JOURNAL OF MATHEMATICAL PSYCHOLOGY 39, 216–233 (1995)

Categorization as Probability Density Estimation

F. GREGORY ASHBY AND LEOLA A. ALFONSO-REESE

University of California at Santa Barbara

A category can be represented as a probability density function (pdf), that is, as a set of exemplars along with the probability or likelihood that each is selected as a stimulus. This article examines the relation between categorization and pdf estimation. We first discuss the differences between classifiers that know the true category pdfs and classifiers that must estimate these functions from trial-by-trial feedback. Consistency is shown to be the key statistical property that guarantees two such classifiers will reasonably agree. Parametric and nonparametric pdf estimators are interpreted from the perspective of the categorization process. It is shown that the prototype model and several decision-bound models of categorization are parametric, whereas most exemplar models are nonparametric. The exemplar models are shown to be equivalent to a classifier that uses the minimum variance unbiased estimator of the category baserates and a nonparametric pdf estimator that is one of the most commonly used estimators of professional statisticians (i.e., the kernel estimator). It is also shown that in most applications, the exemplar models predict essentially optimal performance. Finally, the implications of these results are discussed and an alternative approach to the study of categorization is suggested.

In a categorization task, the experimenter identifies two or more sets of objects or events called categories. The members of a category are called exemplars. On each trial, a stimulus is chosen by randomly selecting an exemplar from one of the categories. The subject's problem is to name the category to which the stimulus belongs.

The set of exemplars comprising the relevant categories will vary in one or more attributes. Most currently popular categorization theories assume that the psychological effects of each stimulus can be described by a vector containing the perceived magnitude of every attribute that varies. In other words, each stimulus can be represented as a point in a multidimensional psychological space. According to this model, a category can be represented as a probability distribution, that is, as a set of points along with the probability or likelihood that each is selected as a stimulus.

To solve the categorization problem, many theories assume the subject computes a set of values that measure the strength of association between the stimulus and each of the relevant categories. For example, in prototype theory, the strength of association is the similarity between the stimulus and the category prototype. In exemplar theory, the strength of association is the sum of similarities between the stimulus and all encoded exemplars of the category. This article shows that in the currently popular categorization models, these values are proportional to estimates of the category probability density functions (pdfs).

As a result, categorization models can be described and compared by using the statistical language of probability density estimation. An immediate advantage of such an approach is that the many results described in the extensive pdf estimation literature can be used to provide new insights into categorization performance. Also, by viewing categorization as pdf estimation, psychologists benefit from important distinctions made in the statistical literature that are not currently emphasized in the categorization literature. For example, one important distinction between pdf estimators is whether they are *parametric* or *nonparametric*. In categorization this distinction corresponds to whether subjects make strong assumptions (corresponding to parametric models) or almost no assumptions (corresponding to nonparametric models) about category structure.

This article proceeds as follows: First, the basic structure of categorization models is described in terms of their representation, category access, and response selection assumptions. Next, we elaborate on the notion of categorization as probability density estimation. We discuss the differences between classifiers that know the true category pdfs and classifiers that must estimate these functions from the trial-by-trial feedback provided by the experimenter. It is shown that *consistency* is the key statistical property that guarantees two such classifiers will reasonably agree. The third section defines parametric and nonparametric pdf estimation and relates these statistical terms to the categorization process. The fourth section discusses parametric categorization models. In particular, it is shown that the prototype model and several decision-bound models are parametric. The fifth section shows that most popular exemplar models of categorization are nonparametric. In particular, we show that they are equivalent to a classifier that uses the minimum variance unbiased estimator of the category baserates and a nonparametric pdf estimator that is one of the most widely used estimators of

Correspondence and reprint requests should be sent to F. Gregory Ashby, Department of Psychology, University of California, Santa Barbara, CA 93106. Email: ashby@psych.ucsb.edu.

0022-2496/95 $6.00

professional statisticians. We also show that in most applications these models predict essentially optimal performance. Finally, the last section discusses implications of these results and suggests an alternative approach to studying human categorization performance.

CATEGORIZATION MODELS

Virtually all categorization models make assumptions about (i) the representation of stimuli, exemplars, and categories; (ii) the information that is accessed from the category representations and the computations that are performed on this information; and (iii) response selection, that is, how a response is selected after the requisite information has been collected and computed.

Almost all currently popular categorization models assume a numeric representation. In particular, they assume that stimuli and exemplars may be represented as points (or a probability distribution of points), in a multidimensional psychological space. Therefore, this article assumes that such a numeric representation is possible. Let the vector $\mathbf{x}' = [x_1, x_2, ..., x_m]$ denote the coordinates of stimulus X in the m-dimensional psychological space (where the prime denotes vector or matrix transpose). In most models the vector $\mathbf{x}$ is a constant, but in the decision-bound model $\mathbf{x}$ is a random vector. The vector $\mathbf{x}$ and the other notation used in this article are defined in Table 1.

The category access assumptions (i.e., assumptions of type ii) delineate the various categorization theories. We focus on three different types of models: prototype, exemplar, and decision-bound models. Prototype models assume the category representation is dominated by the prototype, which is usually defined as the most typical, or representative, category member (Posner & Keele, 1968, 1970; Reed, 1972; Rosch, 1973, 1977). On each trial, the subject is assumed to access the prototypes of all competing categories. Next, the similarity is computed between each prototype and the stimulus (or more technically, the percept elicited by the stimulus).

Exemplar models assume that every exemplar of all relevant categories is accessed. Next, for each category, the similarity between the stimulus and each exemplar is computed, and finally, all these similarities are summed (Brooks, 1978; Estes, 1986, 1994; Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986). Decision-bound models assume the subject constructs a decision bound that partitions the perceptual space into response regions, one for each relevant category. On each trial, the subject determines the region in which the stimulus representation falls and then emits the associated response (Ashby, 1992a; Ashby & Gott, 1988; Ashby & Lee, 1991, 1992; Ashby & Townsend, 1986; Maddox & Ashby, 1993).

Response selection models are either *deterministic* or *probabilistic*. Deterministic models assume that, if on different trials the perceptual information is the same and the same information is accessed from memory, then the subject always selects the same response. Deterministic response selection rules are of one basic type. Consider a task with categories A and B. Let $g(\mathbf{x})$ be some function of the stimulus representation with the property that a stimulus

TABLE 1

Notation Used in This Article in Order of Appearance

Symbol	Description
$\mathbf{x}$	coordinates of stimulus X in psychological space
x_k, y_k	coordinates of stimulus X and Y, respectively, along dimension k
X	stimulus
m	number of dimensions of psychological space
A, B	categories
$g(\mathbf{x})$	discriminant function
δ	response criterion
$P(A), P(B)$	*a priori* probability that stimulus is from categories A and B, respectively
$f_A(\mathbf{x}), f_B(\mathbf{x})$	probability density functions of categories A and B, respectively
$L(\mathbf{x})$	likelihood ratio
S_{XA}, S_{XB}	strength of relationship between stimulus X and categories A and B, respectively
$M(\mathbf{x})$	matching probability
β_A, β_B	response biases toward categories A and B, respectively
N	total number of stimuli in an experiment (i.e., total sample size)
N_A, N_B	number of stimuli in categories A and B, respectively
$\hat{P}(J), \hat{L}_N(\mathbf{x}),$ $\hat{M}_N(\mathbf{x}), \hat{f}_J(\mathbf{x})$	estimators of various functions
θ, ξ	parameters
$\hat{\theta}_N, \hat{\xi}_N$	parameter estimators
h_k	kernel width on dimension k
$K(\mathbf{x})$	kernel function
$k_k(x_k)$	marginal kernel function on dimension k
$\sum_A, \sum_B$	covariance matrices for categories A and B, respectively
σ^2	variance
I	identity matrix
π	probability of encoding a new exemplar
$P(R_A \mid X)$	probability of responding A on trials when stimulus X is presented
η_{XY}	similarity between stimulus X and stored exemplar Y
Y	stored exemplar
s_k	component similarity function
q_k	value of similarity function along dimension k when $x_k \neq y_k$
α	similarity function exponent
d_{XY}	distance between stimulus X and stored exemplar Y
w_k	attention weight for dimension k
C	nonnegative scaling parameter measuring overall discriminability
r	Minkowski distance function exponent
p_k	context model and array model parameter indicating kernel function value for $x \neq 0$
μ_A, μ_B	category mean vectors
$\mathbf{w}_k$	normalized eigenvectors
y_{i_k}	coordinate of stored exemplar Y_i along dimension k

with coordinates $\mathbf{x}$ is more likely to be a member of category A when $g(\mathbf{x})$ is negative and a member of category B when $g(\mathbf{x})$ is positive. For example, in a prototype model $g(\mathbf{x})$ might equal the difference in similarities between the stimulus and the two category prototypes. The deterministic decision rule[1] is to

$$\text{Respond } A \text{ if } g(\mathbf{x}) < \delta; \quad \text{Respond } B \text{ if } g(\mathbf{x}) > \delta, \quad (1)$$

where δ is the response criterion. In some models δ is a random variable (see, e.g., Maddox & Ashby, 1993).

The optimal classifier—that is, the device that maximizes categorization accuracy—uses a rule like Eq. (1). Let $P(J)$ be the *a priori* probability that the stimulus is from category J and let $f_J(\mathbf{x})$ be the likelihood[2] of stimulus X given that it is member of category J (so $f_J(\mathbf{x})$) is the category J pdf. Then it is well known that the most accurate of all decision strategies is the rule (see, e.g., Ashby, 1992b; Green & Swets, 1966)

$$\text{Respond } A \text{ if } L(\mathbf{x}) = \frac{f_A(\mathbf{x})}{f_B(\mathbf{x})} > \delta; \quad \text{Respond } B \text{ if } L(\mathbf{x}) < \delta, \quad (2)$$

where $\delta = P(B)/P(A)$. The Eq. (2) rule is sometimes written as

$$\text{Respond } A \text{ if } g(\mathbf{x}) = -\log L(\mathbf{x}) < -\log \delta;$$
$$\text{Respond } B \text{ if } g(\mathbf{x}) > -\log \delta.$$

The function $L(\mathbf{x})$ is known as the *likelihood ratio*. Because the rule is based on the true category pdfs and the true baserates, we call it the *true likelihood ratio rule*.

