation. It may be that these two features partition the stimulus set into a set of plausible groupings without the features themselves being meaningful. For instance, Italy, Germany and Spain are the only nations that are both "in" the two features. These three nations are better understood as possessing a common 'western Europe' feature. The distinctive features representation can mirror this structure, though in a less convincing manner. Psychologically speaking, it is undesirable to recover uninterpretable features. Furthermore, it may be argued (see Lee & Navarro, 2002 for instance) that a feature should capture a regularity that may be attended-to individually, and therefore be separable in Garner's (1974) framework. The distinctive features approach does not allow the western Europe regularity to be attended-to separately from the other groupings that emerge from these two features.

The failure of the TCM representation to include the 'developed vs developing' distinction is revealing, arising as it does from the value of ρ. Since the data set is dominated by common features, the TCM is required to employ a common features bias. However, the 'developed vs developing' feature does not make sense as anything but a purely distinctive feature, since any common features component makes one half (either developed or developing) more salient than the other. The TCM is therefore incapable of expressing this distinction while still capturing the common features that predominate in the data.

The MCM, on the other hand, expresses the common features and the distinctive features in an interpretable fashion. It is interesting to compare the 'world powers' common feature to the 'developed vs developing' distinctive feature. As previously suggested, the 'developed vs developing' feature is inherently distinctive: two developed nations need not be similar, but a developed nation and a developing nation are necessarily different. In contrast, two nations are more alike if they are both major world powers, but this has no implications regarding two nations that are not. It is precisely this kind of structure

that the MCM was designed to capture.

Regarding the Two Conditions

Although the dissimilarity condition provided very little evidence on which to distinguish between similarity models, this by itself raises an interesting issue. Despite having roughly the same precision as the similarity-condition data, only a single feature could be extracted. Therefore, either the dissimilarity-condition data encodes very little structure, or the structure that exists cannot be effectively modelled by any of the featural similarity models considered here. Given that over 70% of the variance remains unexplained the latter seems more probable. Whichever is the case, it is evident that the two conditions are not simple inverses of one another. It is therefore appropriate to ask why the two conditions differed so fundamentally.

In terms of informal feedback, difficulties with the dissimilarity task were reported far more frequently than with the similarity task. This may reflect the domain and the structure of the task. For instance, consider the four nations Japan, Iraq, Italy and Jamaica, no two of which may appear particularly alike. In order to make a response, a participant must seek some basis upon which to make a decision. In the similarity condition, a single commonality allows a decision (e.g., that Japan and Italy were both part of the Axis in World War II). However, it is more difficult to find a simple decision rule in the dissimilarity task, or rather, there are too many such rules (e.g., Japan is industrialised, whereas Jamaica is not; Iraq is Islamic, whereas Italy is Christian). Ironically, because there are always more differences than similarities between countries, a commonality is more decisive than a distinction. Correspondingly, commonalities play a more important role in the decision-making process, and the similarity task becomes the easier of the two.

A second relevant factor is the presentation of countries in lists of four. Suppose

for the moment that a particular participant attempts to make all decisions based on geographical knowledge, and is presented with Italy, Spain, Nigeria, and Zimbabwe. If asked to select the most similar nations, the geographical features reduce the potential responses to 'Italy & Spain' and 'Nigeria & Zimbabwe'. However, in the dissimilarity task, the other *four* options are left. In general, any individual piece of knowledge will most likely be more helpful in the similarity task than the dissimilarity task.

4.10 Ratio Models: The Road Less Travelled

The four similarity models tested in this chapter are not the only ways of constructing a featural model. An alternative approach is to base clustering models on Tversky's (1977) Ratio Model,

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

Although it is easy to specify a summed saliencies functional form for Λ it is not immediately obvious how best to choose values for the parameters α and β. For instance, choosing $\alpha = \beta = \frac{1}{2}$ yields the model employed by Eisler and Ekman (1959)

$$\hat{s}_{ij} = 2\frac{\Lambda(\mathbf{f}_i \cap \mathbf{f}_j)}{\Lambda(\mathbf{f}_i) + \Lambda(\mathbf{f}_j)}.$$

Therefore, the corresponding clustering model would be,

$$\hat{s}_{ij} = \frac{\sum_k w_k f_{ik} f_{jk}}{\sum_k w_k (f_{ik} + f_{jk})},$$

according to which similarity is given by dividing the common features saliency by the saliency of the first stimulus plus the saliency of the second. Alternatively, Gregson (1975) employed a model corresponding to $\alpha = \beta = 1$,

$$\hat{s}_{ij} = \frac{\Lambda(\mathbf{f}_i \cap \mathbf{f}_j)}{\Lambda(\mathbf{f}_i \cup \mathbf{f}_j)}.$$

The clustering model yielded under a summed saliencies function is

$$\hat{s}_{ij} = \frac{\sum_k w_k f_{ik} f_{jk}}{\sum_k w_k \max(f_{ik}, f_{ik})},$$

which compares the common features saliency to the sum of the saliencies of the features possessed by at least one stimulus. As a third possibility, Bush and Mosteller (1951) employed the asymmetric model that results when $\alpha = 0$, $\beta = 1$,

$$\hat{s}_{ij} = \frac{\Lambda(\mathbf{f}_i \cap \mathbf{f}_j)}{\Lambda(\mathbf{f}_i)}.$$

The resulting clustering model,

$$\hat{s}_{ij} = \frac{\sum_k w_k f_{ik} f_{jk}}{\sum_k w_k f_{ik}}$$

normalises the common features saliency by the saliency of one of the two stimuli. This model happens to correspond to the probability of generalising from stimulus i to stimulus j under Tenenbaum and Griffiths' (2002a) Bayesian theory of generalisation.

Each of these three models makes sense in its own right. That being said, the plausibility of these three cases does not imply that the Ratio Model itself is a good model. After all, the findings presented here suggest that, even though two special cases of Tversky's Contrast Model are interpretable psychological models (the common features model and the distinctive features model), the Contrast Model is problematic. As Arabie (1994, p101) observes in a different context, "[a] generalization is not necessarily more elegant than its special cases". To illustrate this, substituting $\alpha = \frac{1}{4}$ and $\beta = 3$ yields the model,

$$\hat{s}_{ij} = \frac{\Lambda(\mathbf{f}_i \cap \mathbf{f}_j)}{\Lambda(\mathbf{f}_i \cap \mathbf{f}_j) + \frac{1}{4}\Lambda(\mathbf{f}_i - \mathbf{f}_j) + 3\Lambda(\mathbf{f}_j - \mathbf{f}_i)},$$

which lacks a natural interpretation. In short, there may be good reason to adhere to the special cases of the ratio model.

4.11 Summary & General Discussion

Four featural theories of stimulus similarity – the common features model, the distinctive features model, Tversky's Contrast Model, and a Modified Contrast Model – were evaluated in this chapter. Concrete versions of each general theory were produced, and applied to several sets of empirical data using the stochastic hillclimbing algorithms and Geometric Complexity Criteria. Of the stimulus domains, the kinship terms and the countries are best regarded as conceptual stimuli, whereas the cartoon faces are largely perceptual in nature. The kinship terms and faces stimuli are both highly constrained stimuli, in that the kinship terms all refer to very specific relationships and the variation in the faces was fairly limited. In contrast, the countries domain is highly naturalistic and require more general knowledge. Furthermore, the collection methodologies differed across domains: the kinship data involved a sorting task, the faces were judged on a ratings scale, and the countries data were obtained using a forced-choice decision task. Given this diversity, it is difficult to attribute consistent findings to domain-specific or methodological factors.

In general, the four models all lead to representations that provide a good fit to the data in a relatively parsimonious manner. While there are systematic differences in the complexity of the four models, the major difference between them lies in the interpretability of the representations, and the ability to capture important qualitative characteristics of the data. The common features model and the distinctive features model discover important but different regularities in the data sets, confirming the notion that both are required in a featural model of similarity. The TCM, however, fails to include important features such as the 'developed vs developing' feature in the similarity-condition countries representation, or the gender distinction in the kinship data. This shortcoming can be seen to result from the reliance on the decision variable ρ which is

applied uniformly to all features. The success of the MCM in capturing precisely such aspects occurs because it embeds the commonality or distinctiveness of a feature within the representation, and allows individual features to differ in this regard. It may not be too strong to claim that the MCM constitutes something like best of both worlds, where the TCM is closer to the worst.

5. Prototype Space Scaling

This chapter pursues a variant of multidimensional scaling (MDS, see Section 2.2) that allows a single point to represent any number of stimuli. Accordingly, the dissimilarity between two stimuli represented by the same point is given by the additive constant: in all other regards the similarity model is identical to MDS. This similarity model is mathematically appealing, as it limits the number of free parameters in the representation. This advantage notwithstanding, it is more important to introduce a solid psychological foundation for this representational model. The model disrupts the correspondence between stimuli and points, so it no longer makes sense to refer to the derived space as a stimulus space. Instead, the natural interpretation of a point that stands for a number of stimuli, even though those stimuli differ, is a *kind* of thing, a class or a category.

Rosch (1978) provides a useful conceptualisation of categorical structure, in terms of a horizontal dimension and a vertical dimension[1]. The vertical dimension describes the level of generality of a category. For example, since chairs are a type of furniture, the category of "furniture" is more general (and hence higher up) than the category of "chair". The horizontal dimension describes the internal structure of categories at the same level of generality, such as "table" and "chair". Together, these two dimensions yield a taxonomy of categories, as illustrated in Figure 5.1.

In what has become known as the prototype view of categorisation, Rosch and others

[1]The word "dimension" should not be taken too literally: the two dimensions merely denote qualitatively different regards in which two categories may differ.

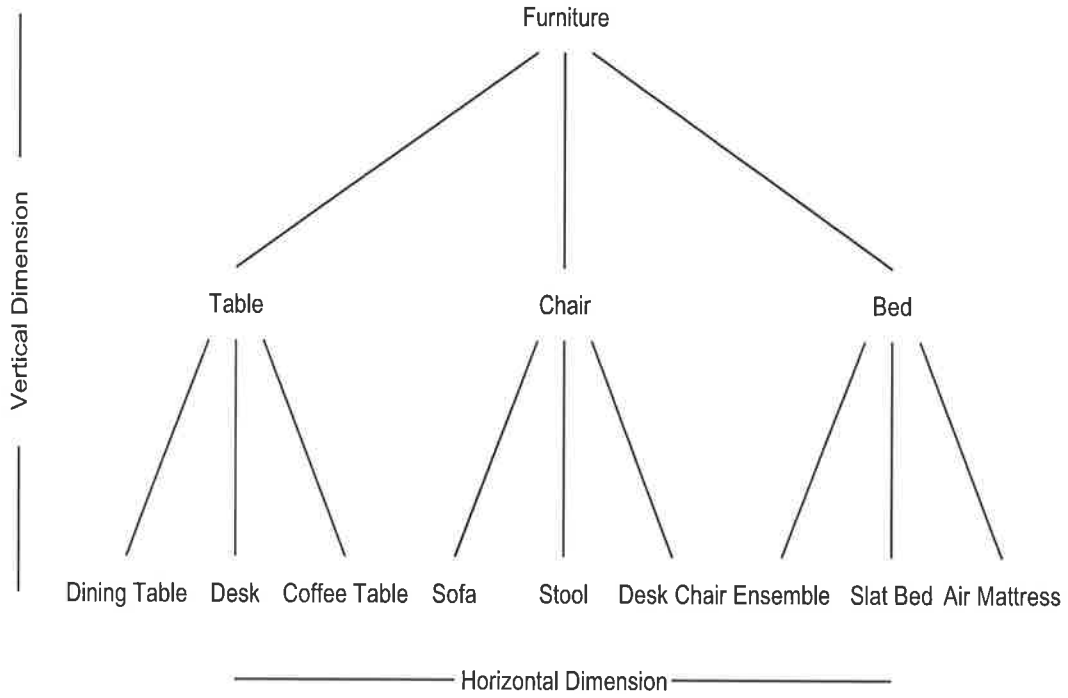

Figure 5.1: Rosch's vertical and horizontal dimensions of categorical structure.

have argued that the internal structure of a category (the horizontal dimension) can be characterised in terms of a single, idealised instance of the category, the *prototype* (e.g., Rosch, 1975, 1978; Rosch & Mervis, 1975; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Mervis & Rosch, 1981; Smith & Medin, 1981). The prototype view has several desirable features: for example, it predicts the observations that categories do not have clear-cut boundaries (Wittgenstein, 1953), that some members are better examples than others (e.g., a robin is a better example of a bird than a penguin; see Rosch, 1975 and Rosch & Mervis, 1975), and that better examples are processed faster (see Mervis & Rosch, 1981). Although the prototype view does have drawbacks, it remains a powerful tool for understanding categorical structure (see Komatsu, 1992 for an overview).

As Rosch (1978) observes, this qualitative account of prototypes provides some constraints on the representation of categories, but does not restrict such representation

to a single formalism. The spatial approach developed in this chapter can be seen to be an example of a prototype representation: several stimuli are represented by a single point, which may be labelled a prototype[2]. It is for this reason that the representations derived under this model are referred to as *prototype spaces*, and the techniques for extracting these spaces as *prototype scaling*. Nevertheless, it is important to recognise that this is merely one way of instantiating these ideas.

5.1 Dissimilarity Between Prototypes

Irrespective of the manner in which prototype scaling algorithms are developed, it is important to formulate the prototype dissimilarity model appropriately. Prototype scaling involves partitioning the stimulus set into a number of mutually exclusive categories, and therefore requires a psychologically plausible measure of the dissimilarity between categories.

From a numerical standpoint, Gordon (1999) observes that when partitioning a set of objects into classes, one could either minimise the heterogeneity of the classes or maximise their isolation. In this sense, a heterogeneity measure is a measure of the extent to which the members of a class differ from one another, and an isolation measure is a measure of the extent to which two classes differ from one another. These two goals are broadly compatible. There are examples of isolation and heterogeneity measures that are genuine inverses, so that – for a fixed number of classes – optimising one necessarily implies optimising the other. However, this situation is the exception rather than the rule: usually, there is some difference between a heterogeneity measure and an isolation measure.

From a psychological standpoint, it is argued that maximising isolation is the ap-

[2]A note on terminology: the term "prototype" is used here to refer to the typicality information represented by a point in the space, whereas the words "cluster", "class" or "category" generally refer to the set of stimuli denoted by such a point.

propriate framework for prototype scaling. Since it is the between-category structure that is modelled by prototype scaling, the measure to optimise should be a measure of between-cluster variation (i.e., isolation), however moot the point may be from an algorithmic standpoint. Furthermore, one may wish to remain open to the suggestion that within-category structure differs fundamentally from between-category structure (see, for instance, Barsalou, 1989). This point is significant inasmuch as one may subsequently wish to analyse the internal structure of the categories using other tools (such as trees). Hence it makes sense to (in theory) leave this internal structure of the categories alone, and work only with the between-category variation.

Two candidate measures for prototype dissimilarity are considered. These measures are based on two models of inter-class similarity considered in papers by Osherson, Smith, Wilkie, Lopez, and Shafir (1990) and Tenenbaum and Griffiths (2001). In the first, the empirical similarity between two categories is the mean similarity between stimuli in different categories (the *average similarity model*), and in the second, the similarity between two classes is given by the similarity of their two most similar members (the *maximum similarity model*). If similarity and dissimilarity are assumed to be linearly related (i.e., $s_{ij} = c - d_{ij}$ for some sufficiently large constant c), the mean similarity model is also a mean *dissimilarity* model, and the maximum similarity model becomes a minimum dissimilarity model. These two models are referred to by Gordon (1999) as *cut distance* and *split distance* respectively. Formally, the dissimilarity between two groups of stimuli X and Y under a mean dissimilarity model is

$$d_{XY} = \frac{1}{n_X n_Y} \sum_{i \in X} \sum_{j \in Y} d_{ij},$$

where n_X and n_Y represent the number of stimuli in each cluster. Contrastingly, the cluster dissimilarity is

$$d_{XY} = \min_{i \in X, j \in Y} d_{ij}$$

if a minimum dissimilarity model is used. On the whole, the two measures are quite similar, and may yield similar results, but differences can occur. The mean suggests "distance between centers", whereas the minimum suggests "distance between edges". The difference between the two measures is illustrated in Figure 5.2, in which the dotted line shows the mean distance between three (spatially represented) categories, and the unbroken line shows the minimum distance between them. For the mean measure, the three distances are ordered $a > b > c$, but for the minimum distance, the ordering becomes $C > A > B$. For the mean measure, a is only marginally longer than b, but both are substantially longer than c. For the minimum measure, A and B differ only slightly, but C is substantially longer. So, with this in mind, the orderings might be better expressed as $a \approx b > c$ and $C > A \approx B$. This possibility is just as important as the first, since it suggests that, for both measurements, one proximity estimate is quite different to the others, but in one case it is larger, and in the other it is smaller.

Are the squares in Figure 5.2 closer to the circles and triangles than they are to each other, or are the squares further away? Returning to the original notion of categories, it is possible to argue for either. Suppose the categories represented mammals (circles), birds (triangles) and dinosaurs (squares). If asked to compare a typical mammal to a typical bird and a typical dinosaur, one might compare a dog, a robin, and a tyrannosaurus rex, and conclude that dinosaurs lie a long way from the other two (i.e., $a \approx b > c$). Alternatively, by making edgewise comparisons, one might compare an ostrich to a velociraptor, a triceratops to a rhinoceros, and a bat to a sparrow. In this case, the dinosaurs could end up closer to the other two categories than they are to each other (i.e., $C > A \approx B$). The crucial feature of this example is the fact that $c \approx C$: the prototypical comparison between birds and mammals (in this example) is not far different to the edgewise comparison, whereas for comparisons involving dinosaurs, this is not the case ($a > A$, $b > B$). The dinosaur category in this example is far more heterogeneous

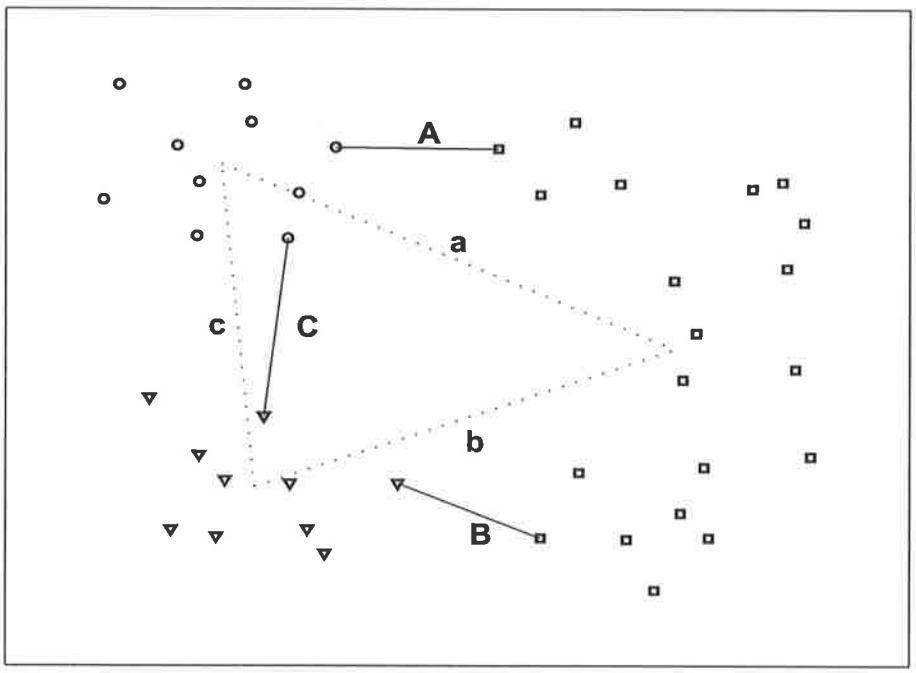

Figure 5.2: Mean (dotted line) and minimum (unbroken line) distances between three spatially displayed categories, corresponding to the circular, square and triangular markers.

than the mammal or bird category, and so can have outliers quite close to the mammal and bird categories while still having its centroid some distance away. This phenomenon does not rely on the categories having different shapes: the same phenomenon would occur if all three were circular, but the dinosaur category had a much larger diameter.

These matters notwithstanding, there is a practical justification for adopting the average similarity model in a prototype scaling context: the maximum similarity model does not satisfy the triangle inequality, and is therefore not a distance metric. In cases when the average similarity model is inappropriate due to violations of the triangle inequality, it makes sense to represent the data using some other representational formalism (e.g., featural representation). Hence, the analyses presented in this chapter use the average dissimilarity model, and the prototype spaces are interpretable as representations in which *typicality* information about mutually exclusive categories is spatially represented.

5.2 Prototype Scaling Algorithms

Having developed a psychological theory for prototype spaces, this section introduces three algorithms for prototype scaling. Again, the distinction is emphasised between the psychological model and the fitting algorithm (e.g., Kruskal, 1964b; Shepard & Arabie, 1979): the psychological implications of these algorithms are minimal.

5.2.1 HEAPS

The first approach to prototype scaling is the most computationally expensive: it is therefore referred to as a Highly Expensive Algorithm for Prototype Scaling (HEAPS). It starts with a single cluster to which all stimuli belong, embedded as a single point in a one-dimensional space (which, allowing for an additive constant, explains 0% of the variance). There are two nested loops within HEAPS, the inner one adding clusters, the outer one adding dimensions. For a given number of classes m and dimensions k,

HEAPS uses a stochastic hillclimbing approach to search through the set of possible partitions. That is, it maintains a vector assigning each stimulus to a cluster, and "flips" stimuli into different clusters in search of a better partition, restarting whenever a better one is found, and accepting the best seen partition after a certain number of local minima (10 in this case). The adequacy of a partition is assessed by performing a multidimensional scaling on the dissimilarities between the m classes. Since every candidate partition is evaluated using MDS, this procedure is very expensive computationally, and will not be viable for large problems.

5.2.2 LEAPS

The second algorithm, called LEAPS (a Less Expensive Algorithm for Prototype Scaling), also treats the partitioning and scaling as a single optimisation problem. Like HEAPS, LEAPS starts with a one cluster, one dimensional solution and adds clusters and dimensions. For a given partition, LEAPS finds the best MDS solution, and then for these co-ordinates finds the best partition. If the new partition differs from the old one, the process repeats. A local minimum results when neither the co-ordinates nor the partition changes, and the process continues for an arbitrary number of local minima (10 in this case). LEAPS is less computationally expensive than HEAPS, but may be more likely to become permanently stuck in a globally suboptimal local minimum.

5.2.3 CAPS

The third approach to fitting prototype models, CAPS (which is a Cheap Algorithm for Prototype Scaling), uses the raw proximities to find the partition, then scales the resulting prototypes using MDS. This approach has the advantage that, since the two aspects are separate from one another, only one MDS is ever performed, so the computation required is greatly reduced. In the analyses presented in this chapter, the partition is found

using a modification of average-link clustering (Hartigan, 1975). As with average-link clustering, the two existing classes that are most proximal to one another (calculated using the mean dissimilarity measure) are merged at each step of the clustering. This process is carried out for $n - m$ iterations, at which point there exist m disjoint clusters. Those remaining clusters constitute the average-link partition. This average-link partitioning method maximises the dissimilarity between the two most similar clusters. If one wished instead to maximise the average cluster dissimilarity (or squared dissimilarity), the average-link partition could be used as an initial solution for a stochastic hillclimbing process, flipping stimuli in and out of clusters until an optimum is reached. However, this approach is substantially more expensive computationally: since the average-link approach produces reasonable solutions to the problems considered in this chapter, this complication is not introduced here.

5.3 Monte Carlo Study III: Culling the Weak

This section presents a Monte Carlo evaluation of HEAPS, LEAPS and CAPS. The aim of this study was simply to compare the performance of the three algorithms on a straightforward task, in order to eliminate candidates that fail to perform sufficiently well.

5.3.1 Design

The demonstration involves a data set of 25 items that properly belong to one of five mutually exclusive categories. The stimuli can be represented perfectly in a two dimensional space, shown in Figure 5.3. The simulation involved adding Gaussian noise ($\mu = 0$, $\sigma = .05$) to the data, and then attempting to recover the five categories and the locations of their prototypes in the space. This procedure was repeated 15 times, since each of the three algorithms has a stochastic element (though in the case of CAPS, this

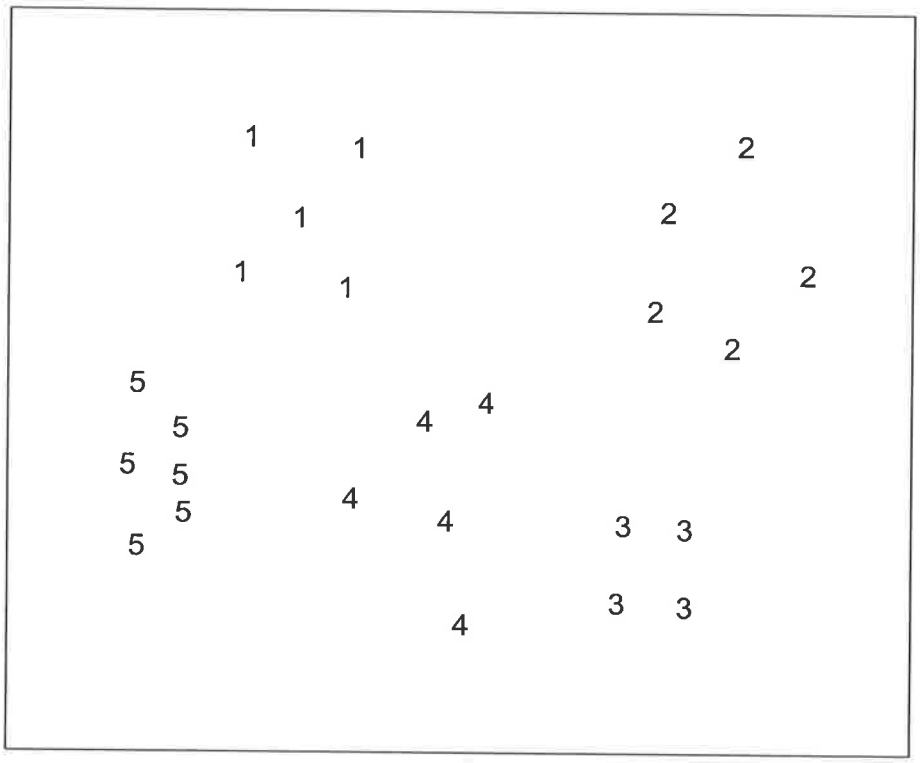

Figure 5.3: The toy data set used in the Monte Carlo study. Each of the 25 stimuli belongs to one of 5 categories.

is purely for the scaling component).

5.3.2 Results

This analysis considers only the performance of the three algorithms at finding an appropriate two-dimension five-cluster solution. HEAPS produced solutions with an average VAF of 64.5% (std dev = 3.1%), providing slightly superior fits to LEAPS, with a mean VAF of 61.9% (std dev = 1.1%). Neither, however, produced solutions as good as CAPS, with a mean of 79.9% (std dev $< 10^{-7}$%). However, these numbers should be interpreted with some caution. The partition-finding procedure within CAPS is deterministic, and in this case finds the correct partition. The low standard deviation results from the fact that with the partition correctly identified, the problem is reduced to a

5-point MDS in two dimensions, which is easily accomplished. Neither HEAPS nor LEAPS managed to partition the data set correctly on any of the 15 trials. Since the average-link procedure used by CAPS produced the correct answer in a deterministic fashion, CAPS made no misclassifications on any trial. After finding the optimal correspondence between derived clusters and true categories, it was found that HEAPS made between 6 and 11 misclassifications, with a mean of 8.35 (std dev = 1.36), whereas LEAPS made 9 misclassifications on every trial.

5.3.3 Discussion

This demonstration suggests that neither HEAPS nor LEAPS perform sufficiently well to merit further consideration as useful approaches to prototype scaling. Although treating the partitioning and scaling aspects as a single optimisation task has the intuitive appeal of allowing the partition and the spatial solution to co-evolve (and potentially achieve superior solutions), in practice the combined problem appears to be a difficult one, and more vulnerable to local minima problems. So, although the approach embodied by CAPS may cause it to become permanently stuck in a suboptimal partition, it appears that in practice it suffers from this problem far less than HEAPS or LEAPS. On the basis of the results of this evaluation, it is argued that neither HEAPS nor LEAPS represent sufficiently promising algorithms to warrant further investigation. Both are computationally expensive and neither fits the data as well as CAPS. Given that the problem lies with local minima, and that local minima problems may be more severe with real data than artificial data, this difference is likely to be more pronounced in practice. Consequently, subsequent analyses use only the CAPS algorithm.

5.4 Complexity, Precision and Categorical Information

In the previous section the appropriate number of prototypes and the dimensionality of the embedding space were already known. This is not the case when modelling empirical data, so some means of specifying the number of prototypes and dimensions is required. Once again the tools of precision and model selection are employed. Given that the Geometric Complexity Criterion is not easily specified for spatial representations (see Section 3.4), the Bayesian Information Criterion (BIC, see Section 3.1.2) is used to guide model selection in prototype scaling. Fortunately, Lee (2001a) has argued that the number of free parameters is a reasonable approximation of model complexity for spatial representations. If m denotes the number of prototypes and k denotes the number of dimensions, the first point is fixed at the origin (without loss of generality for translation-invariant metrics such as the Minkowski family), and the number of free parameters is $k(m-1)+1$. Therefore, the BIC for any given prototype space is given by,

$$\mathrm{BIC} = \frac{1}{s^2}\sum_{i<j}\left(d_{ij}-\hat{d}_{ij}\right)^2 + (k(m-1)+1)\ln\frac{n(n-1)}{2},$$

where n is the number of stimuli and s denotes the precision to which the data is to be modelled. If a small value is chosen for s, more clusters and dimensions will be preferred by the BIC.

When setting s, two factors require consideration. Firstly, it is inappropriate for s to denote a level of precision greater than that of the data itself, σ. For instance, if the estimated data precision $\hat{\sigma}$ is measured at 0.05, then s should not be set below 0.05. In previous chapters the modelling precision s has been assumed to be identical to the data precision estimate $\hat{\sigma}$ because the aim has been to model all of the regularity present in the data but not the noise. In the case of prototype scaling, this may not always be true. Prototype scaling is concerned with capturing and representing categorical or prototypical information about stimuli. Since it will frequently be the case that the data

reflect information about individual stimuli as well as categorical information, it will sometimes be useful to set $s > \hat{\sigma}$. To return to Rosch's (1978) notion of horizontal and vertical dimensions (see Figure 5.1), it may be useful to think of a prototype space as a representation of the horizontal dimension, and some particular level of generality (vertical dimension) specified by s.

5.5 Three Illustrative Applications

In this section CAPS is applied to three sets of similarity data, in order to discuss the implications of the derived representations for prototype scaling. Given the very high consistency of CAPS across runs, the results displayed in this section reflect only a single run, though informal investigation suggested that these results are grossly typical of the behaviour of CAPS on these data sets.

5.5.1 Risks

The first data set analysed is Johnson and Tversky's (1984) risk similarity data, described in Section 2.1. CAPS was applied to the data, with the number of clusters ranging from 1 to 18 (a full stimulus space), and the embedding space ranging from 1 to 4 dimensions (using the Euclidean metric). The Variance Accounted For by each of these solutions is shown in Figure 5.4. Without the raw data, empirical precision estimates could not be obtained, but Figure 5.5 shows the trade-off between fit and complexity assuming modelling precision values s of 0.05, 0.10, 0.15 and 0.20. With this in mind, Figure 5.6 shows the six cluster two dimensional representation preferred when $s = 0.10$, which explains 57.9% of the variance. The "diseases" category contains the stomach cancer, lung cancer, heart disease, leukemia and stroke stimuli; the "accidental falls" category contains only the accidental falls stimulus; "vehicle accidents" contains traffic and airplane accidents; toxic chemical spills and nuclear accidents make up the "environmental

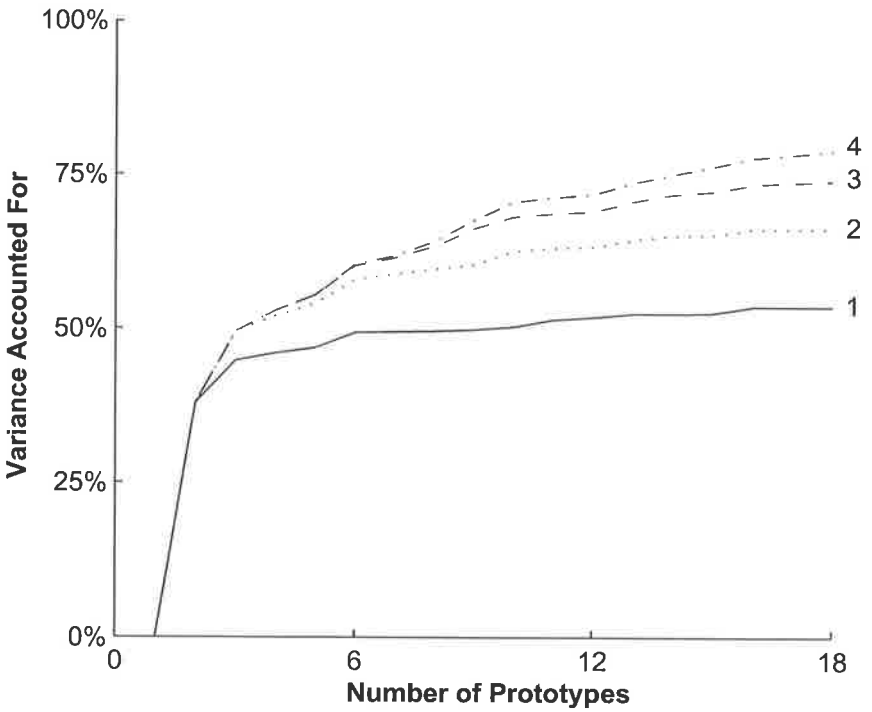

Figure 5.4: Variance Accounted For by prototype space representations of Johnson and Tversky's (1984) risk data. The number of prototypes m ranges from 1 to 18, and the number of spatial dimensions k ranges from 1 to 4. The number of dimensions is marked adjacent to the plots.

damage" category; war, homicide and terrorism count as "acts of violence"; and "natural disasters" include fire, lightning, tornados, floods and electrocution.

The corresponding two dimensional stimulus space is shown in Figure 5.7, and explains 66.1% of the variance. Five of the six categories correspond to connected, bounded regions in the space, in line with Shepard's (1987) identification of such regions with natural kinds. The vehicular accidents category may be disjoint because of the association between airline accidents and things like lightning and terrorism, as well as the link between traffic accidents and toxic chemical spills.

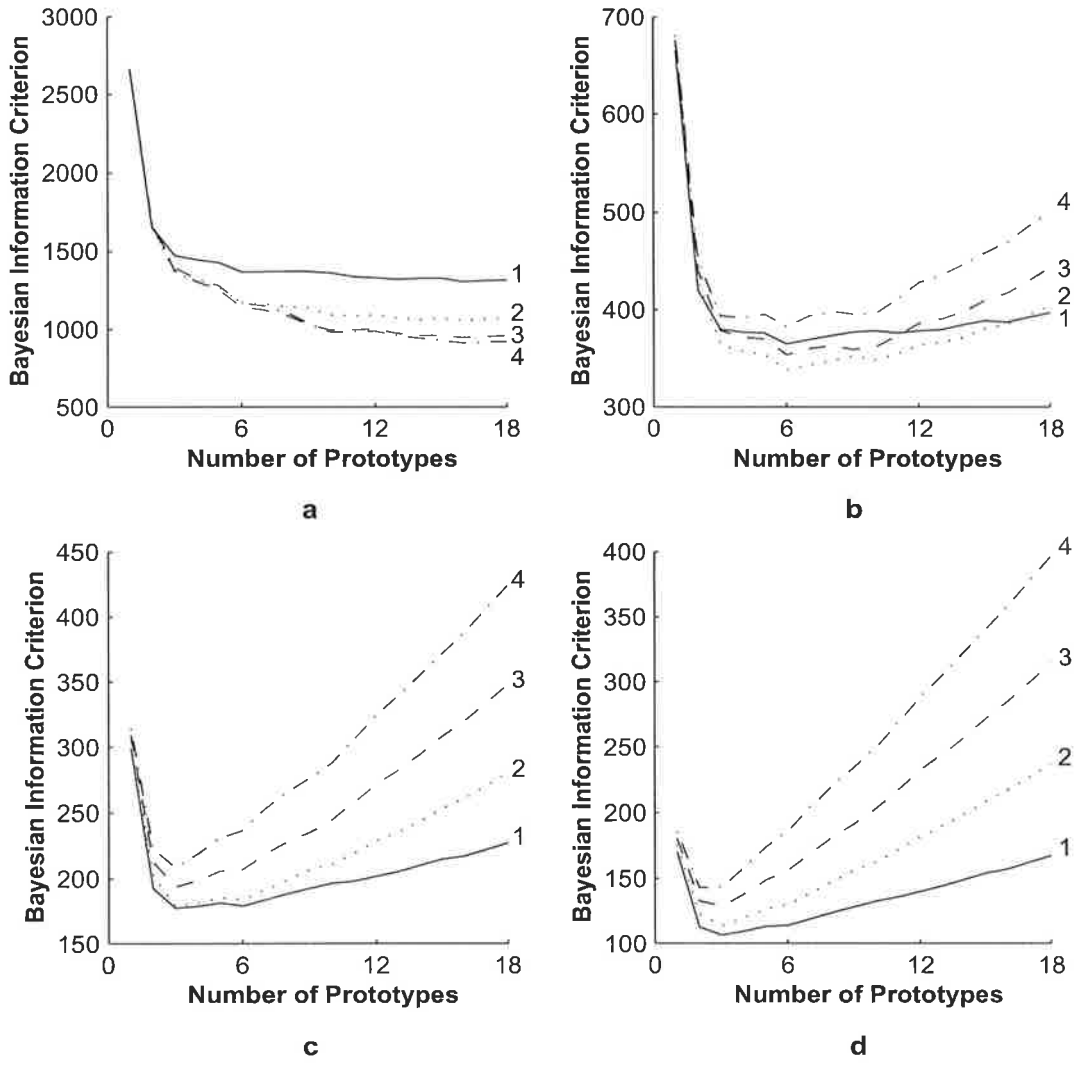

Figure 5.5: BIC values for prototype space representations of Johnson and Tversky's (1984) risk data, using modelling precision s of 0.05 (panel a), 0.10 (panel b), 0.15 (panel c) and 0.20 (panel d). The number of prototypes m ranges from 1 to 18, and the number of spatial dimensions k ranges from 1 to 4. The number of dimensions is marked adjacent to the plots.

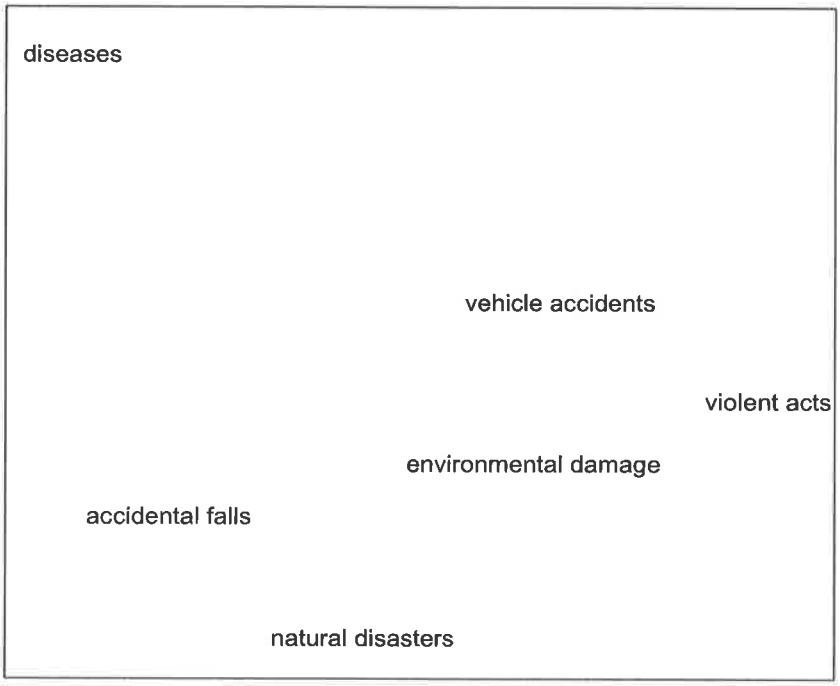

Figure 5.6: The six prototype, two-dimensional representation preferred by the BIC at a modelling precision of $s = 0.10$, explaining 57.9% of the variance.

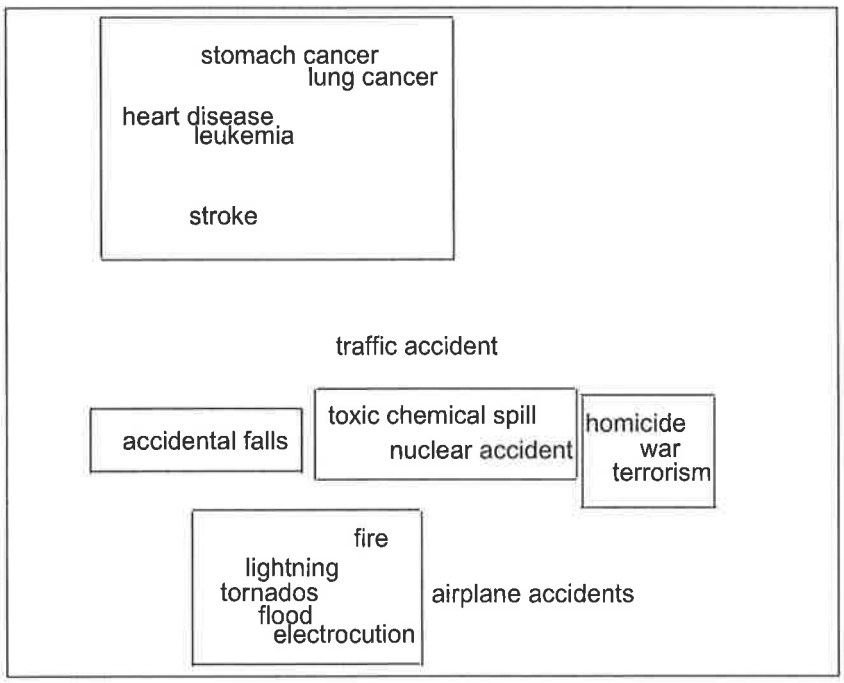

Figure 5.7: The two dimensional stimulus space representation of Johnson and Tversky's (1984) risk data, explaining 66.1% of the variance. Five of the six categories in Figure 5.6 form bounded and connected regions in this space.

5.5.2 Animals

The second data set examined was the unpublished data collected by O'Doherty and Lee (2002; see Section 2.1 for details) on the similarity of 21 animals ($\hat{\sigma} = 0.18$). The data were analysed using CAPS, with the prototype space ranging from a null model ($m = 1$) to a full stimulus space ($m = 25$), and the number of dimensions ranging from 1 to 4 using the Euclidean distance metric. The VAF and BIC values (using $s - \sigma$) for the prototype scaling algorithms are shown in Figure 5.8. The preferred representation has one dimension and four categories, though the five category representation explaining 57.9% of the variance is shown in Figure 5.9. This representation contains classes that might be labelled "mammals", "birds", "other flying things", "reptiles" and "wet things" (the four category representation is obtained by merging "reptiles" with "other flying things"). The one dimensional stimulus space which explains 60.5% of the variance produces the following stimulus dimension: Koala (0.17), Chimpanzee (0.18), Elephant (0.20), Camel (0.20), Cow (0.20), Zebra (0.20), Horse (0.20), Lion (0.20), Cat (0.20), Dog (0.20), Chicken (0.29), Eagle (0.32), Bat (0.35), Dragon (0.36), Snake (0.37), Scorpion (0.40), Butterfly (0.41), Bee (0.42), Frog (0.44), Shark (0.46) and then Goldfish (0.47). Therefore the categorical information extracted in this representation groups neighbouring items in the stimulus space into the same category.

5.5.3 Plants, Animals and Colours

The third data set was reported by Cooke et al. (1986, see Section 2.1) and consists of a set of 25 concepts. Once again, data were analysed using CAPS, with the prototype space ranging from a null model ($m = 1$) to a full stimulus space ($m = 25$), and the number of dimensions ranging from 1 to 4 using the Euclidean distance metric. The Variance Accounted For by these representations is shown in Figure 5.10. The three dimensional model with $m = 25$ explaining 76% of the variance is indistinguishable from the MDS

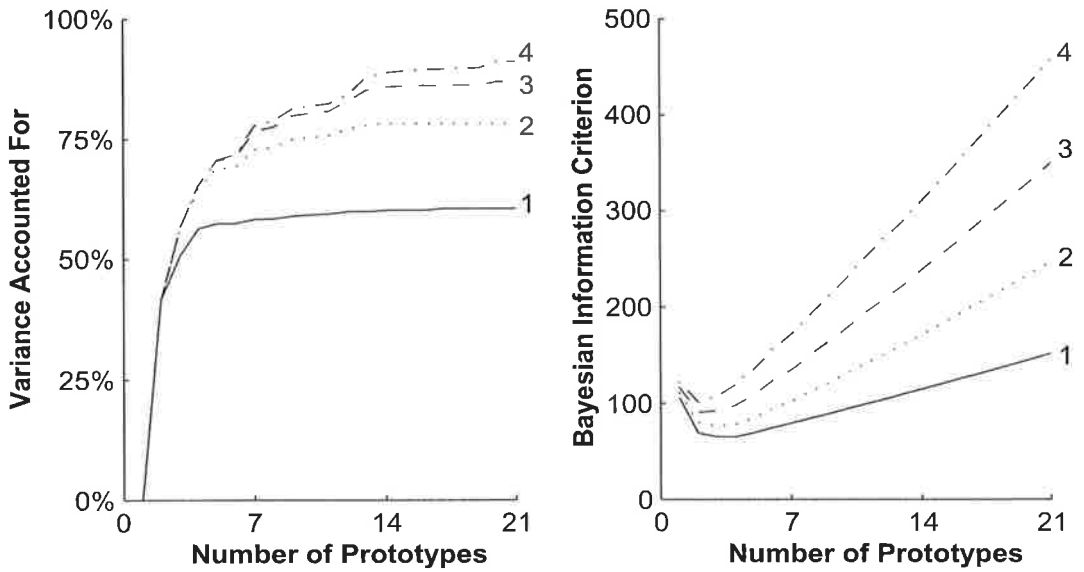

Figure 5.8: VAF (left panel) and BIC (right panel) values for prototype space representations of O'Doherty and Lee's animal data, using modelling precision $s - \sigma = 0.18$. The number of prototypes m ranges from 1 to 21, and the number of spatial dimensions k ranges from 1 to 4. The number of dimensions is marked adjacent to the plots.

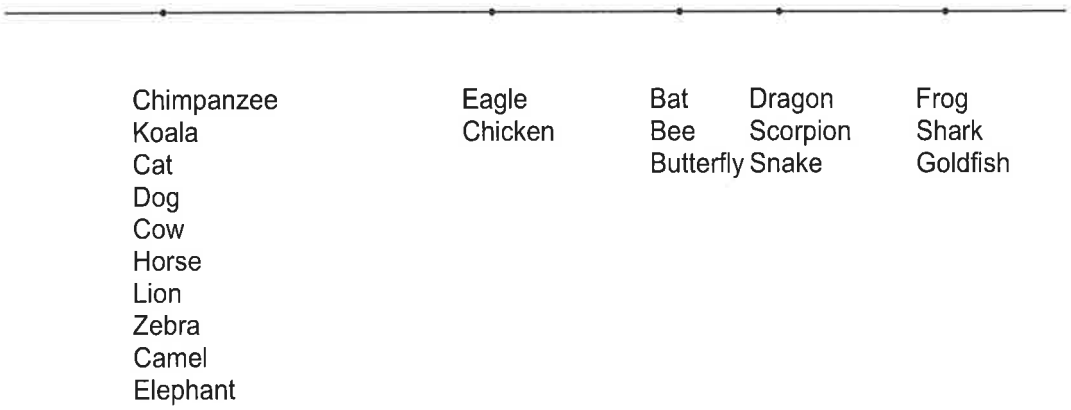

Figure 5.9: The 5-category one dimensional representation of O'Doherty and Lee's animal data, explaining 57.5% of the variance.

representation reported by Cooke et al. (1986). Figure 5.11 displays the BIC values for each of the CAPS-derived representations, assuming s-values of 0.05, 0.10, 0.15 and 0.20. A precision value of 0.05 corresponds to a high modelling precision, and the preferred representation is a full stimulus space with at least four dimensions. At this level of precision prototype scaling reduces to MDS. At more moderate levels of precision ($s = 0.10$ and $s = 0.15$), category-level information emerges, with these precision levels favouring two-dimensional spaces with 19 and 6 categories respectively, explaining 61.1% and 45.4% of the variance. The six prototype representation is displayed in Figure 5.12. For comparative purposes, the two-dimensional stimulus space for this data set is displayed in Figure 5.13. Once again, it is evident that five of the six categories form bounded and connected regions in this space.

Category labels for the representation shown in Figure 5.12 might be assigned as follows: category 1 is "mammals", category 2 is "red things", category 3 is "birds", category 4 is "things specific to deer", category 5 is the exemplar "frog" and category 6 is "plants". However, a number of somewhat arbitrary category assignments seem to have been made. For example, "colour" and "green" have been placed in the same category as the plants, but "red" has not. It seems likely that the requirement that categories be mutually exclusive is unwarranted for this data set. This suggests that another approach should be used to represent these data; for instance, a featural approach such as those discussed in Chapter 4.

5.6 Summary & General Discussion

The research presented in this chapter is preliminary in nature. The psychological basis of prototype spaces was discussed, and a procedure for deriving these spaces was tested. The application of these ideas to three data sets indicates that the approach has merits, though remains largely undeveloped. Future work in this area might con-

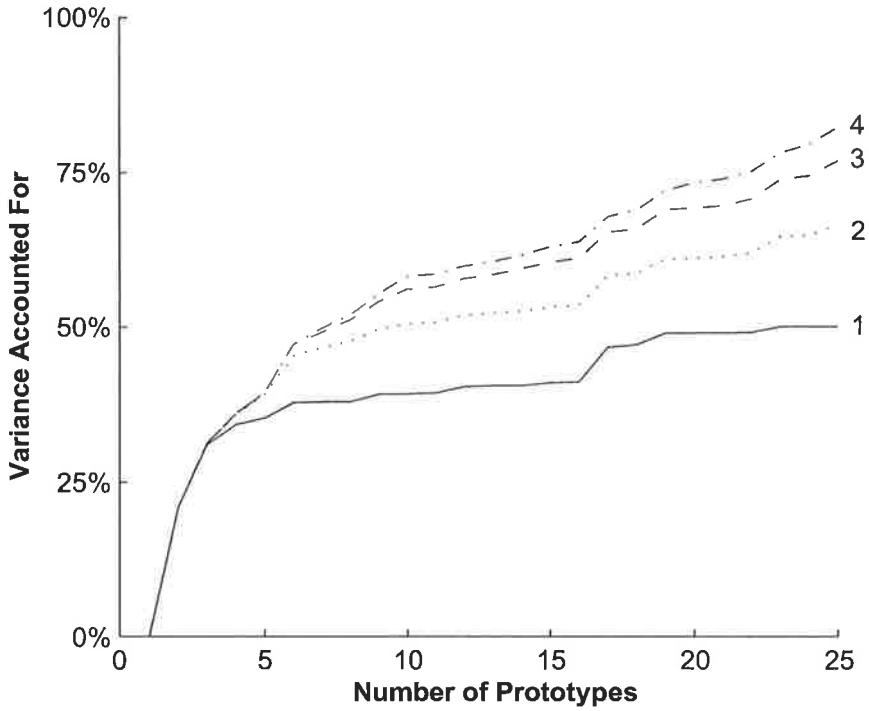

Figure 5.10: Variance Accounted For by prototype space representations of Cooke et al.'s (1986) concept data. The number of prototypes m ranges from 1 to 25, and the number of spatial dimensions k ranges from 1 to 4. The number of dimensions is marked adjacent to the plots.

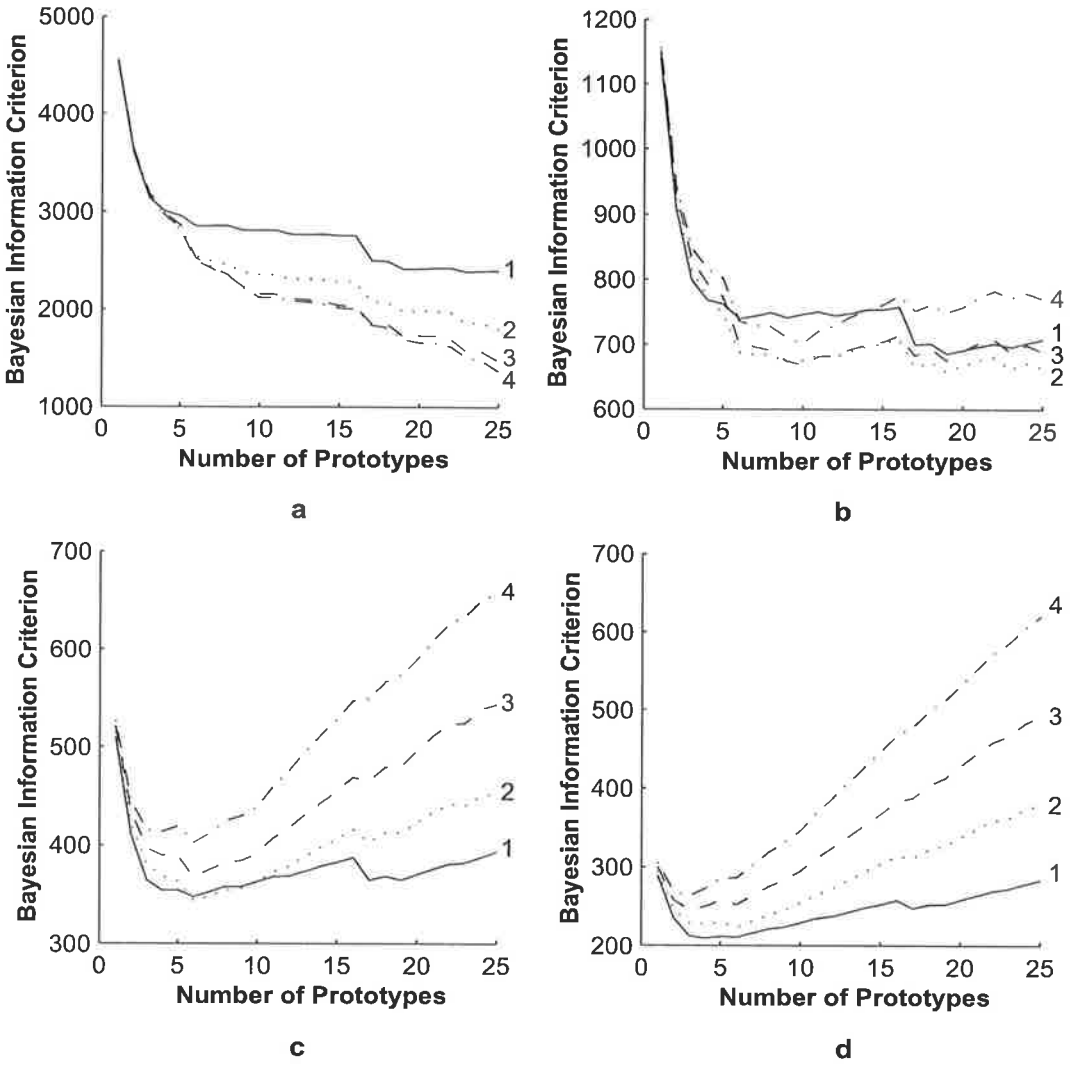

Figure 5.11: BIC values for prototype space representations of Cooke et al.'s (1986) concept data, using modelling precision s of 0.05 (panel a), 0.10 (panel b), 0.15 (panel c) and 0.20 (panel d). The number of prototypes m ranges from 1 to 25, and the number of spatial dimensions k ranges from 1 to 4. The number of dimensions is marked adjacent to the plots.

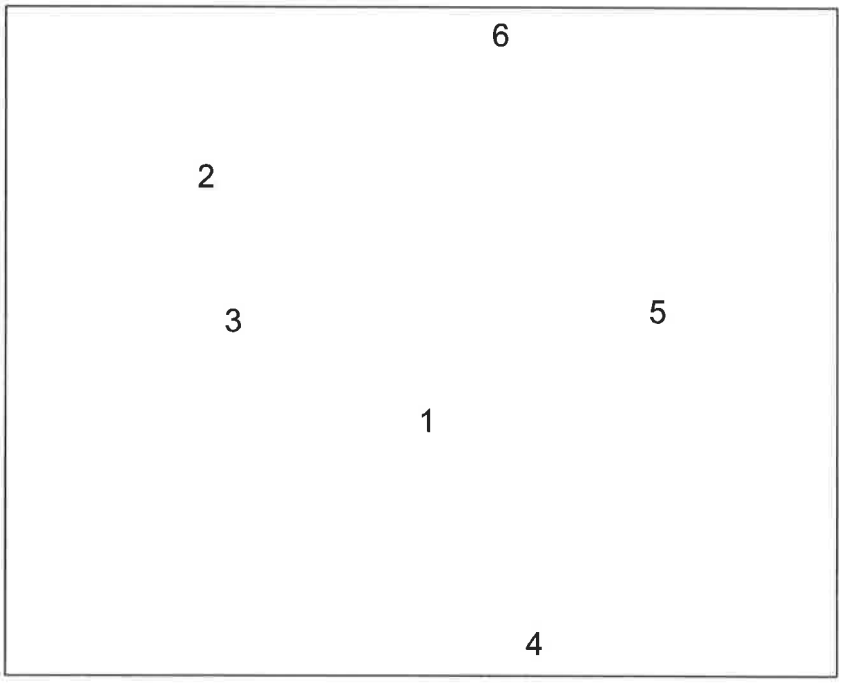

Figure 5.12: The six prototype, two-dimensional representation preferred by the BIC at a modelling precision of $s = 0.15$, explaining 45.4% of the variance. The prototypes plotted consist of the following stimuli: *Category 1* consists of "living thing", "animal", "mammal", "hairs", "dog", "deer" and "bats"; *Category 2* consists of "blood" and "red"; *Category 3* consists of "bird", "feathers", "robin" and "chicken"; *Category 4* consists of "antlers" and "hooves"; *Category 5* consists of "frog"; and *Category 6* consists of "plant", "leaves", "tree", "cottonwood", "flower", "rose", "daisy", "colour" and "green".

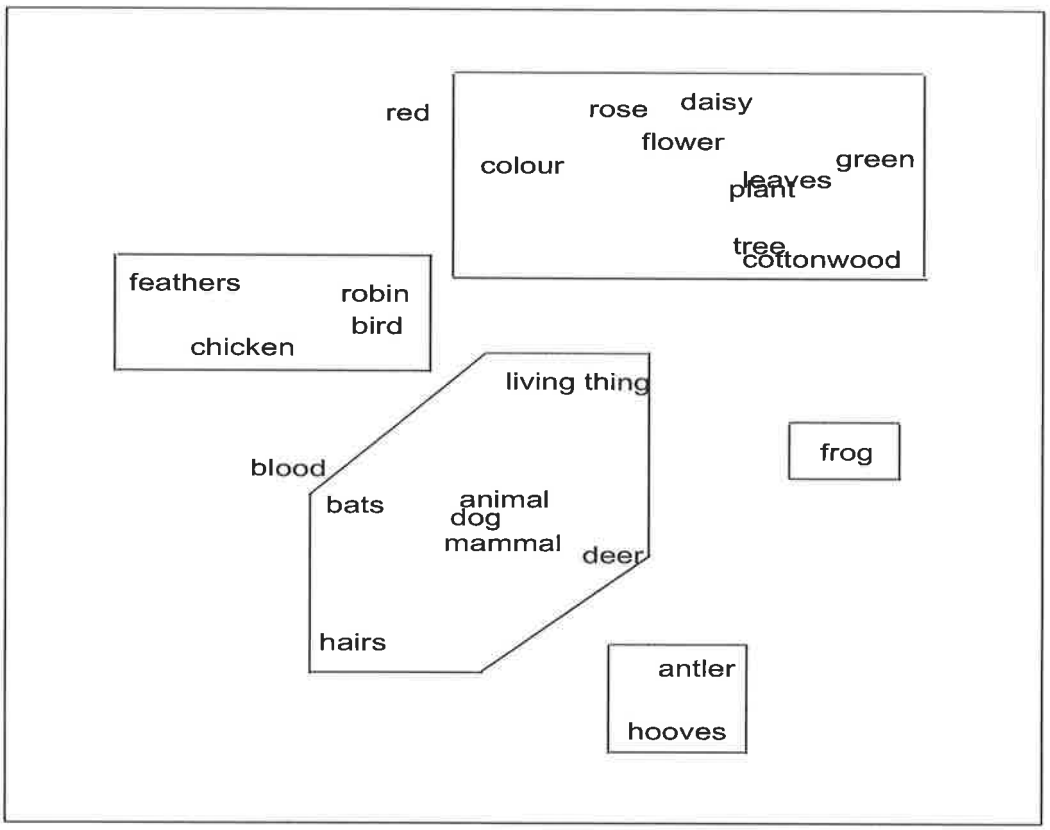

Figure 5.13: The two-dimensional stimulus space (MDS) representation of the domain, explaining 66.6% of the variance. Five of the six categories displayed in Figure 5.12 can be characterised as connected, bounded regions in the space, and are depicted by boxes enclosing a group of stimuli.

sider distributing the representation of stimuli over several prototypes. Inasmuch as concepts such as 'green' and 'tree' do not belong in the same space, a single prototype space could be replaced with multiple category spaces. Alternatively, the prototype approach could be combined with other representational theories. For example, once the stimulus set has been represented in a prototype space, the individual categories could be represented as trees. Psychologically, such an approach would treat categories as hierarchically-organised structures, but assumes that the relationship between categories is better characterised by a "proximity" relationship.

6. Similarity as a Decision Process

Most of the similarity models discussed in this thesis share a common philosophy. They assume that the decision process involved in making a similarity judgement is sufficiently simple that it makes little or no contribution to the data (though see Goldstone, 1994). Correspondingly, similarity judgements are assumed to reflect a response based primarily on some underlying informational structure: the representation. This has proven to be a highly successful research strategy, and an argument can be made to the effect that well-founded representational models of similarity afford an understanding of the principles that govern human similarity judgements (see Tenenbaum & Griffiths, 2002b, for a similar argument regarding generalisation). Without in any way disregarding the strengths of this approach, it is equally interesting (even complementary) to examine the decision processes involved, and consider the manner in which human cognition approximates rational ideals. This chapter outlines just such an approach, though the work is purely theoretical, and is restricted to considering candidates for a similarity decision model.

The models developed in this chapter were motivated by an examination of the task involved in the countries experiment (Section 4.9), in which participants chose one response from a fixed set of alternatives. Therefore, they are most naturally suited to an experimental methodology of this kind, where there is a limited set of discrete responses, known to the participant in advance. Furthermore, the decision process is assumed to operate on a set of discrete environmental cues. In this respect, these decision models

can be regarded as extensions of the featural approach to similarity modelling. However, there is no reason in principle why the models cannot be extended to account for other sorts of tasks, or to other representational structures, though this is not done here.

6.1 Heuristic Decision Models

Gigerenzer and Todd (1999) introduce a framework for modelling human decision making that draws heavily on the ideas of Egon Brunswik and Herbert Simon. Brunswik (1943), like Gibson (1979) after him, argued that an understanding of human decision making required an understanding of the structure of real world environments (though unlike Gibson, Brunswik did not limit investigation solely to environmental structure). People, he observed, are sensitive to inter-correlations between environmental cues, and exploit a range of cues to make accurate decisions. In a similar vein, Simon (1956, 1972) assumes that naturalistic decision making involves the exploitation of the non-arbitrary structure of the environment, though his "bounded rationality" approach incorporates the information processing limitations of the organism when making decisions.

In developing their approach, Gigerenzer and Todd (1999) argue that using small numbers of environmental cues *is* rational. If real environments possess a correlated cue structure, then the return on investment for examining large numbers of cues is minimal. They therefore assume that the limitations on human information processing are a rational, evolved adaptation to real environments, enabling people to make fast and accurate decisions using a minimum of information. Therefore, they claim that people adopt simple heuristics to make decisions, and their research programme investigates these heuristics and the kinds of environmental structures in which such heuristic decision making is rational. Indeed, despite the speed and simplicity of the models, there is evidence to suggest that, in a range of real environments, heuristic models can be just as accurate as decision rules that use all available information (Czerlinski, Gigerenzer,

& Goldstein, 1999).

The combination of speed, simplicity and accuracy is a formidable one, and suggests the idea of a fast and frugal similarity model. Such a model needs to describe how people use a set of features sparingly to make quick and accurate decisions. Therefore the two elements of a heuristic similarity model are a limited information search and a decision rule to end the search.

6.1.1 Feature Sampling Processes

One of the principles introduced by Gigerenzer and Todd is that real-world decision making involves a *search* for information, either through memory or through external sources. Very rarely does the stimulus itself provide the information required. For example, in the countries experiment, participants had to search through memory to recall information about the various nations. Even when all the information required is present in the stimulus, as often occurs when using artificial stimuli, an argument can be mounted to the effect that people *examine* that information serially when making decisions about the stimulus. Therefore, this process is most naturally characterised as a search through the feature set **F** one feature at a time, until a decision can be made.

Leaving the decision rule aside momentarily, consider the three candidates for this search process proposed by Gigerenzer and Goldstein (1996, 1999): *The Minimalist* (MIN), *Take The Last* (TTL), and *Take The Best* (TTB). These heuristics were designed to model two-choice decision tasks, like trying to pick the larger of two cities. They assume that people examine environmental cues one by one until some cue allows a decision to be made. The MIN heuristic is the simplest, in which cues are drawn randomly without replacement from the feature set. The TTL procedure starts with the cue that allowed the most recent decision to be made, then the cue that made the decision before that (unless it is the same cue), and so on. Cues that have never made a

decision are drawn last, in a random order. TTB assumes that there exist some *validities* that provide information on which are the better cues, and draws the cues in order of decreasing validity (breaking ties randomly).

These strategies can be imported without modification into a similarity model, by substituting the word "feature" for the word "cue". If no cue validity information exists, then the MIN or TTL strategies can be employed. However, since TTB relies on cue validities, some notion of feature validity is required. The most plausible candidate for the role of cue validity are the saliency weights w, since a highly salient feature has a larger effect on similarity than a less salient one.

6.1.2 Decision Procedures

The decision procedure for a binary-choice task such as those discussed by Gigerenzer and Goldstein (1999) is fairly simple. Cues are considered to have some evidence value, in that they may support one, both or neither of the two possible answers. Therefore, if a cue provides evidence for only one of the answers, then that answer is given and the search process is terminated. To state this more concretely, imagine asking an American whether Melbourne is bigger than Brisbane. They may have only a sketchy knowledge of Australian cities, and employ a TTB heuristic in the following manner. The first cue they think of might be "Is it Sydney?", since Sydney is Australia's largest city. However, although this cue provides positive evidence in favour of Sydney, it does not help with Melbourne or Brisbane. However, the next cue might be "Has it hosted an Olympics?", which has a positive value for Melbourne but not Brisbane. The search stops at this point and the person answers "Melbourne". Making decisions in this manner works because things like hosting the Olympic Games correlate fairly well with city size. In contrast, imagine the same person was asked whether Whyalla is bigger than Mount

Gambier[1]. It is quite likely that no cue that they recall would enable a rational decision, and a random answer would be expected.

The same decision rule applies to similarity decisions. Imagine instead that the (now somewhat harassed) American were asked whether they thought Sydney is more like Melbourne than Berlin is like St Petersburg. Again there are two responses, though the question now requires a similarity decision. It is again plausible to assume that a heuristic search such as TTB is employed. The theoretically important question involved in modelling the decision rule is to state when a feature provides evidence favouring a particular response. Motivated by featural theories of similarity, three psychologically plausible decision rules are suggested, called the *Common Features* (CF) rule, the *Distinctive Features* (DF) rule, and the *Contrast Model* (CM) rule. The CF-rule states that if two cities share the feature, then the feature provides similarity evidence in favour of that response: if only one city has that feature, or neither feature does, then no evidence is provided. In contrast, the DF-rule says that a feature provides evidence if both cities have the same value on the feature (if they both have it or both do not), but not if only one city has the feature. The CM-rule assumes (like the Modified Contrast Model proposed in Chapter 4) that some features are commonalities and others are distinctions. Therefore, the CM-rule generalises the CF and DF approaches by stating that the CF-rule should be applied to commonalities, and the DF-rule to distinctions.

These decision rules become slightly more complicated when more than two response options are available. It is highly unlikely that there exists a single feature that provides evidence in favour of only one response, particularly when there are a large number of potential responses. One possible solution to this difficulty is suggested by the Categorisation By Elimination model (Berretty, Todd, & Blythe, 1997; Berretty, Todd, & Martignon, 1999), as well as Tversky's (1972) Elimination By Aspects and the fast

[1]The two largest regional centers in South Australia. Both have around 25,000 people.

and frugal E-mail prioritisation algorithm proposed by Lee, Chandrasena, and Navarro (in press). Suppose that a feature has been drawn using some search procedure (be it TTB, TTL, or MIN). If the feature provides no evidence for any of the options, then the cue is discarded and a new one is drawn. If it provides evidence for one or more of the response options, then all other options are eliminated, and a new cue is drawn. This process continues until only one response option remains, at which point the search terminates and the last remaining option is chosen as the response. If there are still multiple response options left once all the cues have been drawn, one of the remaining options is chosen randomly.

In order to provide a concrete example of how this multi-choice similarity decision is assumed to operate, consider the task involved in the countries experiment. Suppose a participant were asked to select the two most similar nations from a list consisting of Spain, Italy, Germany, and Zimbabwe. Furthermore, suppose that the cues available to them are those present in the common features representation displayed in Figure 4.17. Given this, it makes sense to adopt the TTB search heuristic, and employ the CF decision rule. The first cue drawn would be the southern Africa feature (Nigeria, Zimbabwe), which does not provide evidence for any of the response options. However, the second cue drawn would be the western Europe feature (Italy, Germany, Spain), which provides evidence for three of possible responses 'Italy & Germany', 'Italy & Spain', and 'Germany & Spain'. The other three possibilities (those involving Zimbabwe) are eliminated, and the search continues. Inspection of Figure 4.17 shows that none of the other features will provide evidence for any of these three over the other, so one of these would be chosen at random. Therefore, this *TTB-CF heuristic* predicts that these three options are equally similar, which happens to be the same prediction made by the common features model.

Formalising these heuristics is simple enough. Suppose there are x possible re-

sponses, denoted $\mathbf{r} = \{r_1, r_2, \ldots, r_x\}$, where r_p denotes the response corresponding to the pth pair of stimuli. If there are m features belonging to the feature matrix $\mathbf{F}$, the $m \times x$ *evidence matrix* $\mathbf{E} = [e_{kp}]$ can be specified, such that e_{kp} is 1 whenever the kth feature provides evidence for the pth stimulus pair, -1 whenever it provides evidence against that pair, and 0 if it does neither. So, if the pth stimulus pair consists of the ith and jth stimuli, the CF-rule states that

$$e_{kp} = \begin{cases} 1 & \text{if } f_{ik} = f_{jk} = 1 \\ 0 & \text{otherwise} \end{cases}$$

whereas the DF-rule states that

$$e_{kp} = \begin{cases} -1 & \text{if } f_{ik} = f_{jk} \\ 0 & \text{otherwise} \end{cases}$$

If the CM-rule is applied, then those rows of $\mathbf{E}$ that correspond to common features are specified by the CF-rule, whereas the distinctive features rows will be described by the DF-rule. Thus, using some search heuristic (e.g. TTB), one examines the rows of $\mathbf{E}$ one by one, eliminating response options with lower evidence values, until only one option remains.

6.1.3 Non-Compensatory Environments

In the previous example the TTB-CF heuristic made the same prediction as the common features model. It can be seen that these two models will make the same predictions when saliency weights are *non-compensatory*. If the weights are ordered such that $w_1 > w_2 > \ldots > w_m$, then they are non-compensatory if $w_k > \sum_{i=k+1}^{m} w_i$. Correspondingly, if the cues are examined in order of decreasing saliency, there is never any point to continue examining cues after a good one is found: even if all remaining cues suggested another response, their combined saliency would still be less than that of the first good cue. Martignon and Hoffrage (1999) make precisely this argument with respect to Gigerenzer and Goldstein's (1999) original formulation.

Nevertheless, although feature saliencies vary in the real world, it is fair to say that *strictly* non-compensatory environments are not universal, so these heuristic models are certainly distinguishable from the representational models discussed elsewhere. For example, the feature weights displayed in Figure 4.17 are not strictly non-compensatory. If the task were to choose the most similar nations from Cuba, Jamaica, China, and Japan, the heuristic and representational models make different predictions. Cuba and Jamaica share one feature, which has a weight of 0.505, whereas China and Japan share two features with weights of 0.262 and 0.371. No other pair shares any feature. The common features model predicts a choice of China-Japan, since $0.262 + 0.371 > 0.505$, whereas a TTB-CF heuristic would predict a choice of Cuba-Jamaica, since the feature they share would be drawn first. The MIN heuristic would choose China-Japan two times out of three, and Cuba-Jamaica the other third. Note that this feature set was derived according to the common features model, so it is unfair to draw any conclusions from this comparison. The important observation in this regard is that heuristic models and representational models can be empirically distinguished.

6.1.4 Summary

Using Gigerenzer and Todd's (1999) heuristics framework as a guideline, a number of possibilities exist for fast and frugal featural similarity models. These models involve two elements: a search heuristic and a decision rule. Plausible candidates for the search heuristic are the TTB, TTL, and MIN approaches. Suggested decision rules employ the same approach as Categorisation By Elimination, where the evidence provided by any particular feature is given by the CF-rule, the DF-rule, or the CM-rule. In a non-compensatory environment, the TTB heuristic is indistinguishable from the corresponding representational model, although no such guarantee exists for TTL or MIN. In any event, there are certainly environments in which the saliency weights are not

non-compensatory.

6.2 Sequential Sampling Models

The heuristic decision models presented in the previous section rely on the notion of "one good reason". As soon as *any* evidence is found favouring one alternative over another, the unfavoured alternative is immediately discarded. If the environment is strictly non-compensatory and the search heuristic is TTB, the first evidence discovered will always be sufficient to support a rational decision. However, saliency weights are not always (or even often) strictly non-compensatory. Furthermore, there is an argument to be made suggesting that at different times people will use more or less evidence to make decisions. Therefore, inspired by models of simple decision making, this section develops the idea of a similarity decision model that integrates evidence from multiple sources, but need not examine every available feature.

The similarity models proposed in this section are based on the Sequential Sampling Models (SSMs) of simple decision tasks (see Luce, 1986 for an overview). These theories were originally proposed to model simple two-choice tasks. For instance, an experiment might present participants with two lines A and B, and ask them to select the longer of the two lines. The central idea is that when people examine the lines, they collect a sample of observations about the difference in length $A - B$. SSMs assume that stimulus representations are noisy, so there is some variability in the sample of $A - B$ observations. Some models assume that the sampling is discrete, and others assume that the sample evolves in continuous time. Since similarity evidence is considered to be provided by a discrete sample of cue validities, this discussion will consider only discrete time SSMs.

The two main classes of discrete-time SSMs are random walk models and accumulator models. Random walk models (e.g., Laming, 1968; Wald, 1947) keep a single

signed tally: every time an $A - B$ difference is observed, the evidence (given by the size of the $A - B$ discrepancy) is added to the total. So if $A > B$ for the current observation, the magnitude of the discrepancy (that is, $|A - B|$) is added to the tally, but if $B > A$, then $|A - B|$ is subtracted. Once the tally reaches a certain threshold value, a decision is made. The major assumption made by random walk models is that evidence for A is exactly the same as evidence against B. However, as Vickers and Lee (1998) observe, the difficulty with this is that it is not easy to extend random walks to multi-choice scenarios, and it is rarely tried.

Accumulator models maintain separate unsigned totals, one counting the amount of evidence favouring $A > B$ and the other counting the evidence favouring $B > A$. In the recruitment, or simple accumulator model developed by La Berge (1962), the totals only count the *number* of observations favouring each alternative. That is, every time A is observed to be longer than B, the $A > B$ tally increases by 1. However, in the generalised accumulator model proposed by Vickers (1979), the magnitudes of the differences $|A - B|$ are added to the tallies. The appeal of accumulator models in this context is the ease with which they may be extended to the kinds of multi-choice decisions that are involved in similarity judgements[2]. Therefore the models developed here adopt an accumulator approach rather than a random walk approach.

6.2.1 Recruitment: A Multi-Cue Minimalist

Suppose that no information about the individual saliency is known. If so, a recruitment model has an intuitive appeal, since every observation is assumed to add the same amount of evidence. The recruitment model is a natural extension of the MIN heuristic, in the following sense. The MIN heuristic draws cues at random from an underlying

[2]Historically, the other major strength of the generalised accumulator model has been the ability to account for the confidence with which people make decisions (e.g., Vickers, 1979; Vickers & Packer, 1982; Vickers, Smith, Burt, & Brown, 1985). However, confidence is not traditionally measured in similarity tasks.

distribution of cues with unknown evidence values, and decides in favour of the first response option to have a single positive cue value. The recruitment model samples from an unknown distribution of $A - B$ values, and accumulates evidence in the form of the number of observations favouring one response or the other. Therefore, the recruitment model is the natural multi-cue version of the MIN heuristic. The *similarity recruitment model* maintains a counter for each response option that initially starts at zero. Like the MIN heuristic, the recruitment model samples features at random without replacement. Whenever a feature is drawn, the tallies for each response option that have positive evidence values on that cue are incremented by 1. At any given moment, the counters reflect the number of features that have provided evidence in favour of a particular response. The decision process terminates when one of the counters reaches a pre-specified threshold. If two or more response options reach the threshold simultaneously, then one might either raise the threshold by one and draw a new cue (a bit like reaching deuce in tennis), or eliminate the options below the threshold, and then draw a new cue to in order to eliminate more cues (i.e., revert to the Categorisation By Elimination mechanism). If the cues run out before a clear winner appears, one could respond randomly, though this seems undesirable. An alternative approach would be to "top up" the counters until one reaches the threshold, either by resampling cues, or by adding increments to all counters.

6.2.2 The Accumulator: Taking The Best Few

When saliency weights are known, it makes sense to assume that each feature provides an amount of evidence in proportion to its saliency, and to draw the cues in order of decreasing validity, in the manner of TTB. However, if the saliency weights are not strictly non-compensatory, or in the extreme case uniformly distributed, it is poor policy to look at only one feature. The natural solution to this problem is to recast TTB as a

similarity accumulator model. As with the recruitment model, a counter is maintained for each option, initially set to zero. However, when a feature provides evidence for a response option, its saliency is added to the total. Thus, more salient cues are weighted more heavily in the decision process. Once again, as soon as one tally exceeds the threshold value the search process terminates and a decision is made. Ties are less likely to occur under an accumulator than under a recruitment model, but if they do the same options are available for breaking them.

Varying the response threshold allows the accumulator model to interpolate smoothly between TTB and a representational theory of similarity such as the common features model. If the threshold is set arbitrarily low, then any evidence at all will send a tally over the threshold, and the model reduces to TTB. If the threshold is set arbitrarily high, then the threshold will never be reached. Since some response must be made, the sensible strategy is to select the response closest to the threshold, so the model reduces to the corresponding representational model (e.g., common features, distinctive features, etc.). Importantly, intermediate thresholds also yield principled models. For example, a moderate threshold might indicate that one very salient feature is sufficient evidence for a decision, but otherwise it will take several lesser features. There is nothing inappropriate about saying that one very good reason is sufficient to make a decision, but in its absence, many less compelling reasons may be required.

6.2.3 Summary

One reason for applying SSMs to similarity decisions is that they have proven to be remarkably successful models of simple decision tasks. The similarity accumulator in particular has a compelling theoretical appeal as a natural generalisation of heuristic models and representational models. The ability to integrate multiple sources of evidence is important in a decision model, as is the ability to terminate the search for evidence

in a principled manner. Furthermore, the SSM approach to modelling similarity makes predictions regarding response times and confidence. It is for these reasons that the similarity accumulator represents an interesting direction for future research.

7. Epilogue

Similarity has been an extensively researched topic in psychology, with a wide range of similarity theories proposed over the years. The six frameworks identified in this thesis – spatial representations, tree structures, featural representations, network models, alignment models, and transformational models – have been developed to different degrees. For instance, spatial representations have a 50 year history, have been applied in a wide range of situations, and can be derived using many different procedures, whereas transformational representation is barely more than a sketch of an idea backed by a few experiments.

The research presented in this thesis addresses a number of issues related to similarity modelling. Broad issues regarding theories of similarity were discussed in Chapters 1 and 2. The important issue of evaluating a representation in terms of data-fit, model complexity, and theoretical interpretability was discussed in Chapter 3, and an approach developed for applying these ideas to similarity modelling. Chapters 4 and 5 examined the featural and spatial approaches, though in somewhat different regards. Several featural models, including the new Modified Contrast Model (MCM) were evaluated at some length in Chapter 4. In contrast, Chapter 5 makes a more modest contribution, introducing the notion of representing prototypical information in a multidimensional space. Finally, in Chapter 6, an approach is outlined for modelling the decision process underlying similarity judgements. In this last chapter, a brief survey is made of the state of the field, and some suggestions are made about directions for future research.

7.1 Similarity Theories: Representations and Decisions

Most theories of similarity are *representational theories*, inasmuch as they specify informational structures (e.g., features, dimensions, etc.) that shape similarity and generalisation. Accordingly, representations can be taken to reflect the principles that underlie similarity judgements, irrespective of how the judgement was made. Representational models therefore permit an explanation of similarity judgements in these principled terms, and shed light on the basic knowledge structures by which people understand their environment.

Each of the representational theories has different strengths and weaknesses. It has frequently been argued that spatial representations are most appropriate for low-level perceptual stimuli, whereas featural representations are better suited to high-level conceptual domains (e.g., Carroll, 1976; Tenenbaum, 1996; Tversky, 1977). Nevertheless, real-world environments will frequently involve perceptual and conceptual elements. Therefore, there is some merit to the idea of combining continuous and discrete elements into a single representation. As Carroll (1976, p. 462) argues: "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". This could be achieved by allowing for "general representations" consisting of spatial dimensions and discrete features. Although preliminary investigations of this kind of similarity model suggests that the idea has some promise (Navarro & Lee, submitteda), it remains largely unexplored territory.

The alternative approach, based on Gigerenzer and Todd's (1999) notion of fast, heuristic decision making, is to examine the cognitive processes by which people use an informational structure to make a similarity judgement, and therefore develop decision models of similarity. It is clear that the two approaches are complementary: representa-

tional models implicitly assume that some decision process is in operation, and decision models rely on underlying representations. The historical focus in similarity modelling has been the representational approach. Inspired by models of analogical reasoning, Goldstone's (1994) SIAM model represents a step towards a developing decision models in this field. The sequential sampling models discussed in Chapter 6 may also be interpreted in this light. Importantly, the sequential sampling models provide a theoretically interpretable framework for interpolating between heuristic decision models and representational similarity models.

7.2 Similarity Modelling and Geometric Complexity

Pinker (1998) has argued that "[p]inning down mental representations is the route to rigor in psychology" (p. 85). Cognitive process models frequently rely on stimulus representations in order to account for observed phenomena. With this in mind, it is important that these informational structures are themselves plausible accounts of mental representations. A cognitive model that employs hand-specified representations is a bit like a house built on quicksand. No matter how soundly built it is, a house is useless if the foundations are sinking.

If the study of similarity affords psychologists the opportunity to specify mental representations in a principled manner, it is crucial that the analysis of similarity data be based on sound principles of scientific inference. As argued in Chapter 3 and by many other authors (e.g., Collyer, 1985; Myung, 2000; Pitt et al., 2002; Roberts & Pashler, 2000), it is bad practice to assume that a theory is a good one simply because it provides a good fit to the data. Choosing between similarity models should be based on quantitative properties such as data-fit and model complexity, and qualitative properties such as theoretical interpretability and psychological plausibility.

The Geometric Complexity Criterion (GCC) employed in this thesis is perhaps the

state of the art for quantitative model selection. By measuring model complexity in terms of the proportion of distinguishable probability distributions indexed by the model that lie close to the true distribution, the GCC enables a principled trade-off between fit and complexity. GCC expressions were derived in Chapters 3 and 4 for several classes of featural models, as well as additive trees. The additive clustering and additive tree frameworks were analysed in some detail, in order to provide an understanding of how the model parameters interact to make a representation more or less complex.

This approach can be extended in fairly obvious ways. Though attempts to find a GCC expression for spatial representations have thus far been unsuccessful, it may not be impossible. Even if analytic methods fail, GCC values can be approximated using numerical methods. Such an approach would be too computationally expensive to employ in a multidimensional scaling context, but numerical evaluations of various spatial representations may enable conclusions to be drawn about the complexity of spatial models generally. Furthermore, it is natural to look for GCC expressions for other representations, such as network or alignment models. In particular, it would be interesting to compare the highly structured SIAM model (Goldstone, 1994) to other approcches by using the GCC. Similarly, if the sequential sampling models proposed in Chapter 6 can be shown to give a good account of empirical data, they could be examined using the GCC as well.

7.3 On Using Representations

As outlined in Chapter 1, similarity-based representations are employed by models of identification, recognition, and categorisation. When employing representations in this regard, it is important to observe that derived representations reflect only the information relevant to the experimental task. For example the feature "cricket playing nation" could only emerge in the countries experiment if more than one such nation (Zimbabwe) were

included in the domain. Correspondingly, it is important to design methodologies that elicit the appropriate information from the data.

Note that a "soccer playing nation" feature did not emerge either, despite the fact that eight of the sixteen nations reached the finals of the 2002 World Cup (Italy, Germany, Spain, Russia, United States, China, Japan, and Nigeria). This feature may not have been sufficiently salient to make a large contribution to participants' decisions, particularly since it correlates strongly with other, more salient features. For the purposes to which the countries data were put, this is not a problem. If, however, one subsequently used this representation in a model of categorisation required to classify nations as "soccer nations" or "non-soccer nations", caution is required. If participants are explicitly asked to classify in this regard, then the saliency of any soccer-related knowledge will rise very sharply. Correspondingly, since similarity always involves some context (e.g., Goldstone et al., 1997; Goodman, 1972), it may not be appropriate to apply a representation derived from a soccer-neutral similarity context to a soccer-related classification task.

Similarity modelling serves a dual purpose, in that representations can be employed in models of other cognitive processes, but are also informative in their own right. A theory of similarity makes a number of assumptions about the nature of mental representation, and so every test of the theory should inform psychologists about the structure of human knowledge (even if only a little). By way of example, consider the extensive analysis of the TCM and MCM presented in Chapter 4. Both of these theories treat a feature as some aspect of the world to which people are sensitive. Furthermore, if a particular feature emerges in the representation of some data set, then that feature is something people use when making decisions: it is a *reason* to make a decision. The TCM treats a feature as an abstract grouping of stimuli, and permits a feature to denote a commonality or a distinction as the task demands. However, this reliance on task demands implies that features cannot be treated differentially, since there is nothing intrinsic to the feature that

dictates how it relates to the environment. In contrast, the MCM assumes that features are intrinsically commonalities or distinctions. The feature itself dictates how it should be used: commonalities only justify a decision when two stimuli share the property, and distinctions justify a decision when two stimuli differ in that regard. The superior performance of the MCM suggests that the *type* of regularity embodied by a feature is a central aspect of representational structures. Therefore, even though neither the MCM nor the TCM can lay claim to being a complete account of similarity, by proposing and testing these theories, it is possible to learn something about the nature of conceptual structure.

7.4 Similarity and the Blue Sky of Cognition

Similarity may be uniquely justifiable as a tool for understanding basic mental structure, in that the sense that "this thing is like that one" is central to cognition. Even an act as "simple" as inferring that "this thing is a chair" is fundamentally reliant on similarity. No two chairs that a person encounters will ever be precisely identical, so in order to identify and categorise a novel object, some generalisation (or "slippage", to use Hofstadter's 2000 term) from the encountered object is required. Therefore, despite the observed differences between the various chairs a person encounters, something about them is identified as being "the same". Given the ubiquity and necessity of this process, the words of William James are appropriate: "This sense of Sameness is the very keel and backbone of our thinking" (James, 1890, p. 459).

Why should this be? Intelligent behaviour relies on the ability to determine when to treat two things alike, despite observable differences between them. Since it is crucial for any organism to discover the regularities that govern its environment, it is logical to assume that the internalisation of that environment should map the physical world in a manner that supports appropriate generalisation from one stimulus to the next (Shepard

1981, 1984, 1987, 1994). Therefore, two stimuli *should* appear more alike to an organism if they are more likely to entail the same consequences. It is therefore appropriate to assume that internal representations reflect such psychophysical mappings, and provide insight into basic cognitive architecture.

The study of similarity is necessarily complicated by the flexible manner in which people perceive likenesses. People use different information to draw different analogies, finding a variety of similarities depending on context. It seems likely that people *build* a representation appropriate to the task at hand, pull it apart as needed, pack new things into it, and discard irrelevant information. Current theories of similarity cannot account for this process, but they do not claim to. Rather, they make a more modest claim, providing an account of the "shape" of the representations that people use to make simple decisions. If a theory of similarity does a good job, it can tell us something about how information is structured, and provide some insight into how things are organised in the mind.

References

Abdi, H., Barthelemy, J. P., & Luong, X. (1984). Tree representations of associative structures in semantic and episodic memory research. In D. E. & J. van Buggenhaut (Eds.), *Trends in Mathematical Psychology* (p. 3-31). Amsterdam: North-Holland.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory.* Budapest: Akademiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute, 50,* 277-290.

Arabie, P. (1994). Pathfinding in quicksand. *Contemporary Psychology, 39*(1), 101.

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

Armstrong, M. A. (1988). *Groups and Symmetry.* New York, NY: Springer-Verlag.

Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology, 63,* 546-554.

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

Balasubramanian, V. (1999). A geometric formulation of Occam's razor for inference of parametric distributions. *Unpublished manuscript.*

Barsalou, L. W. (1989). Intraconcept similarity and its implications for interconcept similarity. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (p. 76-121). Cambridge: Cambridge University Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioural and Brain Sciences*, *22*, 577-660.

Barthelemy, J. P., & Guenoche, A. (1991). *Tree Models of Proximity.* New York: Wiley.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370-419.

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review*, *75*, 127-142.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, *4*, 91-99.

Bellman, R. (1970). *Introduction to Matrix Analysis* (2nd ed.). New York: McGraw-Hill.

Berretty, P. M., Todd, P. M., & Blythe, P. W. (1997). Categorization by elimination: A fast and frugal approach to categorization. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (p. 43-48). Mahwah, NJ: Lawrence Erlbaum.

Berretty, P. M., Todd, P. M., & Martignon, L. (1999). Categorization by elimination: Using few cues to choose. In G. Gigerenzer & P. M. Todd (Eds.), *Simple Heuristics That Make Us Smart* (p. 235-254). Oxford: Oxford University Press.

Borg, I., & Lingoes, J. (1987). *Multidimensional Similarity Structure Analysis.* New York: Springer-Verlag.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information theory. *Journal of Mathematical Psychology, 44*, 62-91.

Brewer, W. F. (1989). The activation and acquisition of knowledge. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 532–545). Cambridge: Cambridge University Press.

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

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review, 50*(3), 255-272.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In F. R. Hodson, D. G. Kendall, & P. Tautu (Eds.), *Mathematics in the Archaeological and Historical Sciences* (p. 387-395). Edinburgh, UK: Edinburgh University Press.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review, 58*, 413-423.

Calvin, W. H. (1996). *How Brains Think: Evolving Intelligence, Then and Now.* London: Wiedenfeld and Nicolson.

Carlton, E. H., & Shepard, R. N. (1990a). Psychologically simple motions as geodesic paths: I. Asymmetric objects. *Journal of Mathematical Psychology, 34*, 127-188.

Carlton, E. H., & Shepard, R. N. (1990b). Psychologically simple motions as geodesic paths: II. Symmetric objects. *Journal of Mathematical Psychology, 34*, 189-228.

Carroll, J. D. (1976). Spatial, non-spatial and hybrid models for scaling. *Psychometrika, 41*(4), 439-463.

Carroll, J. D., & Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology, 31*, 607-649.

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

Carroll, J. D., & Pruzansky, S. (1975). Fitting of hierarchical tree structure (HTS) models, mixtures of HTS models, and hybrid models, via mathematical programming and alternating least squares. In *U.S.-Japan seminar in theory, methods, and applications of multidimensional scaling and related techniques.* San Diego, CA.

Carroll, J. D., & Pruzansky, S. (1980). Discrete and hybrid scaling models. In E. D. Lantermann & H. Feger (Eds.), *Similarity and Choice* (p. 108-139). Bern: Hans Huber.

Clouse, D. S., & Cottrell, G. W. (1996). Discrete multi-dimensional scaling. *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, 290-294.

Collyer, C. E. (1985). Comparing strong and weak models by fitting them to computer-generated data. *Perception and Psychophysics, 38*(5), 476-481.

Cooke, N. M., Durso, F. T., & Schvaneveldt, R. W. (1986). Recall and measures of memory organization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 12*(4), 538-549.

Corter, J., & Tversky, A. (1986). Extended similarity trees. *Psychometrika, 51*, 429-451.

Corter, J. E. (1982). ADDTREE/P: A PASCAL program for fitting additive trees based on Sattath and Tversky's ADDTREE algorithm. *Behavior Research Methods and Instrumentation, 14*, 353-354.

Corter, J. E. (1996). *Tree Models of Similarity and Association.* Thousand Oaks, CA: Sage.

Cox, T. F., & Cox, M. A. A. (1991). Multidimensional scaling on a sphere. *Communications in Statistics: Theory and Methods, 20*(9), 2943-2953.

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

Cunningham, J. P. (1978). Free trees and bidirectional trees as representations of psychological distance. *Journal of Mathematical Psychology, 17*, 165-188.

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In G. Gigerenzer & P. M. Todd (Eds.), *Simple Heuristics That Make Us Smart* (p. 97-118). Oxford: Oxford University Press.

D'Andrade, R. (1978). U-statistic hierarchical clustering. *Psychometrika, 4*, 58-67.

Davison, M. L. (1983). *Multidimensional Scaling.* New York: Wiley.

de Bruijn, N. G. (1958). *Asymptotic Methods in Analysis.* Amsterdam: North-Holland.

de Soete, G. (1983). A least squares algorithm for fitting additive trees to proximity data. *Psychometrika, 48*, 621-626.

Dodwell, P. C. (1983). The Lie transformation group model of visual perception. *Perception and Psychophysics, 34*(1), 1-16.

Eisler, H., & Ekman, G. (1959). A mechanism of subjective similarity. *Acta Psychologica*, *16*, 1-10.

Ekman, G. (1954). Dimensions of color vision. *The Journal of Psychology*, *38*, 467-474.

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port & T. Van Gelder (Eds.), *Mind as Motion: Explorations in the Dynamics of Cognition* (p. 195-225). Cambridge, MA: Bradford.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*, 1-63.

Feger, H., & Bien, W. (1982). Network unfolding. *Social Networks*, *4*, 257-283.

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology*, *41*, 145-170.

Findley, D. (1991). Counterexamples to parsimony and BIC. *Annals of the Institute of Statistical Mathematics*, *43*(3), 505-514.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3-71.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology.* Cambridge, MA: MIT Press.

Forbus, K. D., & Oblinger, D. (1990). Making SME greedy and pragmatic. *Proceedings of the12th Annual Conference of the Cognitive Science Society*, 61-68.

Freyd, J. J. (1987). Dynamic mental representations. *Psychological Review*, *94*(4), 427-438.

Garner, W. R. (1974). *The Processing of Information and Structure.* Potomac, MD: Erlbaum.

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

Gati, I., & Tversky, A. (1987). Recall of common and distinctive features of verbal and pictorial stimuli. *Memory and Cognition, 15*(2), 97-100.

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

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155-170.

Getty, D. J., Swets, J. B., Swets, J. A., & Green, D. M. (1979). On the prediction of confusion matrices from similarity judgements. *Perception and Psychophysics, 26,* 1-19.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception.* Boston, MA: Houghton Mifflin.

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

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In G. Gigerenzer & P. M. Todd (Eds.), *Simple Heuristics That Make Us Smart* (p. 75-95). Oxford: Oxford University Press.

Gigerenzer, G., & Todd, P. M. (Eds.). (1999). *Simple Heuristics That Make Us Smart.* Oxford: Oxford University Press.

Goldman, A. J. (1966). Realizing the distance matrix of a graph. *Journal of Research of the National Bureau of Standards - B. Mathematics and Mathematical Physics, 70B*(2), 153-154.

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 3-28.

Goldstone, R. L. (1998). Hanging together: A connectionist model of similarity. In J. Grainger & A. M. Jacobs (Eds.), *Localist Connectionist Approaches to Human Cognition* (p. 283-325). Mahwah, NJ: Lawrence Erlbaum.

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. (1994). The time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 29-50.

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.

Gordon, A. D. (1999). *Classification* (2nd ed.). Boca Raton, FL: Chapman and Hall.

Greenacre, M. J., & Underhill, L. G. (1982). Scaling a data matrix in a low dimensional Euclidean space. In D. M. Hawkins (Ed.), *Topics in Applied Multivariate Analysis* (p. 183-268). Cambridge: Cambridge University Press.

Gregson, R. A. M. (1975). *Psychometrics of Similarity.* New York: Academic Press.

Grunwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology, 44*, 133-152.

Hahn, U., & Chater, N. (1997). Concepts and similarity. In K. Lamberts & D. Shanks (Eds.), *Knowledge, Concepts and Categories* (p. 43-92). Cambridge, MA: MIT Press.

Hakimi, S. L., & Yau, S. S. (1965). Distance matrix of a graph and its realizability. *Quarterly of Applied Mathematics, 22*, 305-317.

Harary, F. (1964). A graph theoretic approach to similarity relations. *Psychometrika, 29*(2), 143-151.

Hartigan, J. A. (1975). *Clustering Algorithms.* New York, NY: Wiley.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea.* Cambridge, MA: MIT Press.

Hendrickx, M., & Wagemanns, J. (1999). A critique of Leyton's theory of perception and cognition. *Journal of Mathematical Psychology, 43*, 314-345.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations* (p. 75-109). Cambridge, MA: MIT Press.

Hoffman, W. C. (1966). The Lie algebra of visual perception. *Journal of Mathematical Psychology, 3*, 65-98.

Hoffman, W. C. (1968). The neuron as a Lie group germ and a Lie product. *Quarterly of Applied Mathematics, 25*, 423-441.

Hoffman, W. C. (1970). Higher visual perception as prolongation of the basic Lie transformation group. *Mathematical Biosciences, 6*, 437-471.

Hoffman, W. C. (1984). Figural synthesis by vectorfields: Geometric neuropsychology. In P. C. Dodwell & T. Caelli (Eds.), *Figural Synthesis* (p. 249-282). Hillsdale, NJ: Lawrence Erlbaum.

Hoffman, W. C., & Dodwell, P. C. (1985). Geometric psychology generates the visual gestalt. *Canadian Journal of Psychology*, *39*(4), 491-528.

Hofstadter, D. R. (1985). *Metamagical Themas: Questing For the Essence of Mind and Pattern.* New York, NY: Erlbaum.

Hofstadter, D. R. (1995). *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought.* London: Penguin Books.

Hofstadter, D. R. (2000). Analogy as the Core of Cognition. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The Analogical Mind: Perspectives From Cognitive Science* (p. 499-538). Cambridge, MA: Bradford.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint mapping. *Cognitive Science*, *13*, 295-355.

Huba, G. L., Wingard, J. A., & Bentler, P. M. (1981). A comparison of two latent variable causal models for adolescent drug use. *Journal of Personality and Social Psychology*, *40*(1), 180-193.

Hubert, L., Arabie, P., & Meulman, J. (1997). Linear and circular unidimensional scaling for symmetrical proximity matrices. *British Journal of Mathematical and Statistical Psychology*, *50*, 253-284.

Hutchinson, J. W. (1989). NETSCAL: A network scaling algorithm for nonsymmetric proximity data. *Psychometrika*, *54*(1), 25-51.

Imai, S. (1977). Pattern similarity and cognitive transformations. *Acta Psychologica, 41*, 433-447.

Imai, S. (1992). Fundamentals of cognitive judgements of patterns. In H. Geissler, S. W. Link, & J. T. Townsend (Eds.), *Cognition, Information Processing and Psychophysics: Basic Issues* (p. 225-266). Hillsdale, NJ: Lawrence Erlbaum.

James, W. (1890). *The Principles of Psychology.* New York: Dover.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophical Society, 31*, 203-222.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford: Oxford University Press.

Johnson, E. J., & Tversky, A. (1984). Representations of perceptions of risk. *Journal of Experimental Psychology: General, 113*(1), 55-70.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241-254.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (1995). *Essentials of Neural Science and Behavior.* Stamford, CT: Appleton and Lange.

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

Keren, G. (1990). On the intricacies involved in the study of similarity judgments: Comment on Ritov, Gati, and Tversky. *Journal of Experimental Psychology: General, 119*(1), 42-43.

Klauer, K. C. (1988). Representing similarities by ordinal networks. In H. H. Bock (Ed.), *Classification and Related Methods of Data Analysis* (p. 473-477). Amsterdam: Elsevier.

Klauer, K. C. (1989). Ordinal network scaling: Representing proximities by graphs. *Psychometrika*, *54*(4), 737-750.

Klauer, K. C., & Carroll, J. D. (1989). A mathematical programming approach to fitting general graphs. *Journal of Classification*, *6*, 247-270.

Klauer, K. C., & Carroll, J. D. (1991). A comparison of two approaches to fitting directed graphs to nonsymmetric proximity measures. *Journal of Classification*, *8*, 251-268.

Kolb, B., & Whishaw, I. Q. (1996). *Fundamentals of Human Neuropsychology (4th edition).* New York: W. H. Freeman.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, *1*(1), 1-7.

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

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement: Additive and Polynomial Representations, vol 1.* New York, NY: Academic Press.

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

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

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115-129.

Kullback, S. (1968). *Information Theory and Statistics.* New York: Dover.

La Berge, D. (1962). A recruitment theory of simple behaviour. *Psychometrika, 27,* 375-396.

Laming, D. R. J. (1968). *Information Theory of Choice-Reaction Times.* London: Academic Press.

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

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

Lee, M. D. (1999). *Algorithms for Representing Similarity Data* (Technical Report No. DSTO-TR-0152). DSTO, Electronics and Surveillance Laboratory.

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

Lee, M. D. (2001b). 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.*

Lee, M. D. (submitteda). Algorithms for fitting and displaying additive trees with limited complexity. *Manuscript submitted for publication.*

Lee, M. D. (submittedb). A Bayesian analysis of retention functions. *Manuscript submitted for publication.*

Lee, M. D., Chandrasena, L., & Navarro, D. J. (in press). Using cognitive decision models to proritize E-mails. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*.

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

Lewin, K. (1936). *Principles of Topological Psychology.* New York, NY: McGraw-Hill.

Leyton, M. (1985). Generative systems of analyzers. *Computer Vision, Graphics and Image Processing*, *31*, 201-241.

Leyton, M. (1986a). A theory of information structure: I. General principles. *Journal of Mathematical Psychology*, *30*, 102-160.

Leyton, M. (1986b). A theory of information structure: II. A theory of perceptual organization. *Journal of Mathematical Psychology*, *30*, 257-305.

Leyton, M. (1989). Inferring causal history from shape. *Cognitive Science*, *13*, 357-387.

Leyton, M. (1992). *Symmetry, Causality, Mind.* Cambridge, MA: MIT Press.

Lindman, H., & Caelli, T. (1978). Constant curvature Riemannian scaling. *Journal of Mathematical Psychology*, *17*, 89-109.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology* (Vol. 1, p. 103-190). New York: Wiley.

Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization.* New York, NY: Oxford University Press.

Markman, A. B., & Dietrich, E. (2000a). Extending the classical view of representation. *Trends in Cognitive Sciences*, *4*, 470-475.

Markman, A. B., & Dietrich, E. (2000b). In defense of representation. *Cognitive Psychology*, *40*, 138-171.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431-467.

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work? In G. Gigerenzer & P. M. Todd (Eds.), *Simple Heuristics That Make Us Smart* (p. 119-140). Oxford: Oxford University Press.

Massaro, D. W., Cohen, M. M., Campbell, C. S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin and Review*, *8*, 1-17.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, *97*, 225-252.

McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models.* Cambridge, MA: MIT Press.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, *44*, 1469-1481.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (p. 179-195). New York: Cambridge University Press.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89-115.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100*(3), 398-407.

Minsky, M. (1986). *The Society of Mind.* New York: Simon and Schuster.

More, J. J. (1977). The Levenberg-Marquardt algorithm: Implementation and theory. In G. A. Watson (Ed.), *Lecture Notes in Mathematics* (Vol. 630, p. 105-116). New York: Springer-Verlag.

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.

Myung, I. J., & Pitt, M. A. (in press). Model evaluation, testing and selection. In K. Lambert & R. L. Goldstone (Eds.), *Handbook of Cognition.* Sage Publication.

Myung, I. J., & Shepard, R. N. (1996). Maximum entropy inference and stimulus generalization. *Journal of Mathematical Psychology, 40*, 342-347.

Navarro, D. J. (submitted). Regarding the complexity of additive clustering models: Comment on Lee (2001). *Manuscript submitted for publication.*

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.

Navarro, D. J., & Lee, M. D. (in press). Commonalities and distinctions in featural stimulus representations. *Proceedings of the 24th Annual Conference of the Cognitive Science Society.*

Navarro, D. J., & Lee, M. D. (submitteda). Combining dimensions and features in similarity-based representations. *Manuscript submitted for publication.*

Navarro, D. J., & Lee, M. D. (submittedb). Common and distinctive features in stimulus representation: A modified version of the contrast model. *Manuscript submitted for publication.*

Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*, 135-183.

Newell, A. (1990). *Unified Theories of Cognition.* Cambridge, MA: Harvard University Press.

Nocedal, J., & Wright, S. J. (1999). *Numerical Optimization.* New York: Springer-Verlag.

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

Nosofsky, R. M. (1992a). Exemplars, prototypes, and similarity rules. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *Essays in Honor of William K. Estes* (p. 149-167). Hillsdale, NJ: Erlbaum.

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

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 282-304.

O'Doherty, K. C., & Lee, M. D. (2002). The featural representation of animals based on similarity. *Australian Journal of Psychology*(54(1)), 60.

Orth, B. (1988). Representing similarities by distance graphs: Monotonic network analysis (MONA). In H. H. Bock (Ed.), *Classification and Related Methods of Data Analysis* (p. 489-494). Amsterdam: Elsevier.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185-200.

Pinker, S. (1994). *The Language Instinct.* London: Penguin.

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

Pitt, M. A., Kim, W., & Myung, I. J. (in press). Flexibility versus generalizability in model selection. *Psychonomic Bulletin and Review*.

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.

Powell, M. J. D. (1977). Restart methods for the conjugate-gradient method. *Mathematical Programming*, *12*, 241-254.

Pruzansky, S., Tversky, A., & Carroll, J. D. (1982). Spatial versus tree representations of proximity data. *Psychometrika*, *47*, 3-24.

Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science.* Cambridge, MA: MIT Press.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological Methodology* (p. 111-196). Oxford, U. K.: Blackwells.

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

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

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, *30*(4), 629-636.

Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics*, *14*(3), 1080-1100.

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.

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*, *14*.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192-233.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization* (p. 27-77). Hillsdale, NJ: Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573-605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations* (p. 151-193). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. E. (Eds.). (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations* (Vol. 1). Cambridge, MA: MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models* (p. 7-57). Cambridge, MA: MIT Press.

Russell, S. J. (1986). A quantitative analysis of analogy by similarity. In *Proceedings of the National Conference on Artificial Intelligence* (p. 284-288). Philadelphia, PA: AAAI.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika, 42*, 319-345.

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

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

Shannon, B. (1988). On the similarity of features. *New Ideas in Psychology*, *6*(3), 307-321.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325-345.

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125-140.

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27*(3), 219-246.

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

Shepard, R. N. (1981). Psychophysical complementarity. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual Organization* (p. 279-341). Hillsdale, NJ: Erlbaum.

Shepard, R. N. (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thnking and dreaming. *Psychological Review*, *91*(4), 417-447.

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* (p. 53-71). Washington, DC: American Psychological Association.

Shepard, R. N. (1994). Perceptual-cognitive universal as reflections of the world. *Psychonomic Bulletin and Review*, *1*(1), 2-28.

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

Shepard, R. N., & Kannappan, S. (1991). Connectionist implementation of a theory of generalization. *Advances in Neural Information Processing Systems*, *3*, 665-671.

Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive Psychology*, *7*, 82-138.

Shiina, K. (1988). A fuzzy-set-theoretic feature model and its application to asymmetric similarity data analysis. *Japanese Psychological Research*, *30*(3), 95-104.

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129-138.

Simon, H. (1972). Theories of bounded rationality. In C. B. Radner & R. Radner (Eds.), *Decision and Organization* (p. 161-176). Amsterdam: North-Holland.

Smith, E. E., & Medin, D. L. (1981). *Categories and Concepts.* Cambridge, MA: Harvard University Press.

Smith, L. (1989). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (p. 146-178). Cambridge: Cambridge University Press.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1-74.

Sneath, P. H. A. (1957). The application of computers to taxonomy. *The Journal of General Microbiology*, *17*, 201-206.

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy.* San Francisco: Freeman.

Solomonoff, R. J. (1964a). A formal theory of inductive inference. Part I. *Information and Control*, *7*, 1-22.

Solomonoff, R. J. (1964b). A formal theory of inductive inference. Part II. *Information and Control*, *7*, 224-254.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677-680.

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 8* (p. 3-9). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). The rational basis of representativeness. *23rd Annual Conference of the Cognitive Science Society.*

Tenenbaum, J. B., & Griffiths, T. L. (2002a). Generalisation, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 579-616.

Tenenbaum, J. B., & Griffiths, T. L. (2002b). Some specifics about generalization. *Behavioral and Brain Sciences*, *24*, 762-778.

Thelen, E. (1995). Time-scale dynamics and the development of an embodied cognition. In R. F. Port & T. Van Gelder (Eds.), *Mind as Motion: Explorations in the Dynamics of Cognition* (p. 69-100). Cambridge, MA: MIT Press.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association: Theory and Method, 81*, 83-86.

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

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review, 79*, 281-299.

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

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

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

Tversky, A., & Krantz, D. H. (1970). The dimensional representation of the metric structure of similarity data. *Journal of Mathematical Psychology, 7*, 572-596.

van Gelder, T., & Port, R. F. (1995). It's about time: An overview of the dynamical approach to cognition. In R. F. Port & T. van Gelder (Eds.), *Mind as Motion: Explorations in the Dynamics of Cognition* (p. 1-43). Cambridge, MA: MIT Press.

Vickers, D. (1979). *Decision Processes in Visual Processes.* New York, NY: Academic Press.

Vickers, D. (1996). An Erlanger programme for psychology. *Fourth International Social Science Methodology Conference.*

Vickers, D. (2002). *A generative transformational model of human visual perception* (Research Report). Australian Defence Science and Technology Organisation.

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgements: I. Properties of a self-regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences*, *2*(3), 169-193.

Vickers, D., & Packer, J. S. (1982). Effects of alternating set for speed or accuracy on response time, accuracy, and confidence in a unidimensional discrimination task. *Acta Psychologica*, *50*, 179-197.

Vickers, D., Smith, P. L., Burt, J., & Brown, M. (1985). Experimental paradigms emphasising state or process limitations: II. Effects on confidence. *Acta Psychologica*, *59*, 163-193.

Vosniadou, S., & Ortony, A. (1989). Similarity and analogical reasoning: A synthesis. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (p. 1-17). Cambridge: Cambridge University Press.

Wald, A. (1947). *Sequential Analysis.* New York: Wiley.

Weiner-Erhlich, W. K., Best, W. M., & Millwood, J. (1980). An analysis of generative representational systems. *Journal of Mathematical Psychology*, *21*, 219-246.

Wittgenstein, L. (1953). *Philosophical Investigations.* New York: Macmillan.

Young, G., & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, *3*, 19-22.