Mathematical Social Sciences 4 (1983) 25–71
North-Holland Publishing Company

MATHEMATICAL MODELS OF RECOGNITION AND CONFUSION IN PSYCHOLOGY

James T. TOWNSEND and Douglas E. LANDON
Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907, U.S.A.

Communicated by Maria Nowakowska
Received 25 January 1982
Revised 25 June 1982

A wide range of mathematical models of recognition and confusion in psychological experiments is examined. A taxonomy for the classification of these models is proposed. This taxonomy includes discriminant models, feature-confusion models, sophisticated guessing models and choice models. Both the within-class and between-class relationships of these models are examined so as to provide a framework for the application of these models to theories of human pattern recognition. Where appropriate, specific areas of application of these models in the context of human perceptual processing are presented and discussed.

Key words: Recognition; confusion; taxonomy; perception.

1. Introduction

The way in which people might go about recognizing or assigning a classification or name to each of a set of objects has long been of interest to investigators in many fields. In the broad view, pattern recognition plays an exceedingly strategic role in human survival and everyday functioning, from the time an infant comes to recognize his mother's breast to his complex information processing techniques gained through later education. Although much has been learned about the nature of human perception and information processing abilities (e.g., Cornsweet, 1970; Kaufman, 1974; Martindale, 1981), work has only begun in recent years in formulating mathematical models which can be tested statistically against laboratory data (e.g., Geyer and DeWald, 1973; Keren and Baggen, 1981; Townsend, 1971a, b; Townsend and Landon, 1982).

Heretofore, there has been no attempt to survey these models or to provide a scheme wherein they might be classified. The present paper offers an overview of the currently most important mathematical recognition models in psychology and a taxonomy that serves both to relate and differentiate them.

In the area of pattern recognition, as much as any other, there has been a good deal of interchange of ideas among the fields of artificial intelligence and those of psychology and physiology. With regard to psychology, many of the important seminal ideas were arrived at in their present form fairly 'early', for example in the

0165-4896/83/0000–0000/$03.00

1950's and 1960's (e.g., ideas about feature or template processing).[1] The flow of ideas from artificial intelligence approaches in pattern recognition seems to have slowed down in recent years, partly, we suspect, because the overall systems there have become increasingly complex and computationally oriented (see Batchelor, 1978). Some of the more complex approaches still intersect heavily with cognitive psychological work in higher processes, for instance knowledge representation and language learning. However, in the study of such behavior as letter recognition, it behooves psychologists to keep their models as parsimonious as possible, passing to more complicated formulations only when forced to by the data. For that reason the mathematical models below may seem somewhat simple compared with some of the current artificial intelligence approaches to pattern recognition. Nevertheless, certain instances of the models discussed below have done quite well in predicting experimental recognition and confusion data.

The models presented below in the first part of the paper tend to be the closest in structure to the decisional-recognition models associated with communication and signal detection theory of electrical engineering and related fields. These models also bear similarity to the versions of signal detection theory oriented around a two-response experimental situation in psychology (e.g., Green and Swets, 1966). However, while such models are conceptually important in the multiple signal, multiple response context (Green and Birdsall, 1978), it is often easier to build physical systems that operate according to the principles of these models than it is to test such models against behavioral data. Thus, that class of models has not been tested in a statistical manner against human experimental data as much as the models presented in the later portions of the paper, which were constructed with problems of empirical testability more directly in mind. We suspect that the former class will find increasing quantitative employment as psychologists become more adept in evolving testable versions.

Much of early experimental psychology was occupied with the sensation and perception of elementary stimuli and was most often focused upon establishing absolute and difference thresholds. A psychophysics grew up established around binary sensory decision situations, as well as the psychological scaling of unidimensional physical stimuli (see Baird and Noma (1978) for a discussion of classical psychophysics). Just as psychological scaling has been extended to more complicated multidimensional perceptual objects, so does the current paper reflect a growth of theory from the binary perceptual choice situation into multi-object recognition-confusion psychophysics.

We hope to give psychologists a coherant treatment of current quantitatively specified models of recognition and confusion, and those in other quantitatively oriented disciplines an overview of the present status and form of model[illegible] important area of human experimental psychology. Because of their re[illegible]

[1] Many such concepts have a conceptual history that takes them back into the nineteenth century or earlier (see, e.g., James (1890, 1950)).

general decisional and perceptual functioning, it seems likely that the use of such models will expand. Due to the uneveness and diversity of application of the models and because of space limitations, discussion of the actual empirical tests will have to be avoided. However, where possible, references to such work will be given. Reed (1973) and Vickers (1979) provide sources of background ideas on pattern recognition in psychology. Batchelor (1978), Mendel and Fu (1970) and Watanabe (1972) may be consulted for analogous concepts and applications in artificial intelligence. Dodwell (1970) and Harris (1980) examine aspects of pattern recognition from a human developmental and neurophysiological perspective, while Kolers and Eden (1968) attempt to relate biological, psychological and non-living pattern recognition systems to one another.

2. Basic concepts

The typical experiment for recognition (in human observers) usually makes use of some variant of the following format (e.g., Geyer and DeWald, 1973; Townsend, 1971a, b; Townsend and Landon, 1982). Examples are selected from a set of objects that are known beforehand to the observer. The observer is presented with one of the examples on each experimental trial and his or her ask is to give the representative name of that object. The presentation order of the selected objects is usually randomized, although it is common to constrain the presentations so that each example is presented a certain number of times per experimental session. Something about the experimental situation prevents perfect performance, or, more, generally, the observer is not able to pair a particular response with a particular stimulus object with certainty. Thus, confusions are, and the task of the investigator is to develop a theory (i.e., in this paper, a mathematical model) that can explain the pattern of confusions among the various objects as well as the frequencies with which the objects were correctly identified.

The basic black box problem for recognition is shown in Fig. 1; s_i is one of n_s possible stimuli, r_j is one of n_r possible responses and $P(r_j|s_i)$ represents the probability of response r_j given the presentation of stimulus s_i. A particularly common approach in psychology is shown in Fig. 2, where the recognition process is described as being a function of two basic influences, a sensory or perceptual effect

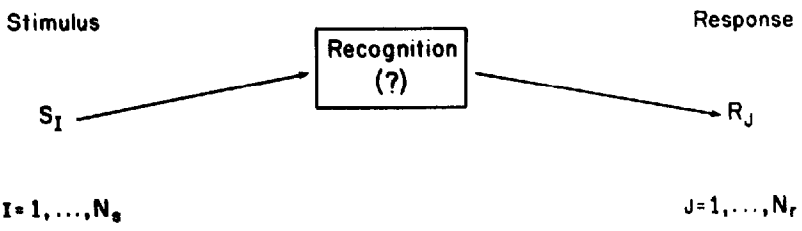

Fig. 1. The black box problem for psychology. How is recognition accomplished?

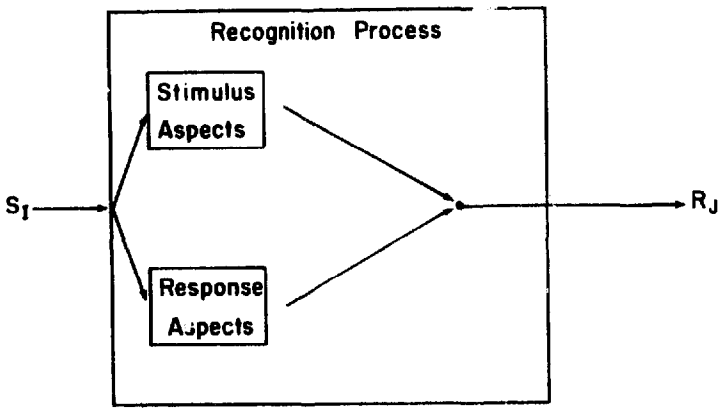

Fig. 2. The separation of stimulus and response aspects in recognition.

of the stimulus and an influence from aspects associated with the response. For instance, significant stimulus aspects might include intensity, frequency, contrast or interstimulus similarity whereas response effects might include rewards or payoffs, relative frequencies of presentations of the different stimuli (resulting in response preferences) or response similarity. Of course, any number of other dimensions of influence might be considered but much of the psychological literature writes the confusion functions in terms of the above two aspects of the recognition setting. In some of the models discussed below, the references to these separate aspects of the recognition situation are submerged or absent. In others, especially those employed to fit psychological data, they are explicit. The general recognition model (occupying the black box in Fig. 1) makes no assumptions except for explicitly presuming that a model *may* take distinct account of stimulus and response factors, as shown in Fig. 2. The particular configuration shown in Fig. 2 is not meant to imply parallel processing of the stimulus vs. response aspects.

The classic psychological explanations of the general black box recognition problem have usually fallen into one of two broadly defined areas: template matching or feature testing. Roughly, 'template matching' connotes the matching of a relatively untransformed or raw sensory image with a set of patterns stored in memory, the so-called templates. Each template represents one of the possible stimuli. A feature model, on the other hand, supposes that the stimulus patterns are made up of a set of more atomic aspects or 'features', and that an observer separately extracts or samples features from a presented stimulus pattern. The observer then matches the sampled feature set against the memory lists of features that constitute each potential stimulus. Clearly, these two kinds of models do not exhaust the universe of recognition models.

Fig. 3 lays out the family tree relating the models discussed in this paper. The first major decision of the scheme in Fig. 3 depends on the fact that the models emanating on the left start with a term (number, vector, etc.) representing an observation event on the part of the observer. In the models on the right, the observation event is either covert, implicit or simply irrelevant, depending on the particular model and its interpretation.

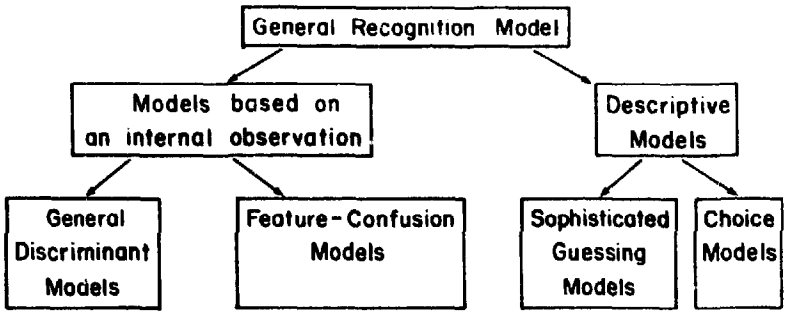

Fig. 3. A taxonomy for models of recognition.

The next division on the left breaks the 'observation' models into discriminant models and feature confusion models. The former supposes that the observation may undergo a transformation after which a distinct function, the 'discriminant', exists for each potential response possibility that maps the transformed observation into a magnitude representable by a real number. After this a selection mechanism picks as the response that alternative associated with the largest magnitude (discriminant). An example of a transformation on the stimulus might be the way in which the sensory apparatus alters the patterned information through transduction (change of the form of the energy, for example, from light to neural pulses) and receptor to brain neural projections. The set of discriminants might correspond to internal pattern matching or statistical functions among other possibilities.

It will be seen below that feature models may be of the discriminant variety but these must be kept logically separate from what we call the 'feature-confusion' models. The latter posit that, in general, there exists a confusion stage consisting of a subset of stimulus alternatives and that the observer responds from this reduced set of possibilities according to a well-defined probability distribution on the reduced set.

The sophisticated guessing models share with the feature confusion models the concept of a reduced confusion set of alternatives or 'candidate responses'. A conditional probability distribution is given leading from the stimulus alternatives to the confusion sets, but *without* reference to an actual observation or mechanism whereby the observer arrives at the confusion set.

Models of the choice variety, depicted on the right of Fig. 3, are based on the idea that the conditional probability of a certain response, given a particular stimulus, is the ratio of the strength of that response to the sum of the strengths of all possible responses.

Introduction of the choice model brings up the question of psychological process interpretations because the choice models arose largely out of measurement considerations (e.g., Luce, 1959, 1963). Other models of the various types also differ somewhat on this dimension, that is, with regard to their natural describability in terms of intuitive processing notions. For instance, the feature confusion models appear relatively high on the process 'scale'. A detailed

physiological model of the recognition process would likely score very high on such a scale. On the other hand until recently, the choice model does not possess viable process interpretation. Some potential solutions to this problem are discussed below.

Finally, it is sometimes possible in special cases or (sometimes fairly bizarre) interpretations to relate two presumably unrelated models as given in Fig. 3, but most of our discussion can proceed apace without reference to such anomalies.

It is extremely important to note at the outset that, in the interest of simplicity, the discussion will be confined to the case where the number of objects in the stimulus population equals the number of responses in the response population and there is a well-specified one-to-one correspondence of the stimuli with the responses. Thus the observer is aware that for each stimulus presentation there is exactly one correct response. Many of the results of this paper generalize immediately to the case where n_s stimuli must be categorized into n_r responses and $n_r < n_s$.[2] However, the question of response selection strategies and their optimality becomes somewhat more subtle.

3. Models based on an internal observation

It is supposed that presentation of a stimulus s_i evokes an internal observation called $\boldsymbol{W}$ (see Fig. 4). For the purposes of this presentation we may assume that the internal observation $\boldsymbol{W}$ is a random number, a random vector or a function-valued object. Thus, in the three respective cases $\boldsymbol{W}$ = random variable, $\boldsymbol{W} = \langle \boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_r \rangle$ (random vector) or $\boldsymbol{W} = \boldsymbol{W}(t)$ (a stochastic process indexed by time t).

We will presume that a suitable probability law is defined on $\boldsymbol{W}$. We may then represent the probability of responding r_j given stimulus s_i by $P(r_j | s_i)$, as an abstract integral over the parts of the space where the response probability conditioned on $\boldsymbol{W} = w$ is nonzero,

$$P(r_j | s_i) = \int_{\{w | P(r_j | w) \neq 0\}} P(r_j | w) \, \mathrm{d}P(w | s_i).$$

In the event that there exists a conditional density on w, then $\mathrm{d}P(w | s_i) = f(w | s_i)\mathrm{d}w$,

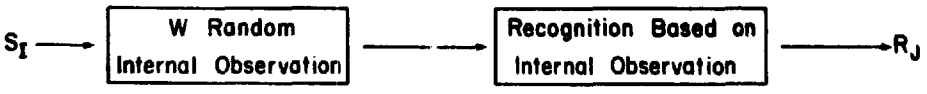

Fig. 4. The general schema for models based on an internal observation.

[2] In a psychophysical setting, when $n_s = n_r = 2$, the situation formally reduces to the classical signal detection situation of the yes-no or two-alternative forced choice variety (see Green and Swets, 1966).

and if $\boldsymbol{W}$ is r-dimensional, then we derive an r-multiple integral.[3] Contrarily, if $\boldsymbol{W}$ takes on discrete values, then the above formula becomes a sum. It is important to note that this formulation assumes that the response is independent of the stimulus, given w.

For example, suppose the observer encodes the randomly varying color (**C**), shape (**SH**) and size (**SZ**) of a visual pattern on separate dimensions, each of which is represented as a real-numbered value. Then $\boldsymbol{W}$ can be described as the random vector

$$\boldsymbol{W} = \langle \boldsymbol{C}, \mathbf{SH}, \mathbf{SZ} \rangle.$$

As noted, $\boldsymbol{W}$ might also be a function of time. For example, the stimulus pattern might be described as a deterministic function, $X(t)$, to which random noise, $\boldsymbol{N}(t)$, is added. $\boldsymbol{W}$ then becomes the random function

$$\boldsymbol{W}(t) = X(t) + \boldsymbol{N}(t)$$

where t represents time and where the stimulus is presented between time $t - T_1$ and T_2. If the observer makes observations at discrete time intervals $t_1, t_2, \ldots, t_r$ where $t - T_1 < t_i < T_2$, $(i = 1, 2, \ldots, u)$, then the internal observation becomes a random vector

$$\boldsymbol{W} = \langle \boldsymbol{W}(t_1), \boldsymbol{W}(t_2), \ldots, \boldsymbol{W}(t_r) \rangle.$$

It is well known from systems theory that, in the absence of noise and under some general assumptions, a finite number of samples taken from a deterministic function (here $X(t)$) completely describes that function (see, e.g., McGillem and Cooper, 1974). It is the addition of noise that makes recognition difficult for the 'best' observer.

In order to make much progress beyond this point, it is necessary for us to be willing to make further assumptions that will both narrow and further detail the classes of models with which we are dealing. This leads to the class of general discriminant models.

4. General discriminant models

The class of models we discuss in this section breaks down the internal processing of the stimulus in a detailed fashion. A schema for the general discriminant model is illustrated in Fig. 5. It is assumed that first of all a transformation of some kind is performed on the internal observation $\boldsymbol{W}$. The result of this transformation is then sent to a set of n other operations, one for each potential response and

[3] Note that we will be using a capital letter $P(\cdot)$ to denote probabilities defined on discrete events, while a lower case $p(\cdot)$ will be used to denote 'probabilities' defined on a continuum (i.e., a continuous density).

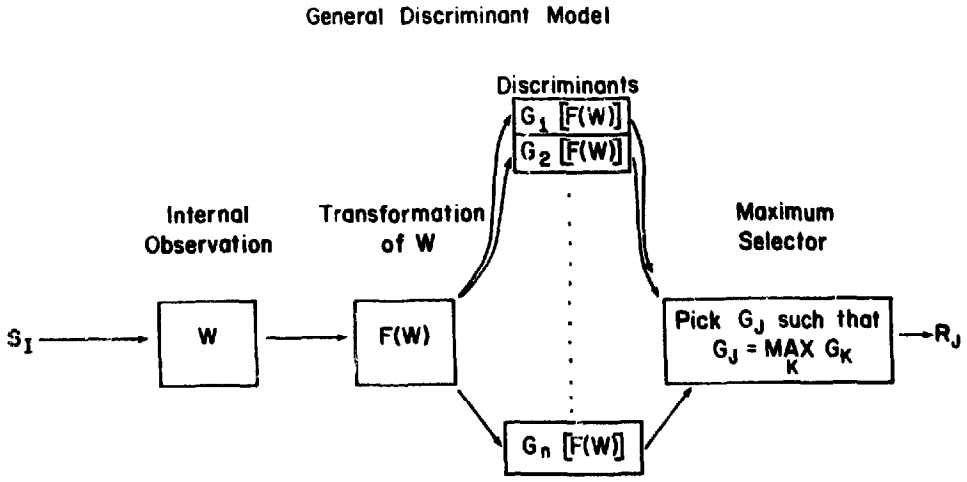

Fig. 5. The general schema for discriminant models.

somehow associated with it. Each of the corresponding n outputs is transmitted to a decision process which selects the response corresponding to the largest of the n outputs from the set of operations.

The n functions $G_k[F(\boldsymbol{W})]$ are known as 'discriminants' because as the kth discriminant function it produces a real number relating the transformed input stimulus ($F(\boldsymbol{W})$) to the kth response alternative. Each discriminant represents the obtaining of information about the relation of the input to the kth possibility, but it may also employ adjunct information such as payoffs, stimulus presentation frequencies and the like, that are pertinent to a specific situation.

To paraphrase the above in more detail, the transformation F first maps the internal observation $\boldsymbol{W}$ into another space of some type suitable for the operation of the discriminants. Each discriminant then maps $F(\boldsymbol{W})$ into a real number. The maximum selector then 'chooses' the largest real number, corresponding to the 'best' discriminant, as the response. Thus, the probability of selecting response r_j given the presentation of stimulus s_i in the general discriminant model is

$$P(r_j|s_i) = \int_{A_j} p[F(w)|s_i]\,\mathrm{d}P(w)$$

where

$$A_j = \{w|G_j[F(w)] = \max_k G_k[F(w)]\}$$

and $P(w)$ is the distribution of $\boldsymbol{W}$. That is, the integration takes place over that portion of the $\boldsymbol{W}$ space which leads to the jth discriminant being a maximum.

In the important circumstance that $F(\boldsymbol{W}) = \langle F_1(\boldsymbol{W}), F_2(\boldsymbol{W}), \ldots, F_m(\boldsymbol{W})\rangle$ (i.e., F is a vector-valued random function) and the discriminant functions are linear combinations of F, as in

$$G_k[F(\boldsymbol{W})] = a_{k1}F_1(\boldsymbol{W}) + a_{k2}F_2(\boldsymbol{W}) + \cdots + a_{km}F_m(\boldsymbol{W}) + a_{k,m+1},$$

then the system associated with this model is sometimes referred to as a *phi-machine.* The overall function of $\boldsymbol{W}$, going through both F and the linear operator, is called a phi-function:

$$\Phi_k(\boldsymbol{W}) = \sum_{i=1}^{m} a_{ki}F_i(\boldsymbol{W}) + a_{k,m+1} = G_k[F(\boldsymbol{W})], \quad 1 \le k \le n.$$

Suppressing the particular discriminant function index leaves the general notation for a phi-machine as

$$\Phi(\boldsymbol{W}) = a_1F_1(\boldsymbol{W}) + a_2F_2(\boldsymbol{W}) + \cdots + a_mF_m(\boldsymbol{W}) + a_{m+1}.$$

An introductory treatise on phi-systems is given by Nilsson (1965). Of interest to psychologists might be the theorems on the capacity of this class of systems. Some of the material on training such systems to recognize various types of patterns might also be of concern as forming incipient models for perceptual learning (see, e.g., Dodwell (1970, pp. 212–228)).

Although the general discriminant model does not by any means exhaust the universe of possible models it does provide a rather general scheme. It has been much investigated by researchers in artificial intelligence, engineering and econometrics to name a few regions of application. Therefore, not all discriminant models need carry any psychological rationale. However, we shall see that some do. For example, the general discriminant model encompasses many of the standard statistical decision models, as will be further explicated below.

There are many particular forms that discriminant models can take that have been of substantial interest in the social and biological sciences, as well as artificial intelligence and decision theory. Perhaps unsurprisingly, several of them emerge as linear discriminant models with special interpretations on the internal observation $\boldsymbol{W}$ and on the linear weightings in the discriminant functions.

4.1. Linear discriminant models

The linear discriminant model can be arrived at in two ways. In one way, it is considered to be a special case of the phi-machine model where F is a vector-valued random function and $F_i(\boldsymbol{W}) = \boldsymbol{W}_i$, so that

$$G_k(\boldsymbol{W}) = a_{k1}\boldsymbol{W}_1 + a_{k2}\boldsymbol{W}_2 + \cdots + a_{kr}\boldsymbol{W}_r + a_{k,r+1}, \quad 1 \le k \le n.$$

The $\boldsymbol{W}_i$ are considered to be real-valued random variables, that is, the vector function F maps the observation $\boldsymbol{W}$ into the vector $(\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_r)$ and the discriminant functions G_k are then linear combinations of the $\boldsymbol{W}_i$'s.

Alternatively, the linear discriminant model can be arrived at by *assuming* that $\boldsymbol{W} = (\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_r)$ (a real-valued random vector) and the $G(\boldsymbol{W}) = \sum_{j=1}^{r} a_{kj}(\boldsymbol{W}_j + a_{k,r+1}$ as before) letting $F(\boldsymbol{W}) = \boldsymbol{W}$. Note that F is not a vector function, and

no transformation is performed on the vector $\boldsymbol{W}$ (i.e., the identity transformation is performed). This also can lead to a linear representation of the discriminant functions as given by $G_k(\boldsymbol{W})$ above.

In both of these cases the $\boldsymbol{W}_i$ are interpreted as the components of the observation $\boldsymbol{W}$. The difference between the two cases concerns where in the general discriminant system the observation $\boldsymbol{W}$ becomes the vector of components $(\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_r)$ necessary for the linear discriminant representation. In both cases, the linear discriminant function, $G_k(\boldsymbol{W})$, is the same.

As is usual in the discriminant type of system, the maximizer selects the response associated with the largest discriminant function. Then, regardless of which case above gave rise to the random vector $\boldsymbol{W}$, the linear discriminant model can be formally written as

$$P(r_j|s_i) = P\left[\sum_{k=1}^{r} a_{jk}\boldsymbol{W}_k + a_{j,r+1} > \sum_{k=1}^{r} a_{mk}\boldsymbol{W}_k + s_{m,r+1} \text{ for all } m = 1, 2, \ldots, r\right]$$

$$= \int \int \cdots \int p(\langle \boldsymbol{W}_1 = w_1, \boldsymbol{W}_2 = w_2, \ldots, \boldsymbol{W}_r = w_r\rangle | s_i)\, \mathrm{d}w_1\, \mathrm{d}w_2 \cdots \mathrm{d}w_r$$

(where the integral is taken over all vectors $w = \langle w_1, w_2, \ldots, w_r\rangle$ such that $G_j(w) = \max_m G_m(w)$)

$$= \text{total probability on the } \boldsymbol{W} \text{ vector space such that } G_j(\boldsymbol{W}) = \max_m G_m(\boldsymbol{W}).$$

Note that for a given stimulus s_i, the *value* of $P(r_j|s_i)$ may differ between the two cases outlined above for representation of $\boldsymbol{W}$ as a random vector. That is, the conditional probabilities, $p(w|s_i)$, may have a different density depending upon how the random vector $\boldsymbol{W}$ is derived in the discriminant model.

The mathematical machinery of linear vector spaces can immediately be brought into describing how (and how well) linear discriminant systems function. We will content ourselves with a few observations. When two regions of the decision space, such that in one r_j is chosen and in the other r_k is chosen, share a common boundary, then it must be the case that

$$G_j(w) - G_k(w) = 0$$

along the boundary. This will of course be true whatever the nature of the G-functions. In the present (linear) instance, this separation surface is a line when $n = 2$, a plane when $n = 3$ and is called a hyperplane when $n > 3$. This result follows from inserting the linear functions into the above equation.

In systems such as these (where $\boldsymbol{W}$ is an r-valued random vector) it makes sense to think of the pattern space as an r-dimensional space where various prototype points, one for each stimulus, represent the central tendency or the 'ideal' of that particular stimulus. Each observation vector $\boldsymbol{W}$ is represented by one of the points in the

space, and is classified as the prototype class r_j according to some a priori decision method (e.g., in the linear discriminant system the decision method is the linear or planar separation surfaces).

One of the most common decision methods is the *minimum distance classifier.* Writing the prototype for s_i as $P_i = \langle p_{i1}, p_{i2}, \ldots, p_{ir} \rangle$ for a given observation $w_i = \langle w_{i1}, w_{i2}, \ldots, w_{ir} \rangle$, the minimum distance classifier computes the Euclidean distance between w_i and P_i, that is, writing $\| \cdot \|$ as the Euclidean norm and $(x - y)$ as the inner product of two vectors x and y we get

$$\| w_i - P_i \| = [(w_i - P_i) \cdot (w_i - P_i)]^{1/2}.$$

Minimizing this distance is equivalent to minimizing its square

$$\| w_i - P_i \|^2 = [(w_i - P_i) \cdot (w_i - P_i)] = (w_i \cdot w_i) - 2(w_i \cdot P_i) + (P_i \cdot P_i).$$

Because $(w_i \cdot w_i)$ provides no helpful information (it is the same for all prototypes), minimizing the whole expression is the same as maximizing

$$(w_i \cdot P_i) - \tfrac{1}{2}(P_i \cdot P_i) = \sum_{j=1}^{r} w_{ij} p_{ij} - \tfrac{1}{2} \sum_{j=1}^{r} p_{ij}^2,$$

so that a linear discriminant function is produced with weights

$$a_{ij} = p_{ij}, \quad 1 \leq j \leq r, \qquad a_{i,r+1} = -\tfrac{1}{2} \sum_{j=1}^{r} p_{ij}^2.$$

Although the Euclidean metric is perhaps most popular, a minimum distance classifier can also employ other distance metrics. Specifically, if $D(w_i, P_i)$ represents the distance between w_i and P_i, any case of the Minkowski metric will do where

$$D(w_i, P_i) = \left(\sum_{m=1}^{r} | w_{im} - p_{im} |^k \right)^{1/k}.$$

The Euclidean distance obtains when $k = 2$. Other commonly employed special cases are the city block distance ($k = 1$) and the square distance ($k = \infty$). However, this does not necessarily mean that any particular case of the Minkowski metric will yield equally satisfactory results in all applications, only that the Minkowski metrics do fit the model requirements of the minimum distance classifiers.

At this point it is important to note that distance metrics have come under considerable debate in the recent psychological literature. The debate revolves around whether the axioms of the classical distance metrics are satisfied in psychological (human) data. These well-known axioms are

$$D(X, Y) \geq D(X, X) = 0, \qquad D(X, Y) = 0 \quad \text{only if } X = Y, \tag{4.1}$$

$$D(X, Y) = D(Y, X), \tag{4.2}$$

$$D(X, Y) + D(Y, Z) \geq D(X, Z). \tag{4.3}$$

The essential elements of this debate can be found in Tversky (1977) and Krumhansl

(1978). Whether or not the distance axioms hold psychologically does not necessarily eliminate linear discriminate models from psychological interpretations, but it may strongly constrain the important special case of the minimum distance classifier.

4.2. Non-linear discriminant models

Not all pattern spaces can be completely separated into n distinct regions such that all the points associated with any particular stimulus pattern (i.e., observations associated with $p(\boldsymbol{W}|s_i)$, $1 \leq i \leq n$) can always be correctly identified by a linear discriminant system. Certainly when $p(\boldsymbol{W}|s_i)$ is continuous throughout the space there will be overlapping distributions and some mistakes will always be made. However, even when the total set of points is finite, they may be arranged in the space so that no linear functions are capable of separating them. An example occurs when no convex surface can separate the patterns of the various stimuli. A set of decision boundaries is convex if and only if any two points of the same region can be connected by a straight line without intersecting a boundary. Fig. 6(A) illustrates a circumstance where only a concave boundary can segregate two pattern classes for the case where $r = 2$. Note that there are s_2 pattern points which, when connected by a straight line, cause the straight line to intersect the boundary. It is straightforward to show that a linear discriminant system always results in convex decision regions (i.e., lines, planes or hyperplanes).

One of the more common non-linear discriminant systems is the *piecewise linear discriminant system*. Such a system differs from the purely linear case primarily in that there are two or more prototype points for each stimulus instead of one. Let $\mathcal{P}_i = \{P_i^{(1)}, P_i^{(2)}, \ldots, P_i^{(q)}\}$ be a set of prototype points corresponding to a stimulus s_i. For a minimum distance classification scheme, the discriminant functions then take the form

$$G_i(\boldsymbol{W}) = \max_k \{G_i^{(k)}(\boldsymbol{W})\}$$

where each $G_i^{(k)}(\boldsymbol{W})$ is defined as

$$G_i^{(k)}(\boldsymbol{W}) = a_{i1}^{(k)} \boldsymbol{W}_1 + a_{i2}^{(k)} \boldsymbol{W}_2 + \cdots + a_{ir}^{(k)} \boldsymbol{W}_r + a_{i,r+1}.$$

The $G_i^{(k)}(\boldsymbol{W})$ are referred to as subsidiary discriminant functions. Finding the maximum of the $G_i^{(k)}(\boldsymbol{W})$ functions for a given $\mathcal{P}_i$, and then selecting the maximum $G_i(\boldsymbol{W})$ is equivalent (as in the purely linear case) to finding the minimum distance of observation w in the space to *any* prototype point. That is, w will be classified as r_j if one of the prototypes of the jth alternative, say the mth, is closer than any other prototypes to w, that is,

$$D(w, P_j^{(m)}) = \min_k \{\min_v [D(w, P_k^{(v)})]\}.$$

An example of the decision surface for a piecewise system when $r = 2$, $\mathcal{P}_1 =$

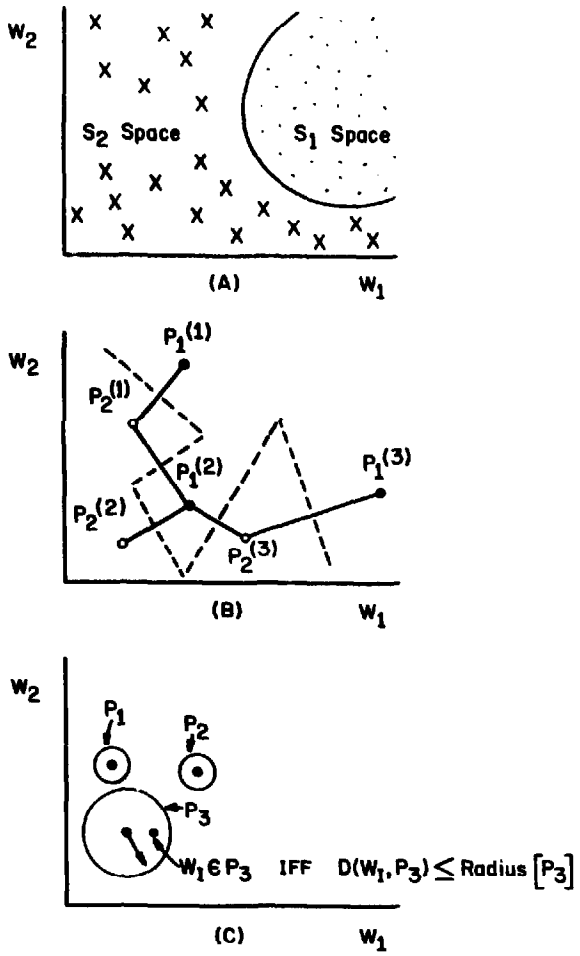

Fig. 6. Discriminant model decision surfaces for the case when $r=2$ (i.e., $\boldsymbol{W}=\langle \boldsymbol{W}_1, \boldsymbol{W}_2 \rangle$). (A) Concave decision surface; (B) A piecewise linear surface where $\mathcal{P}_1=\{P_1^{(1)}, P_1^{(2)}, P_1^{(3)}\}$ and $\mathcal{P}_2\{P_2^{(1)}, P_2^{(2)}, P_2^{(3)}\}$, the dashed line forms the decision boundary: to the right and above, select response r_1, otherwise select r_2; (C) Circular decision surfaces for the prototype points P_1, P_2 and P_3.

$\{P_1^{(1)}, P_1^{(2)}, P_1^{(3)}\}$ and $\mathcal{P}_2=\{P_2^{(1)}, P_2^{(2)}, P_2^{(3)}\}$ is given in Fig. 6(B). Note that the decision surface is linear over limited portions of the space (hence the name piecewise) but is not linear with respect to the decision surface of $\mathcal{P}_1$ relative to $\mathcal{P}_2$.

There exists an infinite number of non-linear discriminant systems. However, one that requires mentioning is where the decision surfaces surrounding the prototypes are circular, as in Fig. 6(C). This is a special case of the system where discriminants have the *quadratic form*,

$$G_i(\boldsymbol{W}) = \sum_{k=1}^{r} a_{kk} \boldsymbol{W}_k^2 + \sum_{k=1}^{r-1} \sum_{m=k+1}^{r} a_{km} \boldsymbol{W}_k \boldsymbol{W}_m + \sum_{k=1}^{r} a_k \boldsymbol{W}_k + a_{r+1}.$$

Although non-linear systems such as discussed above are fairly prominent in artificial intelligence (see, e.g., Batchelor, 1978; Mizoguchi, Shimura and Kakusho, 1980; Rahbar and Mix, 1980), such systems have rarely entered into psychological pattern recognition contexts.

4.3. Statistical decision theory

Statistical decision theory can be employed when the distribution on $\boldsymbol{W}$ is known (e.g. multivariate normal) and there exists a payoff matrix describing the gains and losses associated with the various correct and error responses. The goal within the approach is usually to maximize some aspect of performance (e.g., expected gain, number of correct responses, minimum gain, etc.). It is of great help in certain engineering approaches (just as is ideal detector theory in the $n = 2$ case (Green and Swets, 1966)) since it is possible there to design systems having the desired optimal properties. However, the general theory has not had much application in modeling human recognition since the difficulties that arise in estimating decision boundaries, the dimensionality of the space and parameters of the distributions are rather formidable. Nevertheless, it is an approach of considerable generality under which a number of the other techniques may be subsumed, and is also illustrative of the various facets with which an observer is confronted. Potentially it could also serve as an 'ideal' against which the performance of a real observer can be compared.

It is out of the question to attempt to review all previously suggested possible statistical goals here, so we content ourselves with that goal where the observer attempts to maximize the expected gain for any particular internal observation $\boldsymbol{W} = w$. It will be seen that several other well-known modes of statistical decision making emerge as special cases of some of the present approaches, for example,

(1) maximization of probability correct,
(2) maximum likelihood strategy,
(3) linear discriminant model,
(4) template matching model,
(5) minimum distance classifier.

The statistical decision models assume that $F(\boldsymbol{W}) = \boldsymbol{W}$, and also make use of the so-called gain function, which we will denote as $J(r_j | s_i)$, which gives the amount (utility, etc.) earned when response r_j is made given that stimulus s_i was the one presented. When an observer receives a specific observation $\boldsymbol{W} = w$, he/she wishes to select his/her response in line with a conditional probability distribution that will maximize his/her expected gain. That is, he/she wishes to maximize

$$E(J | w) = \sum_{j=1}^{n} \sum_{i=1}^{n} P(r_j \cap s_i | w) J(r_j | s_i).$$

On the basis of our previous assumption that given $\boldsymbol{W} = w$, the response is independent of the stimulus, we can write

$$P(r_j \cap s_i | w) = P(r_j | s_i \cap w) P(s_i | w) = P(r_j | w) P(s_i | w)$$

and then

$$E(J|w) = \sum_{j=1}^{n} P(r_j|w) \sum_{i=1}^{n} P(s_i|w) J(r_j|s_i)$$

$$= \sum_{j=1}^{n} P(r_j|w) \sum_{i=1}^{n} \frac{p(w|s_i) P(s_i) J(r_j|s_i)}{p(\boldsymbol{W}=w)}$$

utilizing Bayes' theorem.

Now, maximizing this is the same as maximizing the same expression in the absence of the $p(\boldsymbol{W}=w)$ term because the latter term does not affect the choice; hence, the observer in essence wishes to maximize

$$E^*(J|w) = \sum_{j=1}^{n} P(r_j|w) \sum_{i=1}^{n} p(w|s_i) P(s_i) J(r_j|s_i).$$

Note that this is a weighted sum of the amounts earned in making the various responses where the weights are the probabilities of making each response, given the observation $\boldsymbol{W}=w$. That is it averages all the possible 'earnings' each of which is associated with one of the responses. Clearly, the average amount gained will be largest with the weight (i.e., conditional probability) on the responses with the largest average conditional gain. That is, the observer should simply find r_j such that

$$\sum_{i=1}^{n} p(w|s_i) P(s_i) J(r_j|s_i) = \max_k \sum_{i=1}^{n} p(w|s_i) P(s_i) J(r_k|s_i)$$

and then let $P(r_k|w)=0$ for all $k \neq j$ and let $P(r_j|w)=1$. Put in another way, the observer does best for any specific value of $\boldsymbol{W}$ by simply looking for the response with the maximal conditional gain. Note further that this strategy maximizes the expected gain averaged over all possible values of $\boldsymbol{W}$ because if the choice is optimal for a specific $\boldsymbol{W}=w$, no other strategy can produce a better overall return.

This is the same procedure as calculating the discriminant functions, $G_k(w)$, each defined as

$$G_k(w) = \sum_{i=1}^{n} p(w|s_i) P(s_i) J(r_k|s_i), \quad 1 \leq k \leq n,$$

and picking the $G_j(w)$ (i.e., r_j) that is maximal. Thus, the decision theory approach can fit into the discriminant function family. Often, of course, the above discriminant may be quite refractory to employ but there are certain cases of reasonable simplicity.

One special case of considerable import obtains when the payoff matrix is an especially simple one, namely, it is assumed that one unit of pay is gained (any constant amount would do) when any correct response is made and nothing is gained if an incorrect response is made. That is, $J(r_k|s_i)=1$ if $i=k$ and $J(r_k|s_i)=0$ otherwise; this renders

$$G_k(w) = p(w|s_k) P(s_k) = P(w \cap s_k)$$

so that the observer must attempt to maximize the *joint probability* of the observation and the stimulus. Moreover, this case can also lead to a maximization of *probability correct*. Starting with the probability correct conditionalized on a given $\boldsymbol{W} = w$ yields

$$P(\text{correct} \mid w) = \sum_{i=1}^{n} P(r_i \cap s_i \mid w) = \sum_{i=1}^{n} P(r_i \mid w) \cdot P(s_i \mid w)$$

$$= \sum_{i=1}^{n} P(r_i \mid w) \frac{P(s_i \cap w)}{p(\boldsymbol{W} = w)}$$

through the w-conditional independence of stimulus and response assumption and the definition of conditional probability. With respect to the observer, maximizing this expression will only depend upon $P(r_i \mid w)$ because that is the only term that the observer can 'manipulate'. Note also that the denominator, $p(\boldsymbol{W} = w)$, plays no role in the maximization. Therefore, setting

$$P(r_i \mid w) = \begin{cases} 1 & \text{if } P(s_i \cap w) = \max_k P(s_k \cap w), \\ 0 & \text{otherwise} \end{cases}$$

is equivalent to selecting response r_j if and only if the joint probability of s_j and w is maximal. It follows that the overall probability correct must thereby be maximized,

$$\max P(\text{correct}) = \max \int_{\{w\}} P(\text{correct} \mid w) p(\boldsymbol{W} = w) \, dw$$

$$= \int_{\{w\}} \max_i [P(r_i \cap s_i \mid w)] p(\boldsymbol{W} = w) \, dw$$

$$= \int_{\{w\}} \max_i [P(r_i \mid w) P(s_i \mid w)] p(\boldsymbol{W} = w) \, dw$$

$$= \int_{\{w\}} \max_i [P(r_i \mid w) P(s_i \cap w)] \, dw$$

$$= \int_{\{w\}} \max_i [P(s_i \cap w)] \, dw.$$

The first step is valid because the value of the integral will be as large as possible if and only if $P(r_i \cap s_i \mid w)$ is as large as possible for each value of w. The reasoning for the last step is that any deviation from $\max_i P(s_i \cap w)$ would lower the overall integral, so one should set $P(r_i \mid w) = 1$.

Yet a further alternate rule follows from

$$P(s_i \mid w) = \frac{p(w \mid s_i) P(s_i)}{p(\boldsymbol{W} = w)}$$

and maximizing $p(w \mid s_i) P(s_i)$ in the above integral. This decision strategy is then referred to as the *unconditional maximum likelihood* rule. Systems implementing this type of decision are known as Bayes' classifiers for obvious reasons (see Fu,

1970). Furthermore, if $P(s_i) = P(s_j)$ for all i and j, then the decision strategy is called the *conditional maximum likelihood* rule.

Another family of models, the *multivariate normal*, is generated by assuming that $\boldsymbol{W} = \langle \boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_r \rangle$ is a real-valued random vector where $w = \langle w_1, w_2, \ldots, w_r \rangle$ is a specific observation (considered as an $r \times 1$ column vector in the vector and matrix algebra to follow) and letting the $p(\boldsymbol{W} = w \mid s_i)$ be multivariate normal. We can therefore write the density

$$p(\boldsymbol{W} = w \mid s_i) = \frac{}{|D_i|^{1/2}(2\pi)^{r/2}} \cdot \exp\{-\tfrac{1}{2}(w - M_i)^t D_i^{-1}(w - M_i)\}$$

where D_i is the variance/covariance matrix of stimulus s_i for the r-dimensions case, D_i^{-1} its inverse and $|D_i|$ its determinant. M_i is the mean or expectation vector of the distribution and $(w - M_i)^t$ the transpose of the vector representing the difference between the observation vector w and the mean vector M_i. In the two-dimensional case ($r = 2$) this can be written in terms of its parameters more explicitly as

$$p(w = \langle w_1, w_2 \rangle \mid s_i) =$$

$$= \exp\left\{\frac{(w_1 - m_{i1})^2 \sigma_{i22}^2 - 2(w_1 - m_{i1})(w_2 - m_{i2})\delta_{i12}^2 + (w_2 - m_{i2})^2 \sigma_{i11}^2}{-2(\sigma_{i11}^2 \sigma_{i22}^2 - \sigma_{i12}^4)}\right\}$$

$$\times (2\pi)^{-1}(\sigma_{i11}^2 \sigma_{i22}^2 - \sigma_{i12}^4)^{-1/2}$$

where

$$\sigma_{ijk}^2 = \begin{cases} \text{covariance of dimension } j \text{ with } k \text{ for stimulus } s_i, & j \neq k \\ \text{variance of dimension } j(k), & j = k, \end{cases}$$

m_{ij} = expectation on dimension j for stimulus s_i.

An example of a two-dimensional normal model would be auditory signal recognition where frequency and intensity are varied across the n signals and Gaussian noise is added to the stimuli. Then, pitch and loudness would form the psychological dimensions with the covariance terms of D_i standing for the perceptual 'correlation' of pitch with loudness for a given stimulus s_i. Another example will be given in the following section which discusses template models in more detail.

The expected gain maximization problem can also be examined in the multivariate normal model. In the general case $G_j(\boldsymbol{W})$ is obviously a weighted sum of exponentials and may be too refractory to lead very far. However, suppose a correct response earns one unit and an incorrect response earns nothing. Then, as was shown above, the maximizer selects the response associated with the largest joint probability of the observation ($\boldsymbol{W} = w$) and the stimulus. That is,

$$\max_j G_j(w) = \max_j P(s_j) \frac{1}{(2\pi)^{r/2}|D_j|^{1/2}} \exp\{-\tfrac{1}{2}(w - M_j)^t D_j^{-1}(w - M_j)\},$$

which is equivalent in a decisional sense to finding

$$\max_j \ln G_j(w) =$$

$$= \max_j [\ln P(s_j) - \tfrac{1}{2} r \ln 2\pi - \tfrac{1}{2} \ln |D_j| - \tfrac{1}{2}(w - M_j)^t D_j^{-1} (w - M_j)].$$

Dropping the constant ($\frac{1}{2} r \ln 2\pi$) and grouping the terms which do not involve the observation leads to

$$\max_j [C_j - \tfrac{1}{2}(w - M_j)^t D_j^{-1} (w - M_j)]$$

where

$$C_j = \ln P(s_j) - \tfrac{1}{2} \ln |D_j|.$$

This is a quadratic expression because it involves the squares of the entries of the observation vector, and thus would produce *non-linear* decision surfaces.

To be sure, the normal case reduces to a *linear discriminant* model when the further constraint is imposed that the covariance matrices are all identical. This can be seen by expanding the above quadratic form (and letting $D_j = D$),

$$G_j'(w) = \ln P(s_j) - \tfrac{1}{2} \ln |D| - \tfrac{1}{2} w^t D^{-1} w + w^t D^{-1} M_j - \tfrac{1}{2} M_j^t D^{-1} M_j.$$

Observing that the terms $-\frac{1}{2} \ln |D|$ and $-\frac{1}{2} w^t D^{-1} w$ do not depend on the stimulus s_j, the discriminant can be rewritten as

$$G_j''(w) = w^t D^{-1} M_j + \ln P(s_j) - \tfrac{1}{2} M_j^t D^{-1} M_j$$

where

$$w^t D^{-1} M_j = < w_1, w_2, \ldots, w_r > \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1r} \\ d_{21} & & & \\ \vdots & & & \vdots \\ d_{r1} & \ldots & & d_{rr} \end{bmatrix} \begin{bmatrix} m_{j1} \\ m_{j2} \\ \vdots \\ m_{jr} \end{bmatrix}$$

$$= \left\langle \sum_{i=1}^{r} w_i d_{i1}, \sum_{i=1}^{r} w_i d_{i2}, \ldots, \sum_{i=1}^{r} w_i d_{ir} \right\rangle \cdot \begin{bmatrix} m_{j1} \\ m_{j2} \\ \vdots \\ m_{jr} \end{bmatrix}$$

$$= \sum_{k=1}^{r} \left(\sum_{i=1}^{r} w_i d_{ik} \right) m_{jk} = \sum_{k=1}^{r} \sum_{i=1}^{r} w_i d_{ik} m_{jk},$$

which is obviously a linear function of the observation vector components. Thus, the result gives a true *linear discriminant* function of the specific observation w.

Now suppose that

$$d_{ik} = \begin{cases} c & \text{if } i = k, \\ 0 & \text{if } i \neq k \end{cases}$$

where c is a constant. That is, the r dimensions are orthogonal with equal variances. Then

$$G_j(w) = \sum_{k=1}^{r} w_k m_{jk}.$$

This simplified $G_j(w)$ can be interpreted as a *template matching* model where the mean vectors M_k represent the templates. More comments on this will be made in the next section.

Finally, utilizing the simple payoff matrix and assuming that $P(s_i) = P(s_j) = 1/n$ and $D_i = D_j = I$ (the identity matrix) for i and j, then

$$G_j(w) = -\ln(n) - \tfrac{1}{2}\ln(I) - \tfrac{1}{2}[|(w - M_j)|^2].$$

That is, maximizing $G_j(w)$ is minimizing in this case $|(w - M_j)|^2$, the squared distance of the observation from the points M_j. This is therefore a *minimum distance classifier,* where the a priori distributions are normal and the prototype pattern points are just the means of the various distributions.

4.4. Template matching models

It was shown above that, with a sufficiently simple payoff matrix and identical covariance matrices, the normal version of the decision theoretic approach could be interpreted as a template matcher. The idea of template matching has played a significant role in psychological theories of pattern recognition so an amplified discussion seems justified. We shall relate the qualitative background before returning to a more quantitative account.

The template concept has suffered somewhat as a whipping boy in psychology over the past few decades, and the chief reason for that is the very strict version usually brought forth (Neisser (1967) gives representative examples of this perspective). The typical axioms, often not very explicit, are the following.

(1) The memory template is an exact replica of the stimulus.

(2) A positive match occurs between an input pattern and the template if and only if there is complete congruence (a perfect match) between the two, otherwise nothing happens. For example, there is no partial information transmitted.

(3) The input pattern is processed as a unit; thus, all parts are processed simultaneously.

The first two assumptions are the most important and are responsible for the statements that template matches cannot allow for translation, size changes or any rotation of the pattern set, because these would, of course, preclude a perfect congruence match. More general geometric or topological transformations (e.g., nonuniform stretching) are ruled out a fortiori. The third axiom has implications for reaction time, namely, all parts of the pattern must be processed *and completed* simultaneously (a strong type of parallel processing). This will not be discussed further here but for relevant discussions of stochastic parallel vs. serial processing

models in psychology, see Townsend (1972, 1974, 1976a, b) and Townsend and Ashby (1983).

As will be seen below, many feature matching models assume that each 'feature' is sampled and matched from the input stimulus in an all-or-none manner. In this sense, a strict template matching system is nothing but a special case of a feature matching system, where there exists only one feature (the pattern itself) that is tested against a similar memory feature in an all-or-none fashion (Townsend, 1970). Thus, it is little wonder that template models were so easily 'falsified', often in thought experiments and pictorial demonstrations (e.g., Neisser, 1967).

The early ideas of the Gestalt psychologists (e.g., Köhler, 1947) on topological isomorphisms between stimulus patterns and the brain's representation, are qualitatively in line with a more sophisticated interpretation of template matching. In such a scheme the input pattern may be transformed topologically (to oversimplify, any stretching or moving around that does not result in a tear of the surface of the pattern) and then compared with the internal memory pattern. More recent approaches of this type are the so-called normalization theories. Normalization embodies the concept of mentally performing rigid physical transformations, such as rotation or translation, in order to bring one stimulus into congruence with another. Although normalization is not logically necessary for pattern recognition, it does give evidence for the types of processes and internal representations used in human perception (see Cooper and Shepard, 1973; Kubovy and Podgorny, 1981; Neisser, 1967). Additionally, some progress has been made in employing general topological and algebraic notions in a mathematically rigorous fashion to problems in perception (see, e.g., Cowan, 1977; Hoffman, 1966; Zeeman, 1962). In a similar spirit, but from a different vantage point, Cavanagh (1972) has proposed a mathematical treatment of a holographic matching process and Pribram (1971) offers a qualitative account of holographic memory systems. We cannot pursue these matters here, but the incorporation of these kinds of concepts in a reasonable process model may be of import to the field; particularly if means are developed that enable rigorous empirical testability.

On a more prosaic plane, how does the multivariate normal model mentioned in the last section fit in here? Well, it must be admitted that, without some extra machinery, such systems cannot fully answer the queries posed by the anti-templatists. Even though there is a distribution on the possible observations, which to some extent helps to get out of the deterministic bind, the basic 'template', the pattern distribution's mean vector, is, indeed, stuck in a particular position in space and orientation and size.

Engineers and people working in artificial intelligence have taken several paths out of this dilemma, but one of the more common has been to preprocess the input pattern to bring it into standard form, orientation and so on for comparison with the internally stored templates. This is essentially the same idea as the 'normalization' approach mentioned above. Another technique sometimes employed is to carry out a series of comparisons each with a different position,

rotation and possibly size with the internal templates. The best orientation, etc., is chosen for each alternative and then, as usual, the overall maximizer selects the best alternative (the one having the closest match). An interesting set of hypotheses about how size and position invariance in pattern perception might be attained in the brain have been put forth by many workers in the domain of visual spatial frequency analysis and related areas (e.g., Cavanagh, 1978; Köhler, 1947).

Recall that in the multivariate template model, $p(\boldsymbol{W}=w|s_i)$ is multivariate normal and the template is thought of as being represented by the mean pattern vector. The match process takes the form of a cross correlation of the input pattern vector with the mean pattern vector. The mean pattern vector is, as it were, acting as a filter on the input. This notion of filters is more than metaphorical. One way in which the above cross correlator or template matcher is arrived at is by way of filter theory. It can be shown that when a deterministic input pattern goes through a channel obscured by Gaussian noise, then the filters that maximize the signal-to-noise ratio are just the templates represented by the mean pattern vectors. These are called 'matched filters'.[4] Note that no decision rule is yet present. Maximizing the signal-to-noise ratio for any given pattern is optimal from the sensory perspective but bears no *direct* implication from the decision point of view. Nevertheless, it will be seen immediately that the mean vector acting as a cross-correlator permits interpretation as a discriminant function to be potentially employed in a discriminant maximization process.

In general, a filter can be represented as the weighting function in a linear system (McGillem and Cooper, 1974). In the case where $\boldsymbol{W}=\langle \boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_r\rangle$ is a real-valued random vector, we will represent the weighting function for stimulus s_i as $h_i(k)$, $1 \le k \le r$. The output of a pattern filter associated with stimulus s_i can then be written as

$$G_i(\boldsymbol{W}) = \sum_{j=1}^{r} h_i(j)\boldsymbol{W}_j.$$

When this filter is 'matched' to conform to stimulus s_i, h becomes the mean vector of the s_i distribution (Davenport and Root, 1958),

$$G_i(\boldsymbol{W}) = \sum_{j=1}^{r} m_{ij}\boldsymbol{W}_j.$$

Obviously, we used G to represent the output since it is a random discriminant function. The expectation of G when s_i is presented can be found as

$$E[G_i(\boldsymbol{W})] = \sum_{j=1}^{r} m_{ij}E(\boldsymbol{W}_j|s_i) = \sum_{j=1}^{r} m_{ij}^2,$$

[4] The concepts of the 'matched filter' and cross-correlator are widely employed in engineering linear systems theory. With the necessary assumptions (alluded to here and in various portions of this paper), psychological applications are relatively straightforward. A discussion of these concepts can be found in McGillem and Cooper (1974). A well-known psychological application of a cross-correlator is the ideal detector concept in signal detection theory (Green and Swets, 1966).

so that the mean output of this filter or correlator or template matcher in this case is just the sum of the means squared, which can be interpreted as the energy in the pattern qua signal. However, it should be noted that these matched filters are *not* always optimal. As was shown, in many circumstances the optimal discriminant functions in terms of maximizing expected gain will be quadratic (or worse) functions of the observation. It is only in special cases that linear discriminant functions with mean pattern vectors as the discriminants will maximize expected gain.

A final topic to be considered in this section concerns a more down-to-earth example of a template matcher. Suppose the purpose is to recognize members of a set of two-dimensional visual patterns. The internal representation of a stimulus s_j might be given by a function of the points in the plane $S_j(x, y)$ where the function S gives the reflected light intensity at any point (x, y). Under reduced illumination, or when subjected to visual noise, the observation $\boldsymbol{W}$ will be a randomized version of the original stimulus. The random observation $\boldsymbol{W}$ will then be correlated against the internal templates, whichever stimulus it came from. We have been dealing with discrete vector representations but here perhaps a continuous account is called for; the random discriminant (correlator) function for r_j would be

$$G_j(\boldsymbol{W}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S_j(x, y)\, \boldsymbol{W}(x, y)\, dx\, dy$$

where, practically, the integration is just over the useful region. When $\boldsymbol{W}$ is from s_j, then

$$E[G_j(\boldsymbol{W})] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S_j^2(x, y)\, dx\, dy = \text{energy of stimulus } s_j.$$

In practical cases it is usually feasible, if not mandatory, to approximate the continuous $S_j(x, y)$ representation by breaking the plane into a checkerboard grid of cells and represent the stimulus by whether or not each cell is filled in or not (for instance, a quantization decision might be made in each cell such as whether or not it is more than half filled). When the comparison is accomplished for any given template the result is the discrete analogue of the continuous case, except we have the further simplicifation that S and $\boldsymbol{W}$ are now 0–1 functions depending on whether or not each cell is filled (e.g., $S_j(x, y) = 1$ or 0). Therefore,

$$G_j(\boldsymbol{W}) = \sum_x \sum_y S_j(x, y)\, \boldsymbol{W}(x, y) = \text{number of overlapping filled cells.}$$

There are many ways to make correlation devices more sophisticated. Extensions of this kind of idea, employing autocorrelation functions, have also been suggested (e.g., Lappin and Uttal, 1976; Uttal, 1975).

4.5. Feature-discriminant models

Let us first recall that in the most general class of discriminant models, we allowed the observation $\boldsymbol{W}$ to be transformed by F and that the discriminant functions, G, then worked on F, and were followed inevitably by the maximizer.

Features may first of all be roughly characterized as aspects of a pattern that can be used by a recognition system. In turn we may attempt to model their processing as measurements that are done on the observation pattern. Of course, we usually impose a second criterion, namely that such a measurement be on some 'meaningful' aspect of the stimulus. This notion of meaningfulness is difficult to define in a tight manner, but there is little doubt that it carries some force. For example, the spatial template matchers discussed above are not feature testing systems because the points $S_j(x, y)$ are too trivial to deserve the name 'feature'. It is not the intent of this paper to outline the various theoretical and metatheoretical arguments concerning the definition of 'feature' and its application to recognition problems. However, it should be noted that the arguments are not always minor and reflect the fact that feature models (of all types) often stand or fall on the adequacy of their definition of 'feature'.

In any case, within the class of discriminant systems, F can be used to capture the feature processes. That is, let $F(\boldsymbol{W}) = (F_1(\boldsymbol{W}), F_2(\boldsymbol{W}), \ldots, F_m(\boldsymbol{W}))$ where $F_i(\boldsymbol{W})$ corresponds to the measurement of feature i and there are m features sampled or measured on each stimulus presentation. Notice that in general the F_i's could be numbers, sets, etc., but we shall assume that each such feature computation results in a real positive number, possibly representing the degree to which evidence for the feature was found in the observation (for instance, it might be $P(i \mid \boldsymbol{W})$, the a posteriori probability that feature i was present, given the observation). In the event that the features are tested for in an all-or-none manner, the F_i's would be 0 or 1, corresponding to the result of the decision on the presence or absence of the feature.

The result of the transformation is sent to the discriminant functions, each of which can be viewed as computing the similarity of the input pattern to a particular memory pattern. For example, when the F_i's are 0 or 1, then the discriminant functions may calculate the similarity based on feature overlap and nonoverlap. This may be done in many ways (see, e.g., Tversky, 1977). The final step is the decision by the maximizer as to the largest discriminant output. A view of the feature measurements as being themselves produced by little feature templates is given in Townsend and Ashby (1978) and a thorough development of feature processing from a linear systems approach with physiological overtones is found in Mortensen (1978a,b).

Most of the feature models that have been employed in predicting confusion results (e.g., Geyer and Dewald, 1973) have been of a rather different variety than those above, not generally being representable as discriminant models. Those types will be discussed in the next section, and so much time will not be spent with the present discriminant type models. However, we do have to mention one particular

model that has become quite famous, perhaps due somewhat to its whimsical formulation in which the pattern is attacked and devoured by a host of 'demons'. We refer to *pandemonium* (Selfridge, 1959). In pandemonium, the observation is stored and passed on by data demons to the next higher-up demons, the computational demons who perform the F-transformations, serving to measure the features. The degree of evidence for a feature is conveyed to succeeding authorities by howling, the louder the howl by a feature or computational demon the better the evidence for that feature in the input pattern. The succeeding cognitive demons are responsible for gathering information for a particular pattern (e.g., the letter T); there is one demon for each possible stimulus pattern. Finally, each of these demons yells above to the grand high muckety-muck demon, in volume proportional to the evidence (number and strength of features belonging to his pattern) for his particular stimulus alternative. The decision demon 'says' the response corresponding to the loudest pattern yell. Clearly, this fits nicely into the discriminant function scheme where the data demons do little more than represent the observation $\boldsymbol{W}$, the feature demons correspond to the transformation F, the cognitive demons parallel the discriminant functions and, finally, the decision demon is the maximum selector.

5. Feature confusion models

The models discussed here are notable in that a number of its members have been given closed mathematical form and actually fit to confusion data[5] ($P(r_j|s_i)$; i, $j = 1, 2, \ldots, n$), unlike most of the foregoing models. That is, explicit formulae can be given to their predictions, their parameters estimated by one or more methods, and then the resulting numerical predictions tested statistically against the observed $P(r_j|s_i)$ values. Some models of this type can be found in Rumelhart and Siple (1974), Geyer and Dewald (1973), Wandmacher (1976), Mortensen (1978a,b), Townsend, Hu and Ashby (1980) and Townsend and Ashby (1976). Two examples are given below.

The basic structure of these models is given in Fig. 7 where it can be seen that we still have the idea of an observation followed by a (feature measuring) transformation (or more exactly, a set of transformations), F_i. Moreover, the succeeding stage continues to encompass the G-functions, which here act to compute the similarity of $\boldsymbol{W}$ (via the feature computations) to the various pattern alternatives. Now, however, the difference between the present models and

[5] In a typical confusion experiment, the stimuli presented to the subject are usually degraded in some manner. This results in imperfect discrimination by the subject, and the responses made often do not correspond with the stimuli presented. Those confusion data can then be arrayed as a confusion matrix. Usually, the rows of the matrix represent stimuli and the columns the responses in such a way that row i and column j designate the proper response-to-stimulus assignment. Each cell in the matrix contains the proportion of times the subject responded j when stimulus i was actually presented, and is denoted by $P(r_j|s_i)$.

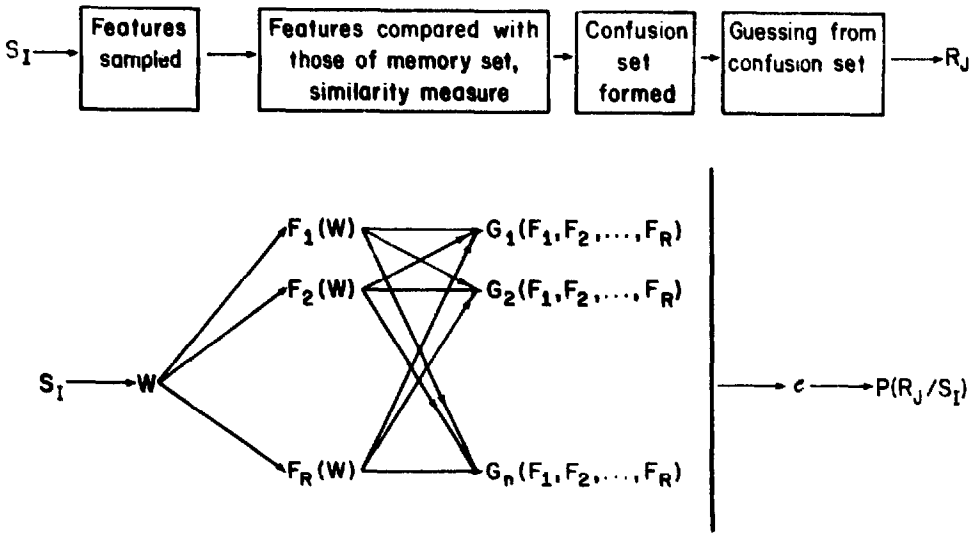

Fig. 7. The general schema for feature confusion models.

discriminant-feature models arises; based on the feature-similarity results, a set of candidate alternatives is formed that typically acts to exclude those possibilities that are sufficiently dissimilar from the input. That is, the observer (in some fashion) establishes some sort of a 'minimum similarity criterion'. *All* discriminant function outputs that exceed this criterion make up a reduced confusion set. The final stage is taken up with a guessing process, at which time one of the candidates from the confusion set is selected as the response. Notice that, in a sense, the feature *discriminant* models are special cases of this scheme, where only that alternative having the maximum G-value is placed in the confusion set. The overall probability of response r_j conditional on a presentation of s_i is

$$P(r_j|s_i) = \sum_{\mathscr{C}} \sum_{\mathscr{F}} \int_{\{W\}} P(r_j|\mathscr{C}) P(\mathscr{C}|\mathscr{F}) P(\mathscr{F}|w) p(w|s_i).$$

In this general formulation allowance is made for the possibility that $\boldsymbol{W}$ is stochastic, the set of feature measurements $\mathscr{F}$ is stochastic and the confusion set $\mathscr{C}$ is stochastic as well of course as the response chosen by guessing from the confusion set. Most models, however, begin with a 'sampled' set of features $\mathscr{F}$. This bypasses $\boldsymbol{W}$ and assumes that the F's are 0 or 1 corresponding to each potential feature (in these cases it may be more reasonable, for bookkeeping purposes, to represent the feature measurements as vectors of 1's and 0's, corresponding to the presence or absence of each feature). This sampled feature set is then compared with the sets of features that make up the various alternatives, and this is deterministic too. Only the last process, the guess, is probabilistic, once the features are known, leaving us with

$$P(r_j|s_i) = \sum_{\mathscr{C}} \sum_{\mathscr{F}} P(r_j|\mathscr{C}) P(\mathscr{C}|\mathscr{F}) P(\mathscr{F}|s_i)$$

where $P(\mathscr{C}) = 1$ for the appropriate confusion set and 0 otherwise. Townsend, Hu and Ashby (1971) present and briefly discuss the standard axioms underpinning notions of feature representation and sampling (i.e., 'extraction') in psychological models of feature processing. Over the past few years direct experimental testing of these axioms has begun (Townsend and Ashby, 1976; Townsend, Hu and Ashby, 1981; Wandmacher, 1976). For example, Townsend and Ashby (1976) were forced to reject the common assumption that features could be lost, but not artificially (i.e., incorrectly) gained from a stimulus. The usual axiom that features are extracted independently of one another was also falsified. Both exemplary models exhibited below make this assumption.

Although there are many extant feature models, we will examine only two representative examples. In one (Rumelhart, 1970, 1971) it is assumed that all features have equal independent probabilities of being sampled. The probability that a feature which is not in a stimulus is sampled is assumed to be zero. Let $\mathscr{F}_{rj}$ be the set of features in response r_j, $\mathscr{F}_{\bar{r}j}$ the set of features not in r_j but that are in some other r_k, $\mathscr{F}_{si}$ the set of sampled features of stimulus s_i, $\mathscr{F}_{\bar{s}i}$ the set of features available in s_i but not sampled and N a function such that $N(\mathscr{F}_j)$ is the number of features in set $\mathscr{F}_j$. In this model the responses r_j entering the confusion set $\mathscr{C}$ arise from presentation of stimulus s_i according to the joint occurrance of two conditions:

$$N(\mathscr{F}_{rj} \cap \mathscr{F}_{\bar{s}i}) = 0, \quad \text{equivalently,} \quad \mathscr{F}_{si} \subseteq \mathscr{F}_{rj}, \tag{5.1}$$

$$N(\mathscr{F}_{\bar{r}j}) - N(\mathscr{F}_{si}) \leq a \tag{5.2}$$

where a is some positive integer constant. That is, response r_j enters $\mathscr{C}$ if all sampled features are present in the feature set of r_j and at most 'a' features in $\mathscr{F}_j$ are missing from $\mathscr{F}_i$.

Response r_j is then selected from $\mathscr{C}$ according to the Bayesian rule,

$$P(r_j|\,\mathscr{C}) = \begin{cases} b_j & \text{if } \mathscr{C} = \emptyset, \\ \dfrac{P(\mathscr{C}|s_j)b_j}{\sum_{k=1}^{n} P(\mathscr{C}|s_k)b_k} & \text{if } r_j \in \mathscr{C}, \\ 0 & \text{if } r_j \notin \mathscr{C} \text{ and } \mathscr{C} \neq \emptyset. \end{cases}$$

The probabilities b_j are interpreted as biases for each of the possible responses.

The second representative model (actually a group of models, see Geyer and Dewald, 1973) assumes that the features each have a unique probability of being sampled, unlike the previous example. The confusion set $\mathscr{C}$ is then formed through the use of a 'hit-ratio', HR, defined as

$$\text{HR} = \frac{N(\mathscr{F}_{si} \cap \mathscr{F}_{\bar{r}j}) - N(\mathscr{F}_{\bar{s}i} \cap \mathscr{F}_{\bar{r}j})}{N(\mathscr{F}_{\bar{r}j})}.$$

Whether or not a response r_j enters $\mathscr{C}$ depends upon the use of the HR. Geyer and Dewald (1973) found that the best fitting version of this model contained an upper and lower HR threshold. If one response alternative exceeded the upper HR

threshold, that response was selected. If several responses exceeded the upper threshold, then the subject utilized a guessing bias for each alternative in order to choose one. If all alternatives fell below the lower threshold, then a response was selected according to the guessing biases. In the case where the maximum HR fell between the thresholds, with probability α the subject resorted to the below threshold guessing strategy, and with probability $1-\alpha$ the response with the largest HR was selected.

6. Descriptive models

The models pictured in this section may be considered less substantive than many of the other models considered above, especially in the sense that they ignore any observation or immediate transformation of an observation. For this reason they will be referred to as descriptive models. This does not mean, however, that the distinction being drawn is between *process* and *non-process* models.

Roughly a process model describes how the cognitive functioning is taking place, whereas a non-process model tends to be mum with regard to how the processing is being accomplished. This has been, and still is, a fuzzy concept, and depends also on an individual scientist's subjective evaluation of any given model. For all that, it is a useful distinction and models can often be graded relative to one another with regard to the degree to which each defines process structure.

With these reservations in mind we consider the class of sophisticated guessing models to be descriptive but reasonably process-oriented, whereas the choice models, to be discussed later, have only recently been given process interpretations. Some potential processing versions of the most important choice model (for confusion experiments) are suggested below.

6.1. Sophisticated guessing models

Fig. 8 illustrates the basic simple configuration of the sophisticated guessing models. As in the contemporary feature-confusion models described above the observer acquires a confusion set representing the alternatives among which he is perceptually confused on a given trial, and, also as in the feature models, he then guesses from the reduced set of possibilities. Unlike the feature models though,

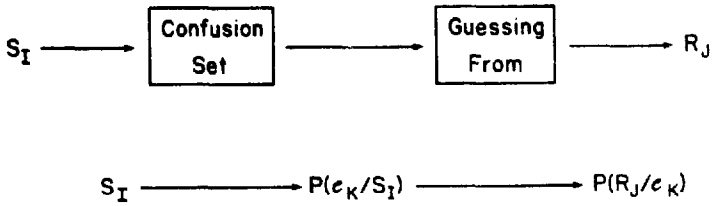

Fig. 8. The general schema for sophisticated guessing models.

there is no intermediate set of features or similarity measurement (not to mention the observation $\boldsymbol{W}$). A probability distribution is placed directly on the confusion sets, conditionalized on the presented stimulus pattern. That is, the probability that confusion set $\mathscr{C}_k$, out of 2^n possible confusion sets, arises from presentation of stimulus s_i (denoted by $P(\mathscr{C}_k|s_i)$) acts as a parameter in the sophisticated guessing models. This result in the response probabilities is given by

$$P(r_j|s_i)=\sum_k P(r_j|\mathscr{C}_k)P(\mathscr{C}_k|s_i).$$

The most general sophisticated guessing model would, in principle, require 2^n-1 sensory or perceptual parameters $P(\mathscr{C}_k|s_i)$ for each stimulus s_i. This follows from the fact the each confusion state is equivalent to a particular subset of the n stimuli, and there are 2^n such subsets. Because $\sum_{k=1}^{2^n} P(\mathscr{C}_k|s_i)=1$ there are 2^n-1 free parameters for each stimulus. Since the observer must guess from all n alternatives if either $\mathscr{C}_k=\emptyset$ or $\mathscr{C}_k=\{\text{the entire response set}\}$, these two states can be collapsed into a single state, reducing the number of sensory parameters to 2^n-2 for each stimulus. In addition, for every possible confusion set of k alternatives there would be $k-1$ guessing parameters, or $\binom{n}{k}(k-1)$ in total. (The guessing parameters must sum to unity, thus $k-1$ instead of k.) The total number of free guessing parameters is therefore

$$\sum_{k=1}^{n}\binom{n}{k}(k-1).$$

Combining the sensory and guessing parameters yields a grand total of

$$n(2^n-2)+\sum_{k=1}^{n}\binom{n}{k}(k-1)=3n2^{n-1}-2n-2^n+1$$

parameters in the general sophisticated guessing model.

It is apparent that the number of free parameters is huge relative to the number of degrees of freedom in an n-stimulus, n-response experiment, which is only $n(n-1)$. For example, when $n=3$ there are 23 free parameters for only 6 degrees of freedom!

The assumptions that are typically used to delimit this class of models into something more tractable are interesting, but not always especially intuitive. Although not much mathematical work on these models exists, it seems that three common assumptions are the following (e.g., Broadbent, 1967; Pachella, Smith and Stanowich, 1978; Smith, 1980; Townsend and Landon, 1982):

$$P(\mathscr{C}_k|s_i)=0 \quad \text{if } r_i\notin\mathscr{C}_k, \tag{6.1}$$

$$P(\mathscr{C}_k|s_i)=P(\mathscr{C}_k|s_j) \quad \text{where } r_i, r_j\in\mathscr{C}_k, \text{ except for } \mathscr{C}_k=\bigcup_{i=1}^{n} r_i \text{ or } \mathscr{C}_k=\emptyset, \tag{6.2}$$

$$P(\mathscr{C}_k|s_i)=P(\mathscr{C}_k|s_j) \quad \text{where } r_i, r_j\in\mathscr{C}_k \text{ for all } \mathscr{C}_k \text{ including } \mathscr{C}_k=\emptyset \text{ (can be substituted for (6.2)).} \tag{6.2'}$$

$$P(r_j|\mathscr{C}_k) = \frac{B_j}{\sum_{\{m|r_m \in \mathscr{C}_k\}} B_m}, \quad 0 \le B_j \le 1, \ \sum_{k=1}^{n} B_k = 1. \tag{6.3}$$

The interpretation of (6.1) is that any confusion set must contain the stimulus pattern that generated it, that is, any confusion set not containing the stimulus has probability 0 of occurrence. The second assumption, (6.2), is a symmetry constraint that requires the probability that a particular confusion state $\mathscr{C}_k$ arises to be the same no matter which stimulus in $\mathscr{C}_k$ was actually presented. The confusion set containing all or no stimuli is excepted. (6.2)′ is a strengthened version of (6.2) where the symmetry must hold even for the confusion state containing all or no members of the stimulus set. Assumption (6.3) makes the response bias of the jth response independent of the confusion state in which r_j is contained.

These restrictions point up some not-so-subtle differences that can arise between the sophisticated guessing models and the feature model approach discussed above. In the context of feature models where features may be gained (e.g., through intruded noise processes) or lost (assuming the gain or loss is all or none), the observer may or may not evaluate his feature sample in a manner consonant with (6.1), (6.2) or (6.2)′ and (6.3). Suppose features are gained and lost from presented stimuli but the observer *believes* that features are only lost. Then he may exclude the actual stimulus from the confusion set (on trials when features were gained) in violation of assumption (6.1). However, assumption (6.2) seems to be typically even more suspect. Consider the same example, with the letters F and E contained in the stimulus set. There $P[\mathscr{C}=\{E,F\}|E]$ will very likely be larger than $P[\mathscr{C}=\{E,F\}|F]$ because when the bottom part of the E is lost, both E and F might seem plausible to our hypothetical observer but when the bottom part is added (by noise) to bottom of an F, then the F may be excluded from the candidate confusion sets. Thus, (6.2) is violated. Other examples can easily be concocted. On the other hand the sophisticated guessing models are not restricted to feature interpretations and may capture interesting psychological structure outside of particularistic viewpoints.

Finally, the third assumption about guessing, (6.3), may be too strong but it is hard to know where else to start at the final decision end of the processes. The Bayesian approach mentioned above in conjunction with one of the feature models has never fared well with human recognition data. As a generalization of (6.3), the observer might revise his guessing probabilities in accordance with whatever confusion state has arisen. This could occur because, even with assumption (6.2) or (6.2)′, a stimulus s_i may give rise to a confusion state $\mathscr{C}_k = \{r_i, r_j, r_m, \ldots\}$ more often than a state $\mathscr{C}_k = \{r_i, r_m, \ldots\}$ which does not contain r_j, thus providing additional information on the likelihood of the occurance of s_i. Again, however, the number of parameters increases dramatically.

Despite these problems, several sophisticated guessing-type models have been proposed. They will be introduced individually.

6.2. The all-or-none model

The all-or-none model contains two confusion state possibilities, one of the states containing a single member, that member being the presented stimulus itself (Townsend, 1971a, b). The other type of state contains all or, equivalently in terms of information transmission, none of the stimulus alternatives. Thus, although it is not logically necessary for perception of a stimulus to be perfect in order that the confusion state contain only the stimulus pattern, sufficient information must have been processed to lead to that event. Similarly, some perception may have occurred when the state contains all of the alternatives, but the information processed cannot have lead to any decrease in the possibilities. A plausible example where the all-or-none model might be appropriate would be a stimulus set where each stimulus was made up of unique features plus features held in common by all of the stimuli. Another example is a strict template type of system where a perfect match occurs between the stimulus and the proper internal pattern or, if not, then guessing at random occurs. Of course, the guessing probabilities need not be uniform; biases can be exerted in the choices. Formally, the confusion set possibilities that have non-zero probability of occurrence are

$\mathscr{C}_i = \{r_i\}$: the complete information confusion set,

$\mathscr{C}_u = \bigcup_{k=1}^{n} r_k$: the zero information confusion set.

The conditional response probabilities for this model are

$$P(r_i|s_i) = N_i + (1-N_i)H_i, \qquad P(r_j|s_i) = (1-N_i)H_j, \quad i \neq j$$

where

$$0 \leq N_i, H_j \leq 1, \qquad \sum_{j=1}^{n} H_j = 1.$$

N_i is, of course, the probability of entering the confusion state $\{r_i\}$ following an s_i presentation and H_i is the probability of guessing r_i when in the zero-information state. It is not the case that $P(\mathscr{C}_u|s_i) = 1-N_i = P(\mathscr{C}_u|s_j) = 1-N_j$ $(i \neq j)$ in general. That is, assumption (6.2) is used rather than (6.2)'. The all-or-none model is probably the best known model of the multiplicative type where $P(r_j|s_i) = a_i \times b_j$. That is, the conditional response probability can be written as a product of a stimulus factor and a response factor. Not surprisingly, these models make very strong predictions. They are discussed in more detail in Falmagne (1972) and Townsend (1978).

6.3. The overlap model

The overlap model was developed as a slightly more complex sophisticated

guessing model which allows two-way confusion sets (Townsend, 1971a, b). It is still considered possible that 'perfect' information could be transmitted leading to a single member confusion state containing only the correct alternative, as in the all-or-none model. In contrast to the all-or-none model, however, no universal confusion state containing zero information was permitted. Instead, the observer can enter any two-way partial information confusion state, each of which contains just two stimulus (response) alternatives. The reason for this was to contrast a pure-guessing type of confusion structure with one based entirely on pairwise confusion.

The confusion state possibilities and the conditional response probabilities for the overlap model are

$$r_{ii} = \{r_i\}, \qquad r_{ij} = \{r_i, r_j\}, \quad i \neq j,$$

$$P(r_i|s_i) = E_{ii} + \sum_{\substack{k=1 \\ k \neq i}}^{n} E_{ik}\left(\frac{B_i}{B_i + B_k}\right), \qquad P(r_j|s_i) = \frac{E_{ij}B_j}{B_i + B_j}, \quad i \neq j$$

where

$$0 \leq E_{ij},\ B_j \leq 1, \qquad E_{ij} = E_{ji},$$

$$\sum_{j=1}^{n} E_{ij} = 1, \quad 1 \leq i \leq n \qquad \text{and} \qquad \sum_{k=1}^{n} B_k = 1.$$

The parameters E_{ij} are interpreted as similarity measures (i.e., the similarity between s_i and s_j) representing the likelihood that the pairwise confusion state r_{ij} arises from presentation of stimulus s_i. The parameters B_j are, of course, the guessing probabilities. Note that the overlap model obeys all of the assumptions outlined above ((6.1), (6.2)', (6.3)). Townsend and Landon (1982) present a proof (provided by Schulze (1973)) that the following conditions on probability entries in a confusion matrix are necessary and sufficient to permit a perfect fit by an overlap model:

(a) *Quasisymmetry*

$$P(r_j|s_i) \cdot P(r_k|s_j) \cdot P(r_i|s_k) = P(r_k|s_i) \cdot P(r_j|s_k) \cdot P(r_i|s_j).$$

(b) *Symmetric cells inequality*

$$P(r_j|s_i) + P(r_i|s_j) \leq 1.$$

(c) *Column constraint*

$$P(r_i|s_i) \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} P(r_i|s_j).$$

Such conditions as these permit the investigator, in principle, to test the model by observing if the conditions hold without estimating the model's parameters or comparing the model's numerical predictions against the data. Further, this type of analysis is an excellent example of where measurement theory can dovetail in an

advantageous way with process-oriented modeling. The resulting synthesis can explicate the constraints that a model puts on a data set, in this case a confusion matrix, and sometimes allow a non parametric test of the model, as alluded to just above (see Townsend and Landon, 1982).

6.4. The generalized overlap model

First suggested as a combination of the all-or-none model and the overlap model (Townsend, 1971a), this model includes all of the possible confusion states of the all-or-none model and the overlap model: single member states, pairwise states and a total confusion zero information state. These and the response probabilities can be formulated as follows:

$$\mathscr{C}_{ii} = \{r_i\}, \qquad \mathscr{C}_{ij} = \{r_i, r_j\}, \quad i \neq j, \qquad \mathscr{C}_u = \bigcup_{k=1}^{n} r_k,$$

$$P(r_i|s_i) = E_{ii} + \sum_{\substack{j=1 \\ j\neq i}}^{n} E_{ij} \frac{B_i}{B_i + B_j} + \left(1 - \sum_{j=1}^{n} E_{ij}\right) B_i,$$

$$P(r_j|s_i) = E_{ij} \frac{B_j}{B_i + B_j} + \left(1 - \sum_{j=1}^{n} E_{ij}\right) B_j, \quad i \neq j$$

where

$$0 \leq E_{ij} = E_{ji}, \; B_j \leq 1,$$

$$\sum_{j=1}^{n} B_j = 1, \qquad \sum_{j=1}^{n} E_{ij} \leq 1, \quad 1 \leq i \leq n.$$

The parameters here are interpreted in the same fashion as for the overlap model. The generalized overlap model follows assumptions (6.1), (6.2) and (6.3) but not, in general, (6.2)′.

6.5. The informed guessing model

Pachella, Smith and Stanovich (1978) explore and apply a case of the generalized overlap model in which it is assumed that $1 - \sum_{j=1}^{n} E_{ij} = \alpha$, a constant for all $1 \leq i \leq n$. That is, this model obeys assumption (6.2)′ rather than just (6.2) and assumes that the probability of entering the zero information state is the same for all stimuli.

More formally,

$$P(r_i|s_i) = E_{ii} + \sum_{\substack{k=1 \\ k\neq i}}^{n} E_{ik} \frac{B_i}{B_i + B_k} + \alpha B_i,$$

$$P(r_j|s_i) = E_{ij} \frac{B_j}{B_i + B_j} + \alpha B_j, \quad i \neq j$$

where

$$\alpha + \sum_{k=1}^{n} E_{ik} = 1, \qquad 0 \le E_{ik} = E_{ki} \le 1, \qquad 0 \le B_j \le 1, \qquad \sum_{i=1}^{n} B_i = 1.$$

Once again, the parameters are interpreted in the same fashion as for the overlap model. It should be pointed out, however, that the parameters of this model are not completely identifiable. That is, a given set of data may be satisfied by more than one set of values of the parameters in the model (see Restle and Greeno (1970) for an introductory account of parameter identifiability). It follows that the generalized overlap model's parameters are also not identifiable.

Schulze (1978) has demonstrated that the informed guessing model holds under the following confusion matrix measurement conditions:

(1) Quasisymmetry (see (a) above).

(2) $$\min_{i,j} \left\{ \frac{P(r_j|s_i) + P(r_i|s_j)}{d_i + d_j} \right\} \ge \max_{i} \left\{ \frac{P_i - 2P(r_i|s_i)}{(n-2)d_i} \right\}$$

where

$$d_i = P(r_i|s_k) \Big/ \sum_{j=1}^{n} \frac{P(r_j|s_k)}{P(r_k|s_j)} \quad \text{for a fixed } k$$

and

$$P_i = \sum_{j=1}^{n} P(r_i|s_j)$$

under the obvious condition that none of the respective denominators are zero. The proof of this assertion will not be presented here. It is available from the present authors upon request.

6.6. *Symmetric sophisticated guessing model*

Smith (1980) has presented and tested what he terms the symmetric sophisticated guessing model. This model is simply a straightforward application of the general sophisticated guessing model outlined above with constraints (6.1), (6.2)′ and (6.3). That is, the confusion state possibilities consist of *any* subset of the stimulus set, which number $2^n - 1$. (The null set or complete stimulus set state are collapsed into a single zero information state.) The conditional response probabilities are

$$P(r_j|s_i) = \sum_{r_{ij}} \left[P(r_{ij}|s_i) \cdot \frac{B_j}{\sum_{(k\ r_k \in r_{ij})} B_k} \right], \quad i \ne j,$$

$$P(r_i|s_i) = 1 - \sum_{\substack{j=1 \\ j \ne i}}^{n} P(r_j|s_i)$$

where

$$\mathscr{C}_{ij} = \{r_i, r_j, \ldots\} = \text{any confusion state containing at least responses } r_i \text{ and } r_j.$$

Although the parameters $P(\mathscr{C}_{ij}|s_i)$ are constrained by the boundary condition

$$\sum_{\mathscr{C}_i} P(\mathscr{C}_{ij}|s_i) = 1 \quad \text{for all } s_i,$$

there remain $2^n - n - 1$ parameters of the form $P(\mathscr{C}_{ij}|s_i)$ and $n-1$ guessing parameters to be estimated for only $n(n-1)$ degrees of freedom.

Despite this obstacle Smith (1980) has shown that the symmetric sophisticated guessing model is testable through the measurement constraints

(1) quasisymmetry (see (a) above),

(2) column constraint (modified),

$$P(r_j|s_j) \geq P(r_j|s_i) \quad \text{for all } s_i, s_j.$$

6.7. The confusion-choice model

This model is somewhat difficult to classify in the present scheme. It bears affinity with the sophisticated guessing models, and so it will be presented in this section (despite its name which suggests a different set of models treated subsequently), but makes more detailed assumptions about how a confusion set of alternatives is arrived at. Yet, these assumptions concern distance computations in a multidimensional space rather than explicit feature formulations, so the model does not easily fit into the substantive feature-processing mold. It is a rather complex model and typically requires rather formidable computer time for its implementation. The reader is referred to the original paper by Nakatani (1972) on which it is based for a detailed treatment. The original model will be outlined in a qualitative way here.

Basically, the idea is that the stimulus and its associated response can be represented by a *single* 'ideal' point in a Euclidean space. When a stimulus s_i is presented, an observation occurs which itself is a point in the same space. The observation point is a random event determined partly by the deterministic (fixed) distances of the 'ideal' point of s_i from all of the other stimuli. To these fixed distances are added independent normally distributed noise sources (one for each distance), each with zero mean and a variance of 1. Each response alternative point (which is coincident with the stimulus 'ideal' point) has associated with it a threshold t_j (for response r_j). If (and only if) the distance of the observed point is less than the threshold associated with any particular response, that response is placed into the confusion state. The probability that an observed point arising from stimulus s_i falls within the acceptance region of response (stimulus) r_j, denoted by L_{ij}, is then

$$L_{ij} = \int_{-\infty}^{t_i - u_{ij}} \phi(z)\,dz$$

where $\phi(\cdot)$ is the normal density function with mean $= 0$ and $\sigma = 1$ and u_{ij} is the Euclidean distance between s_i and r_j. That is, L_{ij} represents the probability that r_j will be an acceptable response (i.e., enter the confusion state) if s_i is presented. Finally the observer chooses the response from among the candidates belonging to the confusion state in accordance with assumption (6.3) of the sophisticated guessing models. That is, each response is associated with a bias strength which is independent of the given confusion state. Its probability of occurrence is given by the ratio of strengths form as in the other sophisticated guessing models presented above.

Note that this model might be interpreted as a feature processing model in cases where the dimensions of the Euclidean space are reasonably described as orthogonally related features. The following may also be ascertained:

(1) The confusion-choice model does not obey sophisticated guessing assumption (6.1), which requires that the presented stimulus always be a member of the confusion state, because there is some finite probability that the observation point for s_i will exceed the threshold for the 'ideal' point of s_i.

(2) The assumptions (6.2) and (6.2)′ are also not satisified, because (a) it may be that $t_j \neq t_k$, that is, two thresholds are not equal so the probability that r_j is included in the confusion state given presentation of s_k may not equal the probability that r_k is included in the confusion state given s_j, and (b) because the distance from s_j to any s_m will not typically equal the distance from s_k to s_m. The probability of the same confusion state given either s_j or s_k will be different in the two cases, even when both r_j and r_k are included in the confusion state.

(3) Finally, condition (6.3) *is* satisfied as noted.

Thus, the confusion-choice model, may, like the feature models, be represented as a process-oriented model, which shares some primitive characteristics with sophisticated guessing models (e.g., the generation of a confusion set, etc.) but not the more constraining assumptions.

Townsend and Landon (1982) have developed and tested a special case of the Nakatani (1972) confusion-choice model. Although it starts at a more macroscopic level than does the original model, and thus loses some of the intricate processing explanation, it redeems itself in its testability against the other similar models of confusion discussed in this paper, and the significant gains made in the ease with which its parameters can be estimated (any extant estimation procedure, e.g., Chandler's (1954) STEPIT routine, may be used). It is, essentially, a general sophisticated guessing model that only uses constraints (6.1) and (6.3), and yet retains the identifiability of its parameters.

The confusion state possibilities consist of *any* subset of the stimulus set, as in the symmetric sophisticated guessing model. The conditional response probabilities are developed as follows:

(1) The acceptance matrix, $\boldsymbol{A}$, is defined as a finite $n \times 2^n$ matrix of probabilities

P_{ik}, where P_{ik} is defined as

$$P_{ik} = P(\gamma_k | s_i) = \prod_{\{j : s_j \in \gamma_k\}} L_{ij} \prod_{\{m : s_m \notin \gamma_k\}} (1 - L_{im})$$

with the constraints that $0 < L_{ij} = L_{ji} < 1$ and $L_{ii} = L_{jj} = 1$, and where $\gamma_k = \{r_i, r_j, \ldots\}$ = any confusion state containing at least responses r_i and r_j. That is, P_{ik} represents the likelihood that confusion state γ_k arises from presentation of s_i, defined in terms of the L_{ij} acceptance probabilities. The constraint that $L_{ii} = 1$ will set the probability $P_{ik} = 0$ whenever the response for the presented stimulus s_i is not included in the confusion set γ_k.

(2) The decision matrix, $\boldsymbol{D}$, is defined as a finite $2^n \times n$ matrix of probabilities Q_{kj}, where each Q_{kj} is defined as

$$Q_{kj} = P(r_j | \gamma_k) = V_j / \sum_{(m : r_m \in \gamma_k)} V_m$$

with the constraints that $0 < V_j < 1$ and $\sum_{j=1}^{n} V_j = 1$. When the confusion state $\gamma_\cdot$ is the null set, $Q_{kj} = V_j$ and when r_j is not a member of the confusion state, $Q_{kj} = 0$. V_j is, of course, interpreted as the response bias for response r_j.

(3) From the above definitions of $\boldsymbol{A}$ and $\boldsymbol{D}$, the conditional response probabilities are given by

$$\boldsymbol{J} = \boldsymbol{A} \cdot \boldsymbol{D} \quad \text{which entails that} \quad P(r_j | s_i) = \sum_k P_{ik} Q_{kj}$$

where $\boldsymbol{J}$ represents a finite confusion matrix. The index k in the summation ranges over the possible confusion states that can arise from the stimulus set. (Note that a simple matrix multiplication is being performed.)

In this special case of the confusion-choice model, the L_{ij} acceptance probabilities are considered as free parameters along with the guessing probabilities. There are $\frac{1}{2}(n(n-1))$ free L_{ij} parameters and $(n-1)$ guessing parameters, which is within the degrees of freedom of any confusion matrix (i.e. $n(n-1)$).

7. Choice models

A very different breed, at least in conception, are the choice models. The original development of the choice model was pursued in the context of measurement theory and by way of placing conditions on the probabilistic structure in choice and decision situations (Luce, 1959); these resulted in the well-known choice axiom (not to be confused with the set-theoretic axiom of the same name). The ideas have since been employed in perceptual recognition situations (e.g., Luce, 1963; Townsend, 1971a, b).

The fundamental outcome of the derivation of the choice model was that the response probability can be expressed as a ratio of the response strength of the considered alternative to the sum of the response strengths of the other alternatives.

It is rare to find use of tests of the choice axiom itself in the recognition context whence the response strength formulation arose (exceptions are Hodge and Pollack, 1962; Townsend and Landon, 1982), and that general formulation may be derived in a number of other ways than from the choice axiom (e.g., Holman, 1979). In any event models of this type continue to be called choice models, primarily because an interpretation of the response strengths in terms of both a stimulus contribution and response contribution was put forth early on by Luce (1959, 1963).

We shall work with the 1963 version (from Luce, 1963) where the overall strength in favor of alternative r_j can be given as a product of a positive similarity of the stimulus presented (s_i) to r_j (U_{ij}), and a bias-strength factor (Y_j). It will be referred to here as the *similarity choice model* (also known as the biased choice model; e.g., Pachella, Smith and Stanovich, 1978). In addition, certain constraints are placed on the parameters; for example, the similarity between a stimulus and itself is equal to one for all stimuli. The conditional response probabilities are then given as

$$P(r_j|s_i) = \frac{U_{ij} Y_j}{\sum_{k=1}^{n} U_{ik} Y_k},$$

$$0 \leq U_{ij} \leq 1,\ 0 \leq Y_j,\ U_{ii} = U_{jj} = 1,\ i,j = 1, 2, \ldots, n.$$

In the event that U_{ij} can be decomposed into a product $U_{ij} = U_i U_j$, a multiplicative confusion model results which is equivalent to the all-or-none model (discussed above) for a single given confusion matrix (Townsend, 1978). Lappin (1978) briefly discusses a possible generalization of the similarity choice model which occurs when the bias parameters are allowed to depend on a pair of responses.

The similarity choice model is well known and has been widely investigated. Shipley and Luce (1964) had some success in fitting the model to their data from discriminations of different weights. Townsend (1971a) and Townsend and Ashby (1976) found that the model performed better than several other confusion models in two separate experiments, one using the upper case English alphabet and the other a synthetic four letter alphabet with varied payoffs. Lupker (1979) and Smith (1980) reported that the choice model predicted their data quite well. Most recently, Townsend and Landon (1982), in an extensive test of both the choice model and the underlying choice axiom, found the choice model to do quite well in predicting their confusion matrices. In fact, despite the more detailed (to date) processing underpinnings, the feature confusion models discussed earlier in this paper have, overall, not tended to predict recognition behavior as well as the similarity choice model. Although occasionally certain sophisticated guessing models vie with the choice model as theoretical alternatives, so far the latter has amassed the most support.

7.1. Processing interpretations of the similarity choice model

Because of the consistent success exhibited by the choice model in outperforming other models in experimental settings, the question of possible interpretations arises for those inclined to information processing characterizations of internal mechanisms. As it turns out, with a little thinking, the problem is not too few interpretations but a superfluity, with little in the way of critical data to decide between them. To be sure, some are more intuitive than others, and some possess defects in structure or generality.

The major aspects of the choice model, as mentioned, are twofold: (i) the probability of any given response alternative can be written in a ratio of strengths form, and (ii) the strength of any alternative is given as a product of a similarity factor and a response bias factor. The main questions, then, concern processing notions that lead to the two separate factors and ways in which the ratio of strengths form of the choice probabilities could arise from processing mechanisms.

One idea to pursue would be to investigate whether this model might be represented in a probabilistic decision space similar to those discussed for discriminant models or in the context of signal detectability theory. Marley (1971, p. 576) has shown that, for the similarity choice model to hold, when there are only two stimuli, the form of any existing probability distribution must be logistic and the decision criterion must be a function both of the choice sensory similarity parameters as well as the bias parameters. Thus, even for $n=2$, the distributions obtained do not readily relate to separate processing stages for sensory and bias factors. When the number of stimulus alternatives is greater than or equal to three, there is no adequate distributional representation at all on a single dimension. Probably a multidimensional representation is needed for $n \geq 3$.

Once one allows the strengths formulated as a product to be taken as given ($U_{ij} \cdot Y_j$), one may refer to the growing body of work relating the ratio form of the choice probabilities to underlying distributions and to general questions of relationships holding between the choice model and the Thurstonian discriminal processes model. Details will not be provided here, but we shall note for the interested reader some of the pertinent references.

The seminal work of Bradley and Terry (1952) in a paired-comparison context provided the impetus for the ratio of strengths form. Soon after Bradley (1953) developed a distribution approach and later (1965) a more complete rationale. Thompson and Singh (1967) present a derivation of the ratio form from asymptotic theorems related in a loose way to semi-neurological notions. More recently, Yellott (1977) has offered a treatment of equivalence relations between Thurstone's and Luce's choice axioms. These, of course, do not directly deal with recognition, or in particular, the similarity choice model, although they do provide avenues of approach.

Grossberg (1969) has developed a neural model for learning with structure quite similar to the 1963 choice model. A measure of association strength between items s_j

and s_k normalized relative to the association of s_j with all the pertinent items, is given in his equation (2) as a function of time t by

$$A_{jk}(t) = O_{jk} Z_{jk}(t) \Big/ \sum_{m=1}^{n} O_{jm} Z_{jm}(t).$$

The C_{jm} terms represent the ease of neural flow between nodes j and m and the $Z_{jm}(t)$ terms represent the amount of associational learning between j and m that has taken place by time t. In transplanting this framework to the letter recognition situation, suppose $O_{jm} = O_m$ for $1 \leq j \leq n$ stands for the bias factor; for instance, it might refer to the neural path from the comparison process to the response r_m. Let $Z_{jm}(t)$ be the neural similarity between s_j (the presented stimulus letter) and r_m (a memory letter) growing as a function of time. We would now assume that at some time $t = T$, the observer makes response r_k with probability $A_{jk}(T) = C_{jk}(T)$, the confusion probability. Although this approach appears to hold some promise it must be noted that in a sense, the problem has just been pushed back one step, for the question as to how the response selection actually gets stochastically generated (with the appropriate probability) has still not been specified. Rather we have simply gained a deterministic normalized measure of response strength.

In the next consideration regarding direct processing interpretations, a simplistic but suggestive parallel model which relates the ratio form to processing latency will be sketched out as if the stimulus set was the upper case English alphabet. It will be referred to as the minimum latency choice model and is presented in Fig. 9.

During the sensory phase the magnitude of the similarity between the presented pattern and each of the memory possibilities is established and, following this, the response bias factor or gain is entered multiplicatively. Each memory alternative has a direct connection with its response in the response selection stage. It is assumed that response is generated which is activated first, with each alternative being

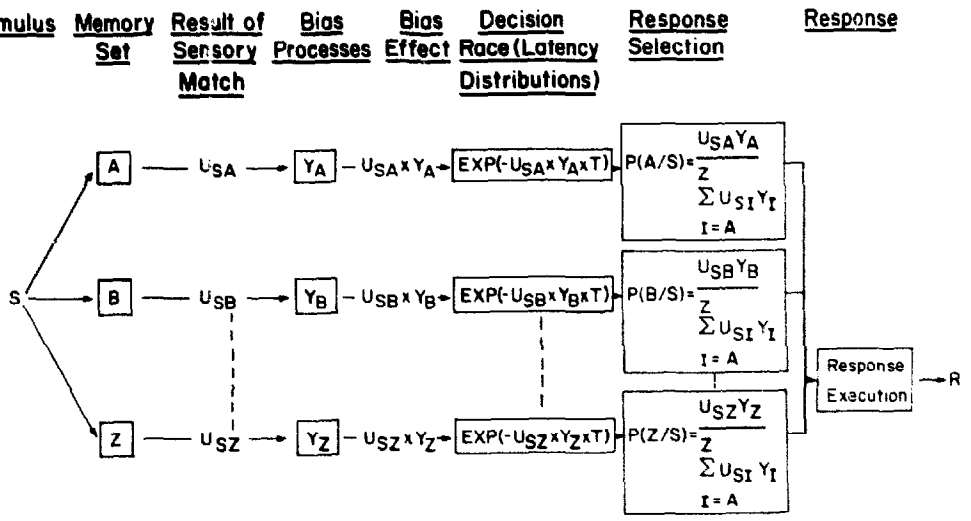

Fig. 9. Schema for the minimum latency choice model.

associated with an exponential probability distribution with rates given by the overall strengths. One minus each of the distribution functions (the so-called survivor function) is shown in Fig. 9 using boxes for brevity.

It can be seen that the form of response probabilities appears in the final boxes. We must observe that the similarity and bias parameters, representing processing outcomes at two different stages here, are treated as deterministic, although it may be that U and Y could be viewed as random variables and still maintain the choice response probabilities as approximations.

The final processing interpretation to be presented is the maximal match choice model (Townsend, Evans and Hu, 1970; Townsend and Landon, 1982). The previous interpretation was based on the *minimum* processing latency, whereas the present one is based on the *maximal match* (overlap, etc.) of the input to the memory patterns. It utilizes an idea of Holman and Marley (cited in Luce and Suppes, 1965), and later Yellott (1977), showing how the double exponential distribution can lead to the ratio-of-strengths formulation. A detailed account of the assumptions of this interpretation and a proof leading to the ratio of strengths form can be found in Townsend and Landon (1982).

An outline of the maximal match choice model for $n = 3$ is given in Fig. 10. As can be seen, the input stimulus s_i is compared (matched) with the memorial

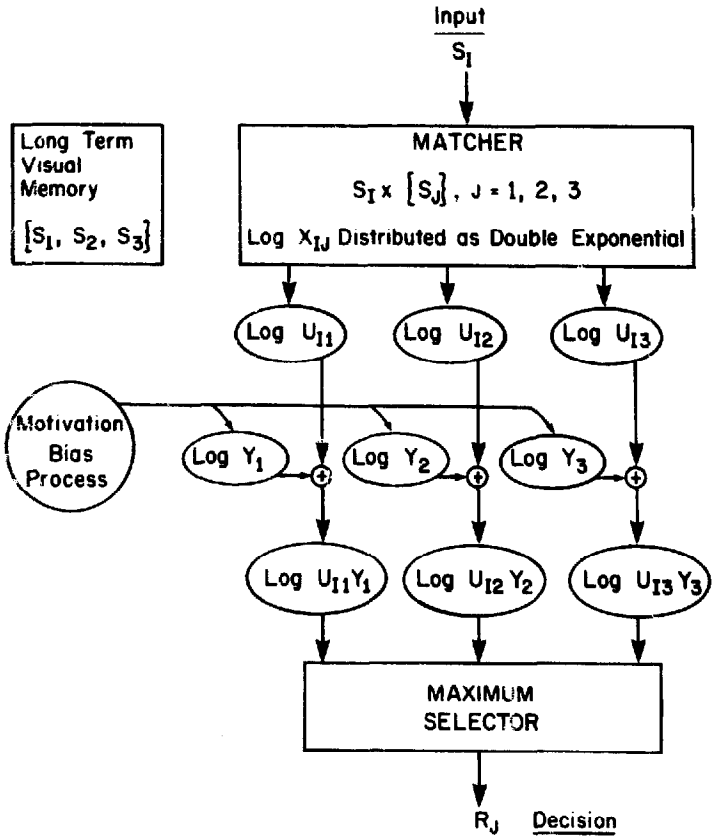

Fig. 10. Schema for the maximal match choice model.

representations. The output of the match is $\log X_{ij}$ where $Z_{ij} = \log X_{ij}$ is distributed as the double exponential, with $\log U_{ij}$ the central parameter. The mean of the bias process is $\log Y_j$ and is added to the output of the matcher. The maximum selector then chooses the appropriate maximal value of $\log U_{ij} Y_j$ as the response. Referring back to Fig. 5, it can be readily seen that the maximal match choice model interpretation is actually a form of discriminant theory.

7.2. Relationships between the similarity choice model and sophisticated guessing models

Unless two formulations are the identical model, there will exist experimental situations where the two will make distinct predictions. However, in many instances, a single experimental confusion matrix cannot decide between two alternate models. For example, it is easy to show that the similarity choice model can predict any single set of data (that is, a single sample confusion matrix) that the all-or-none model can; the latter is mathematically a special case of the former (Townsend and Landon, 1982). However, since the predicted $P(r_j | s_i)$ are distinct functions of different parameter sets in the two models, and because the various parameters are presumed to be functions of different aspects of the environment, experimental conditions can be varied and unique predictions thereby generated. For instance, the stimuli or sensory parameters are usually taken to be functions of intensity, observer sensitivity and the like, whereas the bias parameters are believed to be functions of stimulus presentation frequencies, payoff conditions and such manipulations. Then, when the sensory parameters of one model are written as a function of the parameters of the other, it turns out that the result depends on both the sensory as well as the bias parameters of the other model. This demonstrates that both models cannot predict invariance in, say, the bias parameters, when it is only the stimulus or sensory variables that are experimentally altered. Notice again that such manipulations produce different confusion matrices. The following remarks are confined to the 'space' occupied by a single confusion matrix.

It can be shown that the following two measurement conditions are necessary and sufficient for any single confusion matrix to be represented by the similarity choice model (e.g., Townsend and Landon, 1982):

(1) quasisymmetry (see (a) above),

(2) diagonal maximization,

$$P(r_i | s_i) P(r_j | s_j) \geq P(r_j | s_i) P(r_i | s_j).$$

Townsend and Landon (1982) also show that the similarity choice model can predict any single confusion matrix predictable by any sophisticated guessing model satisfying the sophisticated guessing assumptions (6.1), (6.2)′ and (6.3). This works because such sophisticated guessing models satisfy the two necessary and sufficient conditions given above for the similarity choice model. However, a sophisticated guessing model satisfying assumption (6.2) but not (6.2)′ is not a special case of the

similarity choice model as it does not in general predict quasisymmetry (which is necessary for the choice model; a counterexample to show this is easily constructed), but does predict that

$$P(r_j|s_j) \geq P(r_j|s_i), \quad i \neq j \text{ for all } s_i,$$

which is not true of the choice model.

From these considerations it follows that the overlap model and informed guessing model are special cases of the similarity choice model within the context of a single confusion matrix. Although the all-or-none model does not, in general, satisfy (6.2)', it is still a special case of the similarity choice model (Townsend, 1978). The possibility that there exist more general sophisticated guessing models that contain the similarity choice model remains open.

Finally, Townsend and Landon (1982) suggest that their special case of Nakatani's (1972) confusion-choice model (discussed earlier) may include the similarity choice model, although they were unable to provide a proof. However, they have shown that their special case of the Nakatani model is *not* a special case of the similarity choice model.

8. Summary and discussion

This paper has presented an introduction and survey of some of the mathematical theory of certain specific branches of psychological pattern recognition theories along with an attempt at a taxonomy of those regions covered. The treatment is not exhaustive when one includes theories not yet treated very mathematically or those not yet actually tested against any data. For example, one interesting theory that has not been rigorously applied in psychology is the so-called structural approach in which patterns are given more or less abstract descriptions in terms of 'linguistic' or grammatical relations within the patterns (e.g., Fu, 1974). Reed (1973) considers this and certain other topics not seen here in a qualitative manner.

An area that may ultimately bear importantly on psychological and/or physiological models of human pattern recognition is that of computational geometry (e.g., Minsky and Papert, 1969). That branch of research, defined broadly, seeks to determine the computational 'power' of certain classes of pattern recognition systems in terms of types of geometrical properties (for example, number of angles, congruance, connectedness) they are or are not capable of recognizing. So far it is not in a state that can have much impact on psychological/behavioral models.

As noted earlier even the classes of models that we have discussed differ greatly in the degree as well as the fashion in which they have been applied in psychology. The kinds of models emphasized in the early sections, discriminant models and statistical decision models, have provided a fruitful milieu in which to discuss potential quantitative and qualitative explanations of how persons go about perceiving

patterns of stimulation. However, they have not received much in the way of testing, particularly in the form of parameterized models fit to experimental data. They have received somewhat more attention in the area of categorization where there is a many-one mapping from a set of stimuli to a smaller set of responses, frequently with only two responses (e.g., Rodwan and Hake, 1964; Aiken and Brown, 1971). Even there, it has not usually been feasible to test them directly.

In contrast, the models on which we focused in the latter part of the paper have been typically employed in the n stimuli, n response (with the 1–1 mapping) context and have usually been tested by estimating their parameters and then comparing predicted correct and confusion frequencies with the experimental values (e.g., Geyer and Dewald, 1973; Townsend, 1971a, b; Nakatani, 1972; Wandmacher, 1976; Townsend and Landon, 1982; Rumelhart, 1971).

It would be easy to deride many, if not all, of the models of this paper as being too simplistic. Certainly, even the best of them must fall short of a definitive explanation of pattern recognition, especially in cognitively 'rich' environs where the vagaries of memory, thought, the several perceptual invariances and even emotion may assume importance. At some point attempts will have to be made at incorporating these and similar concepts into mathematical theories of human pattern recognition.

However, it would, we think, be wrong to fail for that reason to pursue the search for an adequate parsimonious mathematical theory that applies in the constricted laboratory context. Very global and complex theoretical structures will not likely be able to provide a very elegant or precise description or explanation of the detailed data at this fine level. Even when they will do so, it will probably be through an advancement of the type of models considered here. We are, in short, attempting an argument for something like a psychophysics of pattern recognition, an expansion from the elementary two stimulus, two response paradigm or the classical scaling domains, but not so large a leap that we leave behind analytically straightforward models. Even the rigorous empirical probing of the models present herein promises to be a challenging endeavor.

Acknowledgement

The authors are grateful for helpful comments made on earlier versions of this manuscript by Eric Holman, Steve Lupker, Joseph Lappin, Uwe Mortensen, Maria Nowakowska, Steven Reed, Han-Henning Schulze, Richard Schweickert. Also we wish to thank Julie McKinsie and Fran Teer for help in preparing the article. Its final stage of production occurred while the first author was a Visiting Scholar at the School of Social Sciences, University of California, Irvine.

References

L.S. Aiken and D.R. Brown, A feature utilization analysis of the perception of pattern class structure, Perception and Psychophys. 9 (1971) 270–283.

J.L. Baird and E. Noma, Fundamentals of Scaling and Psychophysics (Wiley, New York, 1978).

B.G. Batchelor, Pattern Recognition: Ideas in Practice (Plenum Press, New York, 1978).

R.A. Bradley and M.E. Terry, Rank analysis of incomplete block designs Part I. The method of paired comparisons, Biometrika 39 (1952) 324–345.

R.A. Bradley, Some statistical methods in taste testing and quality evaluation, Biometrics 9 (1953) 22–38.

R.A. Bradley, Another interpretation of a model for paired comparisons, Psychometrika 30 (1965) 315–318.

D.E. Broadbent, Word frequency effect and response bias, Psych. Rev. 74 (1967) 1–15.

J.P. Cavanagh, Holographic processes relizable in the neural realm: Predictions of short term memory performance, Ph.D. Dissertation, Carnegie-Mellon University, Pittsburgh, 1972.

J.P. Cavanagh, Size and position invariance in the visual system, Perception 7 (1978) 167–177.

J.P. Chandler, STEPIT: Finds local minima of a smooth function of several parameters, Behavioral Sci. 14 (1954) 81–82.

L.A. Cooper and R.N. Shepard, Chronometric studies of the rotation of mental images, in: W.G. Chase, ed., Visual Information Processing (Academic Press, New York, 1973).

T.N. Cornsweet, Visual Perception (Academic Press, New York, 1970).

T.M. Cowan, Organizing the properties of impossible figures, Perception 6 (1977) 41–56.

W.B. Davenport and W.L. Root, Random Signals and Noise (McGraw-Hill, New York, 1958).

P.C. Dodwell, Visual Pattern Recognition (Holt, Rinehart & Winston, New York, 1970).

J.C. Falmagne, Biscalability of error matrices and all-or-none reaction-time theories, J. Math. Psych. 9 (1972) 206–224.

K.S. Fu, Statistical pattern recognition, in: J.M. Mendel and K.S. Fu, eds. Adaptive, Learning, and Pattern Recognition Systems (Academic Press, New York, 1970).

K.S. Fu, ed., Syntactic Methods in Pattern Recognition, (Academic Press, New York, 1974).

L.H. Geyer and C.G. DeWald, Feature lists and confusion matrices, Perception and Psychophys. 14 (1973) 471–482.

D.M. Green and T.G. Birdsall, Detection and recognition, Psych. Rev. 85 (1978) 192–206.

D.M. Green and J.A. Swets, Signal Detection Theory and Psychophysics (Wiley, New York, 1966).

S.J. Grossberg, Embedding fields: A theory of learning with physiological implications, J. Math. Psych. 6 (1969) 209–239.

C.S. Harris, ed., Visual Coding and Adaptability (Lawrence Erlbaum Associates, Hillsdale, NJ, 1980).

M.H. Hodge and I. Pollack, Confusion matrix analysis of single and multidimensional auditory displays, J. Experimental Psych. 63 (1962) 129–142.

W.C. Hoffman, The lie algebra of visual perception, J. Math. Psych. 3 (1966) 65–98.

E.W. Holman, Monotonic models for asymmetric proximities, J. Math. Psych. 20 (1979) 1–20.

W. James, The Principles of Psychology (Holt, Rinehart & Winston, New York, 1890) (reprinted by Dover Publications, 1950).

L. Kaufman, Sight and Mind: An Introduction to Visual Perception (Oxford University Press, New York, 1974).

G. Keren and S. Baggen, Recognition models of alphanumeric characters, Perception and Psychophys. 29 (1981) 234–246.

W. Köhler, Gestalt Psychology (Liveright, New York, 1947).

P.A. Kolors and M. Eden, eds., Recognizing Patterns: Studies in Living and Automatic Systems (MIT Press, Cambridge, 1968).

C.L. Krumhansl, Concerning the applicability of geometric models to similarity: The interrelationship between similarity and spatial density, Psych. Rev. 85 (1978) 445–463.

M. Kubovy and P. Podgorny, Does pattern matching require the normalization of size and space? Perception and Psychophys. 30 (1981) 24–28.

J.S. Lappin and W.R. Uttal, Does prior knowledge facilitate the detection of visual targets in random noise? Perception and Psychophys. 20 (1976) 367–374.

J.S. Lappin, The relativity of choice behavior and the effect of prior knowledge on the speed and accuracy of recognition, in: N.J. Castellan and F. Restle, eds., Cognitive Theory Vol. 3 (LEA, Hillsdale, NJ, 1978).

R.D. Luce, Individual Choice Behavior (Wiley, New York, 1959).

R.D. Luce, Detection and recognition, in: R.D. Luce, R.R. Bush and E. Galanter, eds., Handbook of Mathematical Psychology Vol. 1 (Wiley, New York, 1963).

R.D. Luce and P. Suppes, Preference, utility, and subjective probability, in: R.D. Luce, R.R. Bush and E. Galanter, eds., Handbook of Mathematical Psychology Vol. 3 (Wiley, New York, 1965).

S.J. Lupker, On the nature of perceptual information during letter perception, Perception and Psychophys. 25 (1979) 303–312.

A.A.J. Marley, Conditions for the representation of absolute judgment and pair comparison isosensitivity curves by cumulative distributions, J. Math. Psych. 8 (1971) 554–590.

C. Martindale, Cognition and Consciousness (The Dorsey Press, Homewood, IL, 1981).

C.D. McGillem and G.R. Cooper, Continuous and Discrete Signal and System Analysis (Holt, Rinehart & Winston, New York, 1974).

J.M. Mendel and K.S. Fu, eds., Adaptive Learning, and Pattern Recognition Systems (Academic Press, New York, 1970).

M.L. Minsky and S. Papert, Perceptrons: An Introduction to Computational Geometry (MIT Press, Cambridge, MA., 1969).

R. Mizoguchi, M. Shimura and O. Kakosho, A new algorithm for constructing piecewise linear discriminant functions, in: Proc. IEEE 5th Internat. Conf. on Pattern Recognition Vol. 1 (IEEE Computer Society, New York, 1980) pp. 666–670.

U. Mortensen, Models of pattern recognition by feature identification and similarity assessment, reprinted from Diskussionsbeitrage Des Fachbereichs Wirtschaftswissenshaften und Statistik Der Universitat Konstanz, 1978a.

U. Mortensen, The activation of features in visual pattern recognition processes, reprinted from Diskussionsbeitrage Des Fachbereichs Statistik Der Universitat Konstanz, 1978b.

L.H. Nakatani, Confusion-choice model for multidimensional psychophysics, J. Math. Psych. 9 (1972) 104–127.

U. Neisser, Cognitive Psychology (Prentice-Hall, Englewood Cliffs, NJ, 1967).

N.J. Nilsson, Learning Machines: Foundations of Trainable Pattern Classifying Systems (McGraw-Hill, New York, 1965).

R.G. Pachella, J.E.K. Smith and K.E. Stanovich, Qualitative error analysis and speeded classification, in: N.J. Castellan and F. Restle, eds., Cognitive Theory Vol. 3 (Lawrence Erlbaum Associates, Hillsdale, NJ, 1978).

K. Pribram, Languages of the Brain (Prentice-Hall, Englewood Cliffs, NJ, 1971).

F. Rahber and D.F. Mix, Pattern recognition based on piecewise linear as quadratic discriminant functions, in: Proc. IEEE 5th Internat. Conf. on Pattern Recognition Vol. 1 (IEEE Computer Society, New York, 1980) pp. 674–676.

S.K. Reed, Psychological Processes in Pattern Recognition (Academic Press, New York, 1973).

F. Restle and J.G. Greeno, Introduction to Mathematical Psychology (Addison-Wesley, Menlo Park, CA, 1970).

A.S. Rodwan and H.W. Hake, The discriminant function as a model for perception, Amer. J. Psych. 79 (1964) 380–392.

D.E. Rumelhart, A multicomponent theory of the perception of briefly exposed visual displays, J. Math. Psych. 7 (1970) 191–218.

D.E. Rumelhart, A multicomponent theory of confusions among briefly exposed alphabetic characters,

Tech. Rept. No. 22, Center for Human Information Processing, University of California, San Diego, 1971.

D.E. Rumelhart and D. Siple, Process of recognizing tachistoscopically presented words, Psych. Rev. 81 (1974) 99–118.

H.H. Schulze, Personal communication, 1973.

H.H. Schulze, Personal communication, 1978.

O.G. Selfridge, O.G. Pandemonium: A paradigm for learning, in: Symp. on the Mechanization of Thought Processes (HM Stationary Office, London, 1959).

E.F. Shipley and R.D. Luce, Discrimination among two- and three-element sets of weights, in: R.C. Atkinson, ed., Studies in Mathematical Psychology (Stanford University Press, Stanford, CA, 1964).

J.E.K. Smith, Models of identification, in: R.S. Nickerson, ed., Attention and Performance Vol. VIII (Lawrence Erlbaum Associates, Hillsdale, NJ, 1980).

W.A. Thompson and J. Singh, The use of limit theorems in paired comparison model building, Psychometrika 32 (1967) 255–264.

J.T. Townsend, Structure issues in pattern recognition and their relation to template matching and feature testing models, Paper presented at the Psychonomic Society Meeting, San Antonio, Texas, November, 1970.

J.T. Townsend, Theoretical analysis of an alphabetic confusion matrix, Perception and Psychophys. 9 (1971a) 40–50.

J.T. Townsend, Alphabetic confusion: A test of models for individuals, Perception and Psychophys. 9 (1971b) 449–454.

J.T. Townsend, Some results on the identifiability of parallel and serial processes, British J. Math. Statist. Psych. 25 (1972) 168–199.

J.T. Townsend, Issues and models concerning the processing of a finite number of inputs, in: B.H. Kantowitz, ed., Human Information Processing: Tutorials in Performance and Cognition (Lawrence Erlbaum Associates, Hillsdale, NJ, 1974).

J.T. Townsend, A stochastic theory of matching processes, J. Math. Psych. 14 (1976a) 1–52.

J.T. Townsend, Serial and within-stage independent parallel model equivalence on the minimum completion time, J. Math. Psych. 14 (1976b) 219–238.

J.T. Townsend, A clarification of some current multiplicative confusion models, J. Math. Psych. 18 (1978) 25–38.

J.T. Townsend and F.G. Ashby, Testing contemporary models of letter recognition, Paper presented at the Ann. Meeting of the Midwestern Psychological Association, Chicago, 1976.

J.T. Townsend and F.G. Ashby, Methods of modeling capacity in simple processing systems, in: N.J. Castellan and F. Restle, eds., Cognitive Theory Vol. 3 (Lawrence Erlbaum Associates, Hillsdale, NJ, 1978).

J.T. Townsend and F.G. Ashby, On the Stochastic Modeling of Elementary Psychological Processes (Cambridge University Press, New York, 1983) in press.

J.T. Townsend, R. Evans and G.G. Hu, Some models in feature processing, Paper presented at the Midwestern Psychological Association Conference, Chicago, 1979.

J.T. Townsend and D.E. Landon, An experimental and theoretical investigation of the constant-ratio rule and other models of visual letter confusion, J. of Math. Psych. 14 (1982) 119–162.

J.T. Townsend, G.G. Hu and F.G. Ashby, A test of visual feature sampling independence with orthogonal straight lines, Bull. Psychonomic Soc. 15 (1980) 163–166.

J.T. Townsend, G.G. Hu and F.G. Ashby, Perceptual sampling of orthogonal straight line features, Psych. Res. 43 (1981) 259–275.

A. Tversky, Features of similarity, Psych. Rev. 84 (1977) 327–352.

W.R. Uttal, An Autocorrelation Theory of Form Detection (Lawrence Erlbaum Associates, Hillsdale, NJ, 1975).

D. Vickers, Decision Processes in Visual Perception (Academic Press, New York, 1979).

J. Wandmacher, Multicomponent theory of perception: Feature extraction and response decision in visual identification, Psych. Res. 39 (1976) 17–37.

S. Watanabe, ed., Frontiers of Pattern Recognition (Academic Press, New York, 1972).

J.I. Yellott, The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution, J. Math. Psych. 15 (1977) 109–144.

E.C. Zeeman, The topology of the brain and visual perception, in: M.K. Tart, ed., Topology of 3-Manifolds (Prentice-Hall, Englewood Cliffs, NJ, 1962).