Probabilistic response selection models assume the subject always guesses, although usually in a sophisticated fashion. Sophisticated guessing, as defined by Broadbent (1967), is a strategy in which the subject guesses more likely responses with greater probability than less likely responses. In other words, if the evidence supports the hypothesis that the stimulus belongs to category A, then a deterministic model predicts that the subject will respond A with probability 1, whereas a probabilistic model predicts that response A will be given with probability less than 1 (but greater than 0.5). Probabilistic response selection models are also of one basic type. Let S_{XA} be a measure of the relationship between stimulus X and category A with the property that larger values of S_{XA} indicate a closer relationship. Then almost all probabilistic response selection models assume the subject uses the rule

$$\text{Respond } A \text{ with probability } M(\mathbf{x}) = \frac{\beta_A S_{XA}}{\beta_A S_{XA} + \beta_B S_{XB}}, \quad (3)$$

where β_J is the response bias toward category J (with $\beta_J \geq 0$). Without loss of generality, one can assume that $\beta_B = 1 - \beta_A$. In many categorization models the response biases are set to $\beta_A = \beta_B = 1$. Equation (3) has many names. It was originally proposed by Shepard (1957) and Luce (1963), so it is often called the Luce–Shepard choice model. But it is also called the similarity-choice model, the biased-choice model, or the relative goodness rule. A special case of Eq. (3) that is of particular interest occurs when the subject's response probability matches the objective posterior probability associated with each relevant category. In this case, Eq. (3) reduces to

$$\text{Respond } A \text{ with probability } M(\mathbf{x})$$
$$= \frac{P(A \mid \mathbf{x})}{P(A \mid \mathbf{x}) + P(B \mid \mathbf{x})} = \frac{P(A) f_A(\mathbf{x})}{P(A) f_A(\mathbf{x}) + P(B) f_B(\mathbf{x})}. \quad (4)$$

We call Eq. (4) the *true probability matching rule* (Estes, 1976; Herrnstein, 1961, 1970). The true probability matching rule is closely related to the true likelihood ratio rule because

$$M(\mathbf{x}) = \frac{P(A) f_A(\mathbf{x})}{P(A) f_A(\mathbf{x}) + P(B) f_B(\mathbf{x})} = \frac{L(\mathbf{x})}{L(\mathbf{x}) + \delta}.$$

CATEGORIZATION AS DENSITY ESTIMATION

In a typical categorization task, a subject attempting to use either the true likelihood ratio rule [Eq. (2)] or the true probability matching rule [Eq. (4)] would not know the true baserates [i.e., $P(J)$] or probability density functions for each category (since these are not usually provided by the experimenter). Instead, the subject would be forced to estimate the baserates and density functions from the trial-by-trial feedback provided by the experimenter. Let N_J be the number of category J exemplars presented so far in the experiment and define $N = N_A + N_B$. Then the minimum variance unbiased estimator of $P(J)$ is $\hat{P}(J) = N_J/N$. Let $\hat{\delta}$ be an estimator of the criterion $\delta = P(B)/P(A)$ and let $\hat{L}_N(\mathbf{x})$ and $\hat{M}_N(\mathbf{x})$ be estimators of $L(\mathbf{x})$ and $M(\mathbf{x})$, respectively, that are each based on a total sample of N exemplars.[3]

[1] Throughout this article we assume the subject guesses if $g(\mathbf{x}) = \delta$. If the category pdfs are continuous, then $P[g(\mathbf{x}) = \delta] = 0$.

[2] More precisely, $f_J(\mathbf{x})$ is the likelihood that the percept $\mathbf{x}$ was elicited by an exemplar from category J. Thus, when we refer to the category pdf, we mean the distribution of all percepts elicited by exemplars from that category.

[3] Most of the popular methods of estimation assume random sampling from some invariant population (i.e., so that statistical independence holds). If these assumptions are violated, the methods will usually fail. For example, if the exemplars are blocked by category during the training sessions, it may be impossible for the subject to estimate accurately the category baserates to be expected during experimental testing. As a consequence, throughout this article we assume stimulus presentation satisfies the usual sampling assumptions.

Given that a subject cannot know the true likelihood ratio, the best a subject can do is to use the rule

$$\text{Respond } A \text{ if } \hat{L}_N(\mathbf{x}) > \hat{\delta}; \quad \text{Respond } B \text{ if } \hat{L}_N(\mathbf{x}) < \hat{\delta}. \tag{5}$$

We call this the *estimated likelihood ratio rule*. Similarly, a subject trying to probability match can, at best, use the *estimated probability matching rule*

$$\begin{aligned} &\text{Respond } A \text{ with probability } \hat{M}_N(\mathbf{x}); \\ &\text{Respond } B \text{ with probability } 1 - \hat{M}_N(\mathbf{x}). \end{aligned} \tag{6}$$

The categorization accuracy of a subject using the estimated likelihood ratio rule or the estimated probability matching rule will depend on how well they estimate δ, $L(\mathbf{x})$, and $M(\mathbf{x})$. In this article, we consider models in which estimators of $L(\mathbf{x})$ and $M(\mathbf{x})$ are constructed from estimators of the category pdfs. Let $\hat{f}_J(\mathbf{x})$ be an estimator of $f_J(\mathbf{x})$ that is based on the N_J exemplars from category J. Then this article considers models[4] in which $\hat{L}_N(\mathbf{x}) = \hat{f}_A(\mathbf{x})/\hat{f}_B(\mathbf{x})$ and

$$\hat{M}_N(\mathbf{x}) = \frac{\hat{P}(A)\hat{f}_A(\mathbf{x})}{\hat{P}(A)\hat{f}_A(\mathbf{x}) + \hat{P}(B)\hat{f}_B(\mathbf{x})}.$$

Note that if the subject estimates the category baserates by using the minimum variance unbiased estimator, then $\hat{M}_N(\mathbf{x})$ becomes

$$\hat{M}_N(\mathbf{x}) = \frac{N_A\hat{f}_A(\mathbf{x})}{N_A\hat{f}_A(\mathbf{x}) + N_B\hat{f}_B(\mathbf{x})}.$$

Under these conditions, the performance of a subject depends on his or her ability to estimate accurately the category baserates and pdfs. Estimation of the category baserates is straightforward, but estimation of the category pdfs is a difficult problem (see, e.g., Myung, 1994). The statistical literature contains many different pdf estimators. Some of these will provide reasonably good estimates of the category pdf, but others will not. A subject who uses the estimated likelihood ratio rule, along with the best possible pdf estimator, will still not respond optimally. Even so, we can expect reasonably good agreement between such a subject and the optimal classifier. On the other hand, a subject who uses a suboptimal rule, or one who uses the estimated likelihood ratio rule along with a poor pdf estimator, should agree poorly with the optimal classifier. Therefore, an important criterion for evaluating the efficacy of a pdf estimator is whether it leads to reasonable agreement between the estimated and true likelihood ratio rules.

What does it mean to say that a subject is in reasonable agreement with the optimal classifier? A fairly strong definition is that, with large enough sample sizes, the subject and the optimal classifier agree on every decision with probability 1. This idea is formalized in our first definition.

DEFINITION 1. The true likelihood ratio rule [Eq. (2)] and the estimated likelihood ratio rule [Eq. (5)] are in *reasonable agreement* if

$$\lim_{N\to\infty} P[\hat{L}_N(\mathbf{x}) > \hat{\delta} \mid L(\mathbf{x}) > \delta] = 1,$$

where N is the sample size.

Definition 1 establishes reasonable agreement as an asymptotic property of categorization performance. As such, it is possible that a subject could reasonably agree with the true likelihood ratio rule, yet perform badly on early training trials. For example, suppose Subjects 1 and 2 both participate in the same categorization experiment and that Subject 1 is in reasonable agreement with the true likelihood ratio rule. Even in this case, it is possible that Subject 2 will significantly outperform Subject 1 during most of the training period. However, if Subject 2 always performs at least as well as Subject 1, then it must be true that Subject 2 is also in reasonable agreement with the true likelihood ratio rule.

With the probability matching rule, the strong form of agreement established in Definition 1 is not possible. This is because the likelihood ratio rule used by the optimal classifier selects a response on every trial, whereas the probability matching rule does not; it only specifies the probability of emitting a particular response. Thus, two subjects given the true category pdfs who are both probability matching could disagree on many trials. As a consequence, we say that a subject using the estimated probability matching rule of Eq. (6) reasonably agrees with the true probability matching rule if, for large sample sizes, the two predict identical response probabilities. This idea is expressed formally in Definition 2.

DEFINITION 2. The true probability matching rule [Eq. (4)] and the estimated probability matching rule [Eq. (6)] are in *reasonable agreement* if

$$\lim_{N\to\infty} P[|\hat{M}_N(\mathbf{x}) - M(\mathbf{x})| < \varepsilon] = 1, \qquad \text{for all} \quad \varepsilon > 0,$$

where N is the sample size.

This second form of agreement is weaker than the first form because if two rules give the same response on every trial (i.e., Definition 1), they necessarily predict the same response probabilities (i.e., Definition 2), but the converse is not true. Two rules could predict the same response probabilities but give different responses on many trials.

Given a definition of reasonable agreement, the next step is to ask what statistical properties the estimators $\hat{\delta}, \hat{f}_A(\mathbf{x})$,

[4] There is no guarantee that this estimation strategy will produce the best estimators.

and $\hat{f}_B(\mathbf{x})$ must satisfy before reasonable agreement is guaranteed. Ideally, we seek the weakest possible property that guarantees reasonable agreement. There are a number of possibilities. Among the strongest is the combination of minimum variance and unbiasedness. Among the weakest is *consistency*. An estimator, $\hat{\theta}_N$, is consistent if it converges to the true parameter value, θ, as the sample size approaches infinity. Consistency is a weak property because even biased estimators can be consistent. In fact, a sufficient condition for consistency is that the bias and variance converge to zero as the sample size approaches infinity. A formal definition of consistency is as follows:

DEFINITION 3. A sequence of estimators $\{\hat{\theta}_N\}$ of a parameter θ is *consistent* if and only if, for every $\varepsilon > 0$,

$$\lim_{N \to \infty} P(|\hat{\theta}_N - \theta| < \varepsilon) = 1.$$

A sequence of estimators $\{\hat{f}_N(\mathbf{x})\}$ of a pdf $f(\mathbf{x})$ is consistent if it is consistent at every value of $\mathbf{x}$.

The two major results in this section show that consistency is a sufficient condition to ensure reasonable agreement both between the estimated and true likelihood ratio rules and between the estimated and true probability matching rules. The proofs of these theorems depend heavily on the following result:

LEMMA 1. *If $\hat{\theta}_N$ and $\hat{\xi}_N$ are consistent estimators of the parameters θ and ξ, respectively, then*

(a) *$\hat{\theta}_N \hat{\xi}_N$ is a consistent estimator of $\theta\xi$,*

(b) *$\hat{\theta}_N + \hat{\xi}_N$ is a consistent estimator of $\theta + \xi$,*

(c) *$\hat{\theta}_N / \hat{\xi}_N$ is a consistent estimator of θ/ξ, if $P(\hat{\xi}_N = 0) < 1$ and $\xi \neq 0$.*

Proof. All proofs are given in the Appendix.

Armed with these tools, we are now in a position to state and prove our first result.

THEOREM 1. *Consider a categorization experiment with two categories, A and B. Suppose $\hat{\delta}$ is a consistent estimator of the criterion δ, and $\hat{f}_A(\mathbf{x})$ and $\hat{f}_B(\mathbf{x})$ are consistent estimators of the category pdfs $f_A(\mathbf{x})$ and $f_B(\mathbf{x})$, respectively. Then the estimated and true likelihood ratio rules [Eqs. (5) and (2)] are in reasonable agreement (Definition 1).*

Thus, consistency is enough to guarantee reasonable agreement. Theorem 2 shows that consistency is also sufficient for reasonable agreement between the estimated and true probability matching rules.

THEOREM 2. *Consider a categorization experiment with two categories, A and B. Suppose $\hat{P}(A)$ and $\hat{P}(B)$ are consistent estimators of the category baserates, $P(A)$ and $P(B)$, respectively, and $\hat{f}_A(\mathbf{x})$ and $\hat{f}_B(\mathbf{x})$ are consistent estimators of the category pdfs $f_A(\mathbf{x})$ and $f_B(\mathbf{x})$, respectively. Then the estimated and true probability matching rules [Eqs. (6) and (4)] are in reasonable agreement (Definition 2).*

Theorems 1 and 2 are important because they establish consistency as the key statistical property that guarantees reasonable agreement with an optimal classifier that is using either the true likelihood ratio rule or the true probability matching rule. In particular, this means that a subject who wishes to respond in an approximately optimal fashion need not estimate the category baserates or pdfs with minimum variance unbiased estimators, or even with unbiased estimators. Instead, only the weak property of consistency is required.[5]

PARAMETRIC VERSUS NONPARAMETRIC DENSITY ESTIMATION

In the statistical literature, an important distinction among pdf estimators is whether they are parametric or nonparametric. Parametric estimators make strong assumptions about the distribution of the observed data, whereas nonparametric estimators make weak assumptions. For example, a parametric estimator might assume that the distribution of the observed data is normal. If so, then the problem of pdf estimation is reduced to the problem of estimating the parameters of the normal distribution (i.e., means, variances, and correlations). In contrast, a nonparametric estimator might assume only that the true pdf is continuous.

Parametric and nonparametric estimators form two ends of a continuum. It clearly is possible to construct estimators that lie between the two extremes of assuming normality and assuming only that the true pdf is continuous. For example, an estimator might assume that the true pdf is in the exponential family, which includes the normal, binomial, Rayleigh, and exponential distributions, among many others. Alternatively, the estimator might assume the true pdf is unimodal and continuous. As we will see, however, the currently popular categorization models are all equivalent to a pdf estimation process in which the estimator is firmly anchored at one end of the continuum or the other.

Nonparametric pdf estimation is a difficult statistical problem. No minimum variance unbiased estimators are known. In fact, no unbiased estimators are known. As a

[5] While preparing this article for publication, we discovered a technical report written in 1951 by Fix and Hodges, which anticipates some of the ideas in this section. In particular, although they used other language, Fix and Hodges proposed a definition of reasonable agreement that is similar to our own. They also identified consistency as the key statistical property of the pdf estimators that guarantees reasonable agreement. However, their development did not allow for unequal category baserates, they had no discussion of the probability matching rules, and they did not prove any of the theorems presented in this article. The Fix and Hodges (1951) technical report has only recently been discovered by the statistical community (Silverman & Jones, 1989).

consequence, many different nonparametric estimators have been proposed (see, e.g., Scott, 1992; Silverman, 1986). The most widely known nonparametric estimator is the relative frequency histogram, but many other nonparametric estimators are known to be more accurate (according to a number of different accuracy criteria; see, e.g., Scott, 1992; Silverman, 1986). For example, the kernel estimator (Parzen, 1962) can be viewed as a generalization of the histogram in which the fixed bins are replaced by non-stationary smooth weighting functions or kernels. In general, kernel estimators lead to substantially smaller mean integrated squared errors than relative frequency histograms (see, e.g., Scott, 1992, Fig. 6.4). Because of its attractive statistical properties and its simplicity, kernel estimators are popular with professional statisticians. In the words of Silverman (1986), "Apart from the histogram, the kernel estimator is probably the most commonly used estimator" (p. 17). He goes on to add that "the kernel method is a good choice for many practical purposes; it is simple and intuitively appealing, and its mathematical properties are quite well understood" (p. 95). A formal definition of the kernel estimator is as follows:

DEFINITION 4. Suppose a random sample of size N has been drawn from some population of m-dimensional vectors with pdf $f(\mathbf{x})$. Denote the ith of these sample vectors by

$$\mathbf{x}_i' = [x_{i_1}, x_{i_2}, ..., x_{i_m}].$$

The *kernel estimator* of $f(\mathbf{x})$ has the form

$$\hat{f}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} K(\mathbf{x} - \mathbf{x}_i),$$

where

(a) $\int K(\mathbf{x})\, d\mathbf{x} = 1$, and

(b) $\int x_k K(\mathbf{x})\, d\mathbf{x} = 0$, for all $1 \leqslant k \leqslant m$.

$\hat{f}(\mathbf{x})$ is a *product kernel estimator* of $f(\mathbf{x})$ if the kernel $K(\mathbf{x} - \mathbf{x}_i)$ equals

$$K(\mathbf{x} - \mathbf{x}_i) = \prod_{k=1}^{m} \frac{1}{h_k} \kappa_k \left(\frac{x_k - x_{i_k}}{h_k} \right). \tag{7}$$

The constant h_i is called the kernel width on dimension i. Also, h_i depends in some way on the sample size N.

Because of condition (a), $K(\mathbf{x})$ is a probability density function if it is non-negative everywhere. Although kernels are not required to be non-negative, many of the most popular kernels are pdfs. For example, a popular choice is to let $K(\mathbf{x})$ be the multivariate normal density function. If the kernel is a pdf, then condition (b) is equivalent to assuming the kernel mean is zero on every dimension. In addition, the $k_k(x_k)$ kernel of the product kernel estimator then corresponds to a marginal density function and the product rule of Eq. (7) is equivalent to assuming statistical independence. The kernel $K(\mathbf{x})$ is a product of the marginal kernels [i.e., the $k_k(x_k)$] because under statistical independence, the joint density equals the product of the marginal densities. If a normal kernel is chosen, the kernel with h_i becomes the standard deviation on dimension i. In most applications, a constant h_i is chosen for all dimensions (i.e., $h_1 = \cdots = h_m$).

In the categorization literature, the parametric versus nonparametric dichotomy corresponds to whether a subject makes strong or weak assumptions, respectively, about the distribution of category exemplars. For example, a subject who assumes category members are normally distributed need only estimate the mean and variance of the exemplars to derive an estimated pdf of that category. Alternatively, a subject who makes no family distribution assumptions of a category might compute a histogram of the exemplars to estimate the pdf of that category.

We are now in a position to re-examine the popular categorization models. In each case, we are interested primarily in two questions. The first is whether the model assumes subjects use parametric or nonparametric category pdf estimators. The second problem is to identify the conditions under which the pdf estimator assumed by the model is consistent. These are the experimental conditions under which the model predicts performance that is essentially optimal. Suboptimal performance by motivated, practiced subjects under these same conditions would therefore be extremely problematic for this class of models.

A number of recent articles have established equivalence relations between the true likelihood ratio rule (i.e., the optimal classifier) and many different categorization models (Ashby & Maddox, 1993; Myung, 1994; Nosofsky, 1990). In general, however, these results are asymptotic in the sense that they assume the subject has had an infinite amount of experience with the categories. Thus, they ignore the statistical problems facing a subject who has only limited experience with the categories (although, see, Myung, 1994). As a consequence, none of these articles draws a connection between categorization and trial-by-trial probability density function estimation. In contrast, most of the results in this article hold exactly for any sample size. The results on consistency are asymptotic, but unlike the asymptotic results of earlier papers, our results place strict requirements on the performance of the subject as he or she approaches that asymptote.

We begin our survey by examining parametric models and conclude with nonparametric models.

PARAMETRIC CATEGORIZATION MODELS

Parametric pdf estimators assume the unknown pdf belongs to a specific family of probability distributions.

Before examining specific parametric categorization models, we examine the conditions under which parametric estimators of the pdf are consistent. These conditions are specified in the next theorem. Since this result is well known in the statistical literature, we state it without proof.

THEOREM 3. *Let $\hat{f}(\mathbf{x})$ be a parametric estimator of the pdf $f(\mathbf{x})$, where the parameters of $f(\mathbf{x})$ are $\boldsymbol{\theta}$ and the estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}$. Then $\hat{f}(\mathbf{x})$ is consistent if:*

(a) *the family of distributions assumed for $\hat{f}(\mathbf{x})$ is the one to which $f(\mathbf{x})$ belongs, and*

(b) *$\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$.*

With most well known parameters (e.g., means, variances, covariances), consistent estimators are easy to find. Therefore, Theorem 3 indicates that the major obstacle to optimal performance by a parametric classifier is the problem of choosing the correct family of probability distributions. In categorization, this problem becomes one of assuming the correct category structure. If the natural categories that an organism must learn have no common structure, then parametric classification makes little sense. If natural categories have many different structures, then a parametric classifier that encounters a new category would often assume an incorrect structure. In this case, we would expect the resulting pdf estimate not to be consistent and suboptimal performance to result. In a world where natural categories have many different structures, nonparametric classification makes more sense.

There is a scenario, however, in which parametric classification is preferred, even if natural categories have no common structure. Most nonparametric pdf estimators require extensive memory and computation (e.g., see Definition 4). With large categories the memory and computational requirements may exceed the capacity of the subject. If so, a simpler alternative is required. One possibility is that the subject estimates only a few moments of the category distributions. For example, the subject may estimate the mean exemplar value on each dimension, the variances, the correlations between dimensions, and the category baserates. Even with these estimates, however, to use either the estimated likelihood ratio rule or the estimated probability matching rule, the subject must assume a distributional family. If the subject's experience with natural categories suggests no solution to this problem, some other criterion for choosing a family of distributions is required. Myung (1994) argued that the best solution to this problem is to assume the category distributions are multivariate normal. Given estimates of the means, variances, correlations, and baserates, the multivariate normal is the maximum entropy inference (assuming the stimulus dimensions are not known to be restricted to some proper subset of the real line). To infer any other family of distributions requires assumptions beyond those needed to infer the multivariate normal. In other words, the multivariate normal distribution is the appropriate noncommittal choice in such situations.

If most natural categories possess a common structure, then parametric classification becomes more desirable than nonparametric classification. For example, it is far easier to estimate an unknown mean and variance than an unknown pdf. Ashby (1992a) argued that many natural categories might be well described by the multivariate normal distribution. First, natural categories tend to have a very large, if not unlimited, number of exemplars. Second, the dimensions of natural categories are often continuous-valued. Third, many natural categories appear to overlap with one another. For example, the category bound between goats and sheep is indistinct (e.g., Schaller, 1979). Finally, many natural categories appear to be symmetrically distributed about some single prototypical member, or at least, subjects often enter categorization tasks with this expectation (e.g., Flannagan, Fried, & Holyoak, 1986; Fried & Holyoak, 1984). The multivariate normal distribution has each of these properties.

As we will now show, many of the currently popular categorization models are equivalent to a parametric classifier that assumes multivariate normal distributions. The next result shows this is true of prototype models.

THEOREM 4. *Consider a categorization task with categories A and B. The prototype model (Reed, 1972) is equivalent to a parametric classifier that uses the true likelihood ratio rule and assumes*

(a) *the category baserates are equal [i.e., $P(A) = P(B)$],*

(b) *the category distributions are multivariate normal,*

(c) *the category covariance matrices are equal (i.e., $\Sigma_A = \Sigma_B = \Sigma$), and*

(d) *one of the eigenvectors of Σ is orthogonal to the minimum distance bound.*

The prototype model predicts that subjects will use the minimum distance bound (e.g., Ashby, 1992a). This is the set of all points equidistant from the two category means. Of course, the subject could construct such a bound only if the population means are known exactly. This is why the prototype model is equivalent to a classifier that uses the *true* likelihood ratio rule. The most widely known conditions under which minimum distance classification is optimal are that $\Sigma_A = \Sigma_B = \Sigma = \sigma^2 I$, where I is the identity matrix. When $\Sigma = \sigma^2 I$, the eigenvectors of Σ can be chosen to have any direction (so long as they are all mutually orthogonal). Thus, one of the eigenvectors of Σ could be chosen to be orthogonal to the minimum distance bound, so the condition that $\Sigma = \sigma^2 I$ is a special case of the conditions detailed in Theorem 4. On the other hand, it is easy to find examples in which the conditions of Theorem 4 are satisfied but $\Sigma \neq \sigma^2 I$. An example is shown in Fig. 1.

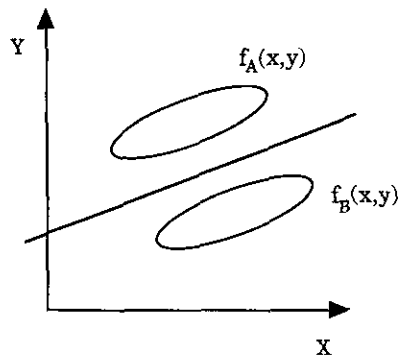

FIG. 1. Contours of equal likelihood and the optimal decision bound from a categorization task with categories A and B. In this experiment, the optimal classifier uses the minimum distance bound even though, within each category, the variances differ across dimensions and the correlation is nonzero.

Theorem 4 states conditions that are sufficient for minimum distance classification to be optimal, but these conditions are not necessary. For example, minimum distance classification is optimal with certain non-normal category distributions.

We turn next to two popular decision-bound models. The general linear classifier assumes only that the subject uses some linear decision bound and the general quadratic classifier assumes only that the subject uses some quadratic decision bound.

THEOREM 5. *Consider a categorization task with categories A and B. The general linear classifier (Ashby, 1992a; Maddox & Ashby, 1993) is equivalent to a parametric classifier that uses the estimated likelihood ratio rule and assumes the category distributions are multivariate normal with equal covariance matrices (i.e., $\Sigma_A = \Sigma_B = \Sigma$). It is also equivalent to a prototype model in which the subject estimates the prototype coordinates.*

Theorem 5 implies that there are two alternative interpretations of the general linear classifier. Both assume the subject estimates parameters of the category pdfs. The first interpretation assumes the subject estimates means, variances, correlations, and category baserates and then uses the estimated likelihood ratio rule, under the assumptions of normality and equal category covariance structures. In the second interpretation, the subject estimates only the category means and then uses the minimum distance rule of prototype theory.

THEOREM 6. *Consider a categorization task with categories A and B; The general quadratic classifier (Ashby, 1992a; Maddox & Ashby, 1993) is equivalent to a parametric classifier that uses the estimated likelihood ratio rule and assumes the category distributions are multivariate normal.*

Thus, the prototype model, the general linear classifier, and the general quadratic classifier are all closely related. They are all equivalent to parametric classifiers that assume the category distributions are multivariate normal. The prototype model and the general linear classifier are most closely related. Under one interpretation, they differ only in whether they assume the subject knows the true coordinates of the category prototypes or must estimate the prototype coordinates from the available data.

NONPARAMETRIC CATEGORIZATION MODELS

Nonparametric classifiers make only weak assumptions about category structure. In general, they do not assume a specific family of probability distributions, although they may assume that the category pdfs are continuous. The first major result of this section shows that a large class of exemplar models are equivalent to a nonparametric classifier that estimates the category pdfs with the widely used kernel estimator.

Although many different exemplar models have been proposed, they all assume that categorization judgments are based on some sort of global match between the representation of the presented stimulus and the memory representations of every exemplar of each relevant category (see, e.g., Brooks, 1978; Estes, 1986, 1994; Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986). In fact, most of the models agree on many details of this global matching process. It is therefore possible to define a *general exemplar model* that contains many of the currently popular exemplar models as special cases.

DEFINITION 5. Consider a categorization task with two categories, A and B. The *general exemplar model* makes the following assumptions:

(i) Each new exemplar encountered in the experiment is encoded into memory with probability π.

(ii) The probability of responding A on trials when stimulus X is presented is

$$P(R_A \mid X) = \frac{\beta_A \sum_{Y \in A} \eta_{XY}}{\beta_A \sum_{Y \in A} \eta_{XY} + \beta_B \sum_{Y \in B} \eta_{XY}}, \qquad (8)$$

where η_{XY} is the similarity between stimulus X and stored exemplar Y and β_J is the response bias toward category J.

(iii) η_{XY} satisfies the following conditions for all values of k:

(a) $\eta_{XY} = \prod_{k=1}^{m} \eta_{x_k y_k}$ (product rule),

(b) $\eta_{x_k x_k} > \eta_{x_k y_k}$, for all $y_k \neq x_k$,

(c) $\eta_{x_k y_k} = s_k(x_k - y_k)$, where s_k is a non-negative, symmetric function, and

(d) $\int_{-\infty}^{\infty} x_k s_k(x_k)\, dx_k < \infty$.

The various exemplar models differ in their assumptions about stimulus and exemplar representation and in how they define the similarity between the stimulus X and the stored exemplar Y (i.e., η_{XY}). Many of the models restrict application to stimuli that vary on binary-valued dimensions. This includes the context model (Medin & Schaffer, 1978), the array model (Estes, 1986, 1994), and several network models (Estes, 1994; Gluck & Bower, 1988; Hurwitz, 1990).

The context model and the array model both assume an unbiased version of the Eq. (8) probability matching rule (i.e., they assume $\beta_A = \beta_B$) and they both assume

$$\eta_{x_k y_k} = \begin{cases} 1, & \text{if } x_k = y_k, \\ q_k, & \text{if } x_k \neq y_k, \end{cases}$$

where $0 \leqslant q_k \leqslant 1$. This definition of similarity satisfies conditions (b), (c), and (d) of the general exemplar model (i.e., Definition 5) with

$$\eta_{x_k y_k} = s_k(x_k - y_k) = \begin{cases} q_k, & \text{if } x_k - y_k = -1, \\ 1, & \text{if } x_k - y_k = 0, \\ q_k, & \text{if } x_k - y_k = 1. \end{cases} \tag{9}$$

Since the context model and the array model both assume overall similarity satisfies the product rule, they are both special cases of the general exemplar model.

A few exemplar models allow continuous-valued dimensions. The most widely known of these are the generalized context model (Nosofsky, 1986) and ALCOVE (Kruschke, 1992). These models assume perfect encoding (i.e., $\pi = 1$) and that

$$\eta_{XY} = \exp(-d_{XY}^{\alpha}), \tag{10}$$

where $\alpha = 1$ (exponential similarity function) or $\alpha = 2$ (Gaussian similarity function) and the distance between stimulus X and stored exemplar Y, d_{XY}, is defined by the weighted Minkowski metric.

$$d_{XY} = C\left(\sum_{k=1}^{m} w_k |x_k - y_k|^r\right)^{1/r}. \tag{11}$$

The non-negative scaling parameter C is a measure of overall discriminability and would be expected to increase with the subject's motivation and experience (Nosofsky, 1986). The exponent r defines the nature of the distance metric. The most popular cases occur when $r = 1$ (city-block distance) and when $r = 2$ (Euclidean distance). The parameter w_k measures the proportion of attention allocated to stimulus dimension k (and thus $\sum w_k = 1$). Finally, α determines the nature of the similarity function.

The generalized context model (Nosofsky, 1986) assumes the same response selection rule as the general exemplar model [i.e., Eq. (8)]. Not all versions of the generalized context model satisfy the product rule, however [i.e., condition (a) of Definition 5]. When similarity is defined by Eqs. (10) and (11), the product rule holds if and only if the exponent of the Minkowski distance metric equals the exponent of the similarity function (i.e., if and only if $r = \alpha$; see Nosofsky, 1984). It is clear that all versions of Eqs. (10) and (11) satisfy conditions (b) and (c) of the general exemplar model. If $r = \alpha$, condition (d) holds because

$$\begin{aligned} \int_{-\infty}^{\infty} x_k s_k(x_k)\, dx_k &= \int_{-\infty}^{\infty} x_k \exp(-C^r w_k |x_k|^r)\, dx_k \\ &= \int_{-\infty}^{0} x_k \exp(-C^r w_k |x_k|^r)\, dx_k \\ &\quad + \int_{0}^{\infty} x_k \exp(-C^r w_k x_k^r)\, dx_k \\ &= -\int_{0}^{\infty} y_k \exp(-C^r w_k y_k^r)\, dy_k \\ &\quad + \int_{0}^{\infty} x_k \exp(-C^r w_k x_k^r)\, dx_k \\ &= -\frac{\Gamma(2/r)}{rC^2 w_k^{2/r}} + \frac{\Gamma(2/r)}{rC^2 w_k^{2/r}} \\ &= 0, \end{aligned}$$

so the integral converges and condition (d) holds. Thus, the generalized context model with $r = \alpha$ is a special case of the general exemplar model.

In virtually all applications of the generalized context model, α is restricted to $\alpha = 1$ or $\alpha = 2$. Therefore, the only versions of the generalized context model that satisfy the definition of the general exemplar model are the version with city-block distance and an exponential similarity function ($r = \alpha = 1$) and the version with Euclidean distance and a Gaussian similarity function ($r = \alpha = 2$). Most applications of the generalized context model have been restricted to these two versions.

Our next result shows that under certain weak conditions, all versions of the general exemplar model are equivalent to probability matching with a kernel density estimator.

THEOREM 7. *Consider a categorization task with categories A and B. The general exemplar model (Definition 5) is equivalent to probability matching on the stored exemplars with an estimator of the category baserates and a product kernel density estimator (Definition 4) if for all values of k,*

$$\int s_k(x_k)\, dx_k = b_k, \tag{12}$$

where b_k is a positive constant that may depend on k, but does not depend on the category. If there is no response bias in the general exemplar model (i.e., $\beta_A = \beta_B$), then the baserate estimator is unbiased and of minimum variance. If there is a response bias (i.e., $\beta_A \neq \beta_B$), then one of the baserate estimators is biased [i.e., either $\hat{P}(A)$ or $\hat{P}(B)$] and the other is unbiased and of minimum variance. In either case, the kernel $K(\mathbf{x})$ equals

$$K(\mathbf{x}) = \prod_{k=1}^{m} \frac{1}{h_k} \kappa_k \left(\frac{x_k}{h_k}\right) = \prod_{k=1}^{m} \frac{1}{b_k} s_k(x_k).$$

The condition expressed in Eq. (12) is actually quite weak in the sense that it is satisfied by most current exemplar models. The function $s_k(x_k - y_k)$ specifies the similarity between two stimulus values (i.e., x_k and y_k) on dimension k. Equation (12) is satisfied so long as the shape of this function does not depend on the category of the stored exemplar Y, although there can be a different similarity function on each dimension. For example, Eq. (12) would be violated if the subject used an exponential similarity function to compute the similarity between the stimulus and exemplars of category A, but switched to a Gaussian similarity function to compute the similarity between the stimulus and exemplars of category B. We know of no current models that allow this kind of flexibility in similarity computation. In particular, as the next two results show, the context model, the array model, and the generalized context model with $r = \alpha$ all satisfy the Eq. (12) constraint.

THEOREM 8. *The context model (Medin & Schaffer, 1978) and the array model (Estes, 1986, 1994) are equivalent to probability matching on the stored exemplars with the minimum variance unbiased estimator of the category baserates and a product kernel density estimator in which the marginal kernel on dimension k equals*

$$\frac{1}{h_k} \kappa_k \left(\frac{x}{h_k}\right) = \begin{cases} p_k, & \text{if } x = -1, \\ 1 - 2p_k, & \text{if } x = 0, \\ p_k, & \text{if } x = 1, \end{cases}$$

where $p_k = q_k/(1 + 2q_k)$ and $0 \leq q_k \leq 1$.

In statistical applications, Parzen kernel density estimators are usually applied only when the data are continuous-valued. With discrete-valued data, minimum variance unbiased estimators are easy to find. This is because when $\mathbf{x}$ is discrete, $f_J(\mathbf{x})$ is a probability rather than a likelihood, which can be estimated in the same fashion as any other probability. Specifically, when $\mathbf{x}$ is discrete, the proportion of total samples from category J exactly equal to $\mathbf{x}$ is the minimum variance unbiased estimator of $f_J(\mathbf{x})$. Even so, there is nothing in the definition of the kernel estimator (i.e., Definition 4) that precludes discrete data and a discrete kernel.

THEOREM 9. *The generalized context model with $r = \alpha$ (Nosofsky, 1986) is equivalent to probability matching with an estimator of the category baserates and a product kernel density estimator if $C > 0$ and all $w_k > 0$. If there is no response bias in the generalized context model, then the baserate estimator is unbiased and of minimum variance. If there is a response bias, then one baserate estimator is unbiased and of minimum variance and the others are biased.*

This theorem has the following two immediate consequences.

COROLLARY 1. *The generalized context model with Euclidean distance (i.e., $r = 2$) and a Gaussian similarity function (i.e., $\alpha = 2$) is equivalent to probability matching with a product kernel density estimator (assuming $C > 0$ and all $w_k > 0$), in which the marginal kernel on dimension k ($1 \leq k \leq m$) is the univariate normal pdf with mean 0 and standard deviation h_k. The parameters of the generalized context model are related to the parameters of the kernel via*

$$C = \frac{1}{\sqrt{2}} \left(\sum_{i=1}^{m} \frac{1}{h_i^2} \right)^{1/2} \quad \text{and} \quad w_k = \frac{\prod_{i \neq k}^{m} h_i^2}{\sum_{i=1}^{m} \prod_{j \neq i}^{m} h_j^2} = \frac{1}{\sum_{i=1}^{m} h_k^2/h_i^2}.$$

COROLLARY 2. *The generalized context model with city-block distance (i.e., $r = 1$) and an exponential similarity function (i.e., $\alpha = 1$) is equivalent to probability matching with a product kernel density estimator (assuming $C > 0$ and all $w_k > 0$), in which the marginal kernel on dimension k ($1 \leq k \leq m$) is the univariate Laplace pdf with mean 0 and standard deviation h_k. Specifically,*

$$\frac{1}{h_k} \kappa_k \left(\frac{x_k}{h_k}\right) = \frac{1}{\sqrt{2}\, h_k} \exp\left(-\frac{\sqrt{2}}{h_k} |x_k| \right), \qquad -\infty < x_k < \infty.$$

The parameters of the generalized context model are related to the parameters of the kernel via

$$C = \sqrt{2} \sum_{i=1}^{m} \frac{1}{h_i} \quad \text{and} \quad w_k = \frac{\prod_{i \neq k}^{m} h_i}{\sum_{i=1}^{m} \prod_{j \neq i}^{m} h_j} = \frac{1}{\sum_{i=1}^{m} h_k/h_i}.$$

Note that in both corollaries, the parameter mappings ensure that $\Sigma w_k = 1$. Also, the mappings clearly illustrate the dependence of the discriminability parameter C and the attention weights w_k on the sample size, N (since h_i depends on N).

The results in this section show that the most popular exemplar models essentially assume that subjects probability match by using the minimum variance unbiased estimator of the category baserates and a product kernel density estimator. Furthermore, Theorem 2 showed that such a subject shows reasonable agreement with the true

probability matching rule when the kernel estimator is consistent. Not all product kernel estimators are consistent, so it is vital that we determine the conditions under which the product kernel estimators assumed by the various exemplar models are consistent. The next result establishes these conditions for the array and context models.

THEOREM 10. *The product kernel density estimator implicitly assumed by the context model (Medin & Schaffer, 1978) and the array model (Estes, 1986, 1994) is the minimum variance unbiased estimator of the category distributions if* $\pi = 1$ *(perfect encoding) and* $q_k = 0$ *for* $1 \leqslant k \leqslant m$ *(perfect discriminability). It is a consistent estimator of the category distributions if* $\pi > 0$ *and*

$$\lim_{N \to \infty} q_k = 0, \qquad \text{for all} \quad 1 \leqslant k \leqslant m.$$

The assumption of the context and array models that only some exemplars are encoded into memory (part i of Definition 5) reduces the effective sample sizes from which the category pdfs can be estimated. For example, in a experiment in which N_A and N_B category A and B exemplars are presented to the subject, respectively (where $N_A + N_B = N$), the context and array models predict that, on the average, πN_A exemplars from category A and πN_B exemplars from category B will be encoded into memory. So long as the probability that an exemplar is encoded does not depend on its similarity to other category exemplars, the assumption of imperfect encoding does not destroy random sampling. In other words, even with imperfect encoding, if the experimenter presents the subject with a random sample of category exemplars, the exemplars actually encoded by the subject will also form a random sample. Thus, the only effect of the imperfect encoding assumption of the general exemplar model is to reduce the effective sample size. Since consistency is an asymptotic property, it is not affected by the imperfect encoding assumption.

Theorem 10 is closely related to a result of Myung (1994, Prop. 4). He showed that if all q_k converge to zero as N approaches infinity and if encoding is perfect, then the context and array models are asymptotically equivalent to a process that infers the category pdfs using maximum entropy inference. The context and array models assume all stimulus features are binary, and in this special case, Myung (1994, Appendix B) showed that maximum entropy inference is always optimal. Thus, for binary features, the general exemplar model can be viewed as assuming subjects either use maximum entropy inference or estimate the category pdfs using a consistent nonparametric pdf estimator.

The proof of Theorem 10 constructs the pdf estimator directly. With continuous category distributions such construction is not possible. Even so, the exact conditions the kernel must satisfy before consistency is guaranteed are well known. These conditions are specified in the next result, which is due to Epanechnikov (1969).

THEOREM 11 (Epanechnikov, 1969). *Suppose* $\hat{f}_N(\mathbf{x})$ *is a product kernel estimator of the continuous pdf* $f(\mathbf{x})$*, with marginal kernel on dimension k equal to*

$$\frac{1}{h_k} \kappa_k \left(\frac{x_k - y_k}{h_k} \right).$$

Suppose, when every $h_k = 1$*, that each* κ_k *satisfies the following conditions*:

(i) $\kappa_k(y)$ *is finite-valued, for all y. Specifically, there exists some constant a, such that* $0 \leqslant \kappa_k(y) < a < \infty$*, for all y.*

(ii) $\kappa_k(y)$ *is a symmetric function about 0. Specifically,* $\kappa_k(y) = \kappa_k(-y)$*, for all y.*

(iii) *The variance associated with* $\kappa_k(y) = 1$*. Specifically,*

$$\int_{-\infty}^{\infty} y^2 \kappa_k(y)\, dy = 1.$$

(iv) *All moments of* $\kappa_k(y)$ *are finite. Specifically, for all n such that* $0 \leqslant n < \infty$*,*

$$\int_{-\infty}^{\infty} y^n \kappa_k(y)\, dy < \infty.$$

Then $\hat{f}_N(\mathbf{x})$ *is a consistent estimator of* $f(\mathbf{x})$ *if, for* $1 \leqslant k \leqslant m$*,*

$$\text{(a)} \quad \lim_{N \to \infty} h_k = 0, \quad \textit{and}$$

$$\text{(b)} \quad \lim_{N \to \infty} N \prod_{k=1}^{m} h_k = \infty.$$

Condition (a) requires that the kernel widths decrease as the sample size increases, but condition (b) says they cannot decrease too quickly. Specifically, the product of the kernel widths cannot decrease at a rate faster than that at which the sample size increases.

Theorem 11 can be used to establish the conditions under which the pdf estimator implicitly assumed by the generalized context model is consistent.

THEOREM 12. *Consider a categorization experiment in which all the category pdfs are continuous. The product kernel density estimator implicitly assumed by the generalized context model with* $r = \alpha$ *(Nosofsky, 1986) is a consistent estimator of the category pdfs if*

(a) *all attention weights are nonzero (i.e., $w_k > 0$, for all k),*

(b) *discriminability is nonzero (i.e., $C > 0$),*

(c) $\lim_{N \to \infty} C = \infty$, *and*

(d)

$$\lim_{N \to \infty} \frac{C^m \sqrt{\prod_{k=1}^{m} w_k}}{N} = 0, \quad \text{if} \quad r = \alpha = 2, \quad \text{or}$$

$$\lim_{N \to \infty} \frac{C^m \prod_{k=1}^{m} w_k}{N} = 0, \quad \text{if} \quad r = \alpha = 1.$$

The sufficiency conditions in Theorem 12 for the consistency of the generalized context model density function estimator are fairly weak. The subject must be able to discriminate between at least some of the stimuli (i.e., $C > 0$) and must allocate at least some attention to every relevant stimulus dimension (i.e., all $w_k > 0$). Condition (c) says that the discriminability parameter, C, must increase with experience. Equations (10) and (11) imply that the similarity between a pair of stimuli decreases as C increases. It makes sense that as a subject gains experience with a pair of categories stimuli begin to look more distinct and, as a result, less similar to other stimuli. Although C must increase with experience, condition (d) says it may not increase too quickly. The parameter mappings in Corollaries 1 and 2 make it clear that the w_k may change with experience, since they are functions of the h_k, which must decrease with experience. No strong conclusions can be drawn, however, about the direction of change of a particular w_k. Whether it increases or decreases with experience depends on the rates of decrease of the various h_k. Nevertheless, the product of the w_k may change with experience. Even so, since the sum of the w_k must equal 1, the product is restricted to the interval $(0, m^{-m})$. Because this interval is narrow, condition (d) of Theorem 12 limits the rate at which C may increase.

The requirement of Theorem 12 that C increase with experience is similar in spirit to the context model and array model consistency requirement of Theorem 10 that each of the q_k decrease with experience. With binary-valued dimensions, only one similarity is important—the similarity between the two levels on that dimension. In the array and context models, that similarity is denoted by q_k. If q_k decreases with experience, then the similarity of every pair of nonidentical values on dimension k decreases. Thus, decreasing q_k has exactly the same effect as increasing C.

When the stimulus dimensions are continuous, maximum entropy inference is not guaranteed to be optimal (Tribus, 1969). Thus, in the continuous case, assuming the subject uses maximum entropy inference is not equivalent to assuming the subject estimates the category pdfs using a consistent nonparametric pdf estimator. Myung (1994, Prop. 3) showed that the generalized context model with $r = \alpha = 2$ is asymptotically equivalent to a maximum entropy inference procedure only if all of the category pdfs are multivariate normal. In contrast, the results of Theorem 12 hold for *any* continuous category pdfs.

DISCUSSION

The results of the previous section show that the context model, the array model, and the generalized context model assume subjects are extremely good at categorization. First, the context and array models and the unbiased version of the generalized context model (with $r = \alpha$) each assume subjects use the minimum variance unbiased estimator of the category baserates. Although the proportion of stimuli belonging to category J may not seem like a complicated statistic, it assumes subjects keep a perfect count of the presented stimuli. For example, suppose a random exemplar is somehow lost from a subject's category representation. If the subject estimates the category baserates in the same fashion as before, the resulting estimates will still be unbiased, but they will no longer be of minimum variance. Second, all three models assume that subjects estimate the category distributions by using a nonparametric estimator that is among the most commonly used in the statistical community. Third, these estimators are consistent under the reasonable assumption that pairwise similarity decreases with training, and in the case of the generalized context model, that the subject is allocating at least some attention to every relevant stimulus dimension.

These three facts, together with Theorem 2, imply that all three exemplar models predict that with enough training, subjects will respond almost optimally in *any* categorization task, no matter how complex. The major source of suboptimal performance occurs because subjects probability match instead of use the likelihood ratio rule.

In contrast, the prototype and decision-bound models make strong assumptions about category structure. Specifically, they are each equivalent to a parametric classifier that assumes the category distributions are multivariate normal. The assumption of normality may be made either because many natural categories have this structure and subjects learn this through experience (Ashby, 1992a), or else because subjects only estimate category means, variances, correlations, and baserates and then infer the multivariate normal distribution through maximum entropy inference (Myung, 1994). In either case, these models predict that subjects will have difficulty in a categorization task with non-normal categories in which the optimal decision bound is neither linear nor quadratic.

On the other hand, note that the prototype and decision-bound models do not necessarily predict optimal performance in any categorization tasks, even those with normally distributed categories. This is because they make no specific

assumptions about how the subject will estimate the unknown category parameters. For example, the models do not rule out the possibility that subjects will act as if they are using estimators that are asymptotically biased. In this sense, these models are less detailed than the exemplar models.

A popular approach to comparing alternative categorization models is goodness-of-fit testing. The results of this article suggest that an alternative approach is to first ask whether human categorization performance is parametric or nonparametric. If it is parametric, what distributional family or families are assumed? What estimators of the unknown parameters does the subject use? Are these estimators unbiased? Are they consistent? If subjects are nonparametric classifiers, do they use kernel density estimators? If so, what type of kernel do they use? Are the kernel estimators consistent? Also, how do subjects estimate category baserates? The answers to these questions require a type of experimental and theoretical approach to the study of categorization different from that currently popular in the literature.

APPENDIX

Proof of Lemma 1. A sequence of estimators $\{\hat{\theta}_N\}$ of a parameter θ is consistent if the sequence $\{\hat{\theta}_N\}$ converges in probability to θ (e.g., see Neuts, 1973), and $\{\hat{\theta}_N\}$ converges in probability to θ if $\{\hat{\theta}_N\}$ converges in measure to θ (e.g., see Royden, 1968). When stated in terms of convergence in measure, the results of Lemma 1 are standard results in measure theory (e.g., see Halmos, 1950, pp. 92–95). Lemma 1 was first proved by Slutsky (1925).

Proof of Theorem 1. Suppose $L(\mathbf{x}) > \delta$. Then $f_A(\mathbf{x}) > \delta f_B(\mathbf{x})$ or, equivalently, $f_A(\mathbf{x}) - \delta f_B(\mathbf{x}) > 0$. Let $\hat{\theta}_N = \hat{f}_A(\mathbf{x}) - \hat{\delta}\hat{f}_B(\mathbf{x})$ and $\theta = f_A(\mathbf{x}) - \delta f_B(\mathbf{x})$. By parts (a) and (b) of Lemma 1, $\hat{\theta}_N$ is a consistent estimator of θ, and for all $\varepsilon > 0$,

$$
\begin{aligned}
1 &= \lim_{N\to\infty} P(|\hat{\theta}_N - \theta| < \varepsilon) \\
&= \lim_{N\to\infty} P(-\varepsilon < \hat{\theta}_N - \theta < \varepsilon) \\
&= \lim_{N\to\infty} P(\theta - \varepsilon < \hat{\theta}_N < \theta + \varepsilon).
\end{aligned}
$$

Since this limit equals 1 for any $\varepsilon > 0$, it must also equal 1 for any ε in the interval $0 < \varepsilon < \theta$. Therefore, for all $0 < \varepsilon < \theta$,

$$
\begin{aligned}
1 &= \lim_{N\to\infty} P(0 < \hat{\theta}_N < \theta + \varepsilon) \\
&\leqslant \lim_{N\to\infty} P(\hat{\theta}_N > 0).
\end{aligned}
$$

But since $P(\hat{\theta}_N > 0)$ is a probability, this last limit cannot exceed 1. Therefore

$$
\begin{aligned}
1 &= \lim_{N\to\infty} P(\hat{\theta}_N > 0) \\
&= \lim_{N\to\infty} P[\hat{f}_A(\mathbf{x}) - \hat{\delta}\hat{f}_B(\mathbf{x}) > 0] \\
&= \lim_{N\to\infty} P\left[\frac{\hat{f}_A(\mathbf{x})}{\hat{f}_B(\mathbf{x})} > \hat{\delta}\right]. \quad \blacksquare
\end{aligned}
$$

Proof of Theorem 2. Under the conditions of Theorem 2, the following three statements follow directly from results (a) and (b) of Lemma 1.

(i) $\hat{P}(A)\hat{f}_A(\mathbf{x})$ is a consistent estimator of $P(A) f_A(\mathbf{x})$,

(ii) $\hat{P}(B)\hat{f}_B(\mathbf{x})$ is a consistent estimator of $P(B) f_B(\mathbf{x})$,

(iii) $\hat{P}(A)\hat{f}_A(\mathbf{x}) + \hat{P}(B)\hat{f}_B(\mathbf{x})$ is a consistent estimator of $P(A) f_A(\mathbf{x}) + P(B) f_B(\mathbf{x})$.

Statements (i) and (iii), together with result (c) from Lemma 1, imply that

$$\frac{\hat{P}(A)\,\hat{f}_A(\mathbf{x})}{\hat{P}(A)\,\hat{f}_A(\mathbf{x}) + \hat{P}(B)\,\hat{f}_B(\mathbf{x})}$$

is a consistent estimator of

$$\frac{P(A)\,f_A(\mathbf{x})}{P(A)\,f_A(\mathbf{x}) + P(B)\,f_B(\mathbf{x})}.$$

$P(A) f_A(\mathbf{x}) + P(B) f_B(\mathbf{x})$ cannot equal zero for any stimulus X, since the stimuli are assumed to be sampled randomly from either category A or category B. The theorem follows immediately. $\blacksquare$

Proof of Theorem 4. Under conditions (a) and (b) of the theorem, the optimal classifier (i.e., a classifier that uses the likelihood ratio rule) uses a decision bound that satisfies (see, e.g., Ashby & Gott, 1988)

$$
\begin{aligned}
0 &= 2(\boldsymbol{\mu}_B - \boldsymbol{\mu}_A)'\,\Sigma^{-1}\mathbf{x} + (\boldsymbol{\mu}_A'\Sigma^{-1}\boldsymbol{\mu}_A - \boldsymbol{\mu}_B'\Sigma^{-1}\boldsymbol{\mu}_B) \\
&= 2(\boldsymbol{\mu}_B - \boldsymbol{\mu}_A)'\,\Sigma^{-1}\mathbf{x} - (\boldsymbol{\mu}_B'\Sigma^{-1}\boldsymbol{\mu}_B - \boldsymbol{\mu}_A'\Sigma^{-1}\boldsymbol{\mu}_A) \\
&= 2(\boldsymbol{\mu}_B - \boldsymbol{\mu}_A)'\,\Sigma^{-1}\mathbf{x} - (\boldsymbol{\mu}_B - \boldsymbol{\mu}_A)'\,\Sigma^{-1}(\boldsymbol{\mu}_B + \boldsymbol{\mu}_A).
\end{aligned}
$$

Denote the diagonal representation of Σ by WDW', where $W = [\mathbf{w}_1, \ldots, \mathbf{w}_m]$ is a matrix whose columns are the normalized eigenvectors of Σ and D is a diagonal matrix containing the eigenvalues of Σ. Using this representation, the optimal bound converts to

$$
\begin{aligned}
&2(\boldsymbol{\mu}_B - \boldsymbol{\mu}_A)'\,WD^{-1}W'\mathbf{x} \\
&\qquad - (\boldsymbol{\mu}_B - \boldsymbol{\mu}_A)'\,WD^{-1}W'(\boldsymbol{\mu}_B + \boldsymbol{\mu}_A) = 0. \quad \text{(A-1)}
\end{aligned}
$$

Now the simple prototype model predicts the minimum distance bound (see, e.g., Ashby, 1992a), which satisfies (see, e.g., Ashby & Gott, 1988)

$$2(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)'\,\mathbf{x}+(\boldsymbol{\mu}_A'\boldsymbol{\mu}_A-\boldsymbol{\mu}_B'\boldsymbol{\mu}_B)=0.$$

By hypothesis, one of the $\mathbf{w}_i$, say $\mathbf{w}_1$, must, therefore, be orthogonal to the hyperplane

$$(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)'\,\mathbf{x}=0.$$

Thus, $\mathbf{w}_1$ must be orthogonal to every vector in this hyperplane. Now the hyperplane is the set $\{\mathbf{x}\mid(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)'\,\mathbf{x}=0\}$ or, in other words, the set of all vectors orthogonal to $\boldsymbol{\mu}_B-\boldsymbol{\mu}_A$. Therefore, $\mathbf{w}_1$ must be coincident with the vector $\boldsymbol{\mu}_B-\boldsymbol{\mu}_A$. Specifically,

$$\begin{aligned}\mathbf{w}_1&=\frac{1}{\sqrt{(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)'\,(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)}}(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)\\&=\frac{1}{c}(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A).\end{aligned}$$

Now Eq. (A-1) can be written as

$$\begin{aligned}&2(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)'\,[\mathbf{w}_1,\ldots,\mathbf{w}_m]\,D^{-1}W'\mathbf{x}\\&\quad-(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)'\,[\mathbf{w}_1,\ldots,\mathbf{w}_m]\,D^{-1}W'(\boldsymbol{\mu}_B+\boldsymbol{\mu}_A)=0.\end{aligned}$$

Note that

$$(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)'\,\mathbf{w}_1=c$$

and for all $i=2,\ldots,m$

$$(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)'\,\mathbf{w}_i=0.$$

Therefore, Eq. (A-1) becomes

$$\begin{aligned}0&=2cd_1^{-1}\mathbf{w}_1'\mathbf{x}-cd_1^{-1}\mathbf{w}_1(\boldsymbol{\mu}_B+\boldsymbol{\mu}_A)\\&=2(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)'\,\mathbf{x}-(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)'\,(\boldsymbol{\mu}_B+\boldsymbol{\mu}_A)\\&=2(\boldsymbol{\mu}_B-\boldsymbol{\mu}_A)'\,\mathbf{x}+(\boldsymbol{\mu}_A'\boldsymbol{\mu}_A-\boldsymbol{\mu}_B'\boldsymbol{\mu}_B),\end{aligned}$$

which is the minimum distance bound. ∎

Proof of Theorem 5. The proof of the first part follows directly from results given by Ashby & Gott (1988). The second part is true, since for any decision bound generated by the general linear classifier, one can identify an infinite number of pairs of points both of which are equidistant from every point on the decision bound. Any one of these pairs could serve as the category prototypes. ∎

Proof of Theorem 6. The proof follows directly from results given by Ashby & Gott (1988). ∎

Proof of Theorem 7. Let N_A and N_B be the number of category A and B exemplars encoded into memory, respectively, and let $N=N_A+N_B$. In a classifier that probability matches with an estimator of the category baserates and a product kernel density estimator

$$P(R_A\mid X)=\frac{\hat{P}(A)\,\hat{f}_A(\mathbf{x})}{\hat{P}(A)\,\hat{f}_A(\mathbf{x})+\hat{P}(B)\,\hat{f}_B(\mathbf{x})},$$

where

$$\hat{P}(J)\,\hat{f}_J(\mathbf{x})=\frac{\hat{P}(J)}{N_J}\sum_{i=1}^{N_J}\prod_{k=1}^{m}\frac{1}{h_k}\kappa_k\left(\frac{x_k-y_{i_k}}{h_k}\right)$$

and where conditions (a) and (b) of Definition 4 are satisfied. In the general exemplar model

$$\begin{aligned}&P(R_A\mid X)\\&=\frac{\beta_A\sum_{Y\in A}\eta_{XY}}{\beta_A\sum_{Y\in A}\eta_{XY}+\beta_B\sum_{Y\in B}\eta_{XY}}\\&=\frac{\beta_A(\sum_{Y\in A}\eta_{XY}/\prod_{k=1}^{m}b_k)}{\beta_A(\sum_{Y\in A}\eta_{XY}/\prod_{k=1}^{m}b_k)+\beta_B(\sum_{Y\in B}\eta_{XY}/\prod_{k=1}^{m}b_k)},\end{aligned}$$

where b_k is defined by Eq. (12), and

$$\beta_J\left(\sum_{Y\in J}\eta_{XY}\Big/\prod_{k=1}^{m}b_k\right)=\beta_J\sum_{i=1}^{N_J}\prod_{k=1}^{m}\frac{1}{b_k}\eta_{x_k y_{i_k}}.$$

Thus, the general exemplar model is equivalent to probability matching with a product kernel estimator if

$$\hat{P}(J)=N_J\beta_J \tag{A-2}$$

and

$$\frac{1}{h_k}\kappa_k\left(\frac{x_k-y_{i_k}}{h_k}\right)=\frac{1}{b_k}\eta_{x_k y_{i_k}}=\frac{1}{b_k}s_k(x_k-y_{i_k})$$

and the following two conditions hold:

$$\text{(a)}\quad 1=\int K(\mathbf{x})\,d\mathbf{x}=\int\prod_{k=1}^{m}\frac{1}{b_k}s_k(x_k)\,d\mathbf{x}$$

$$\text{(b)}\quad 0=\int x_k K(\mathbf{x})\,d\mathbf{x}=\int x_k\times\prod_{i=1}^{m}\frac{1}{b_i}s_i(x_i)\,d\mathbf{x}\quad\text{for all } 1\leqslant i\leqslant m$$

We begin by establishing condition (a). Because of the product rule,

$$\int\prod_{k=1}^{m}\frac{1}{b_k}s_k(x_k)\,d\mathbf{x}=\prod_{k=1}^{m}\int\frac{1}{b_k}s_k(x_k)\,dx_k=1.$$

This latter equality holds since, by hypothesis,

$$b_k = \int s_k(x_k)\, dx_k.$$

Next we show that condition (b) is satisfied.

$$\begin{aligned}
&\int\int\cdots\int x_k \prod_{i=1}^{m} \frac{1}{b_i} s_i(x_i)\, d\mathbf{x} \\
&\quad = \left[\prod_{\substack{i=1 \\ i\neq k}} \int \frac{1}{b_i} s_i(x_i)\, dx_i\right] \frac{1}{b_k} \int x_k s_k(x_k)\, dx_k \\
&\quad = \frac{1}{b_k} \int x_k\, s_k(x_k)\, dx_k \\
&\quad = 0.
\end{aligned}$$

The latter equality holds since, by definition, the integral on the right converges and $s_k(x_k)$ is symmetric about zero.

To complete the proof, we must verify the claims about the baserate estimators. If responding is unbiased then, without loss of generality, we can set $\beta_A = \beta_B = 1/N$. In this case, we see from Eq. (A-2) that $\hat{P}(J)$ is the minimum variance unbiased estimator of $P(J)$. Suppose, however, that $\beta_A \neq \beta_B$. Now

$$E[\hat{P}(J)] = \beta_J E(N_J) = \beta_J N P(J).$$

Thus, $\hat{P}(J)$ is biased if $\beta_J \neq 1/N$. Since $\beta_A \neq \beta_B$, they both cannot equal $1/N$, so either $\hat{P}(A)$ or $\hat{P}(B)$ is biased. However, without loss of generality, one of the β_J can be set to $1/N$, so one $P(J)$ estimator is unbiased and of minimum variance. ∎

Proof of Theorem 8. The context model and the array model are special cases of the general exemplar model. Therefore, to prove the theorem we have to show only that Eq. (12) holds. From Eq. (9)

$$\int s_k(x_k)\, dx_k = \sum_{i=-1}^{1} s_k(i) = 1 + 2q_k,$$

which is a constant that does not depend on the target category. From Theorem 7,

$$\frac{1}{h_k}\kappa_k\left(\frac{x_k}{h_k}\right) = \frac{1}{b_k} s_k(x_k) = \begin{cases} \dfrac{q_k}{1+2q_k}, & \text{if } x_k = -1, \\ \dfrac{1}{1+2q_k}, & \text{if } x_k = 0, \\ \dfrac{q_k}{1+2q_k}, & \text{if } x_k = 1. \end{cases}$$ ∎

Proof of Theorem 9. If $r = \alpha$ the generalized context model is a general exemplar model with overall similarity function

$$\begin{aligned}
S(\mathbf{x}) &= \exp\left(-C^r \sum_{k=1}^{m} w_k\, |x_k|^r\right) \\
&= \prod_{k=1}^{m} \exp(-C^r w_k\, |x_k|^r).
\end{aligned}$$

Therefore, the component similarity function on dimension k equals

$$s_k(x_k) = \exp(-C^r w_k\, |x_k|^r).$$

Because of Theorem 7, to complete the proof we need only establish the condition specified by Eq. (11). Now

$$\begin{aligned}
\int_{-\infty}^{\infty} s_k(x_k)\, dx_k &= \int_{-\infty}^{\infty} \exp(-C^r w_k\, |x_k|^r)\, dx_k \\
&= 2\int_0^{\infty} \exp(-C^r w_k x_k^r)\, dx_k \\
&= \frac{2\Gamma(1/r)}{rCw_k^{1/r}} \\
&< \infty, \qquad \text{for all} \quad r > 0.
\end{aligned}$$

Note also that the value of the integral does not depend on the target category. ∎

Proof of Corollary 1. By Theorem 9, the generalized context model with $r = \alpha = 2$ is equivalent to probability matching with a product kernel density estimator. Since $r = \alpha = 2$, the overall similarity function equals

$$\begin{aligned}
S(\mathbf{x}) &= \exp\left(-C^2 \sum_{k=1}^{m} w_k x_k^2\right) \\
&= \prod_{k=1}^{m} \exp(-C^2 w_k x_k^2) \\
&= \prod_{k=1}^{m} s_k(x_k),
\end{aligned}$$

where $s_k(x_k)$ is the component similarity function for dimension k. Now, let

$$C^2 = \frac{\sum_{i=1}^{m} \prod_{j\neq i}^{m} h_j^2}{2\prod_{i=1}^{m} h_i^2} \qquad \text{and} \qquad w_k = \frac{\prod_{i\neq k}^{m} h_i^2}{\sum_{i=1}^{m} \prod_{j\neq i}^{m} h_j^2}.$$

Then

$$s_k(x_k) = \exp\left(-\frac{1}{2h_k^2} x_k^2\right).$$

Now

$$\int_{-\infty}^{\infty} s_k(x_k)\,dx_k = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2h_k^2}x_k^2\right)dx_k = \sqrt{2\pi}\,h_k.$$

Therefore, by Theorem 7

$$\frac{1}{h_k}\kappa_k\left(\frac{x_k}{h_k}\right) = \frac{1}{\sqrt{2\pi}\,h_k}\exp\left(-\frac{1}{2h_k^2}x_k^2\right),$$

which is the univariate normal pdf with mean 0 and standard deviation h_k. The expression for C^2 can be simplified by dividing the numerator and denominator by $\prod h_i^2$. ∎

Proof of Corollary 2. By Theorem 9, the generalized context model with $r=\alpha=1$ is equivalent to probability matching with a product kernel density estimator. Since $r=\alpha=1$, the overall similarity function equals

$$\begin{aligned} S(\mathbf{x}) &= \exp\left(-C\sum_{k=1}^{m} w_k|x_k|\right) \\ &= \prod_{k=1}^{m}\exp(-Cw_k|x_k|) \\ &= \prod_{k=1}^{m} s_k(x_k). \end{aligned}$$

Now, let

$$C = \frac{\sqrt{2}\sum_{i=1}^{m}\prod_{j\neq i}^{m} h_j}{\prod_{i=1}^{m} h_i} \quad \text{and} \quad w_k = \frac{\prod_{i\neq k}^{m} h_i}{\sum_{i=1}^{m}\prod_{j\neq i}^{m} h_j}.$$

Then

$$s_k(x_k) = \exp\left(-\frac{\sqrt{2}}{h_k}|x_k|\right).$$

Now

$$\int_{-\infty}^{\infty} s_k(x_k)\,dx_k = \int_{-\infty}^{\infty}\exp\left(-\frac{\sqrt{2}}{h_k}|x_k|\right)dx_k = \sqrt{2}\,h_k.$$

Therefore, by Theorem 7

$$\frac{1}{h_k}\kappa_k\left(\frac{x_k}{h_k}\right) = \frac{1}{\sqrt{2}\,h_k}\exp\left(-\frac{\sqrt{2}}{h_k}|x_k|\right),$$

which is the univariate Laplace pdf with mean 0 and standard deviation h_k. The expression for C can be simplified by dividing the numerator and denominator by $\prod h_i$. ∎

Proof of Theorem 10. As in the proof of Theorem 7, let N equal the total number of exemplars encoded into memory. Now, without loss of generality we can assume $h_k=1$ for all k in the kernel assumed by the array and context models. In this case, the models assume

$$\begin{aligned} \hat{f}_N(\mathbf{x}) &= \frac{1}{N}\sum_{t=1}^{N} K(\mathbf{x}-\mathbf{y}_t) \\ &= \frac{1}{N}\sum_{t=1}^{N}\prod_{k=1}^{m}\kappa_k(x_k-y_{t_k}), \end{aligned}$$

where

$$\kappa_k(x_k-y_{t_k}) = \begin{cases} p_k, & \text{if } x_k \neq y_{t_k}, \\ 1-2p_k, & \text{if } x_k = y_{t_k}. \end{cases}$$

Now,

$$\prod_{k=1}^{m}\kappa_k(x_k-y_{t_k}) = \begin{cases} \prod_{k=1}^{m}(1-2p_k), & \text{if } x_k = y_{t_k}, \\ \quad \text{for all } k & \\ p_i\prod_{k\neq i}^{m}(1-2p_k), & \text{if } x_i \neq y_{t_i},\ x_k = y_{t_k}, \\ \quad \text{for all } k\neq i & \\ \quad\vdots \qquad\qquad\vdots & \\ \prod_{k=1}^{m} p_k, & \text{if } x_k \neq y_{t_k}, \\ \quad \text{for all } k. & \end{cases}$$

Let $N_{ij\cdots m}$ equal the number of stored exemplars that match the stimulus on dimension $i, j, \ldots, m$. Note that $i, j, \ldots, m$ need not be consecutive integers. Then

$$\hat{f}_N(\mathbf{x}) = \frac{1}{N}\left[N_{\varnothing}\prod_{k=1}^{m} p_k + \cdots + N_{12\cdots(m-1)}p_m \times\prod_{k=1}^{m-1}(1-2p_k) + N_{12\cdots m}\prod_{k=1}^{m}(1-2p_k)\right].$$

Let $\theta_{ij\cdots m}$ equal the probability that an exemplar randomly drawn from the category matches the stimulus on dimensions $i, j, \ldots, m$. Then by Lemma 1, $\hat{f}_N(\mathbf{x})$ is a consistent estimator of

$$\theta_{\varnothing}\prod_{k=1}^{m} p_k + \cdots + \theta_{12\cdots(m-1)}p_m \times\prod_{k=1}^{m-1}(1-2p_k) + \theta_{12\cdots m}\prod_{k=1}^{m}(1-2p_k).$$

Now $\theta_{12\cdots m}=f(\mathbf{x})$. Therefore, $\hat{f}_N(\mathbf{x})$ is a consistent estimator of

$$f(\mathbf{x})\prod_{k=1}^{m}(1-2p_k)+\theta_{\varnothing}\prod_{k=1}^{m}p_k + \cdots +\theta_{12\cdots(m-1)}p_m\prod_{k=1}^{m-1}(1-2p_k).$$

Each term in this expression is independent of N. Therefore, if any $p_k>0$, then $\hat{f}_N(\mathbf{x})$ is an asymptotically biased estimator of $f(\mathbf{x})$ and thus is not a consistent estimator of $f(\mathbf{x})$. However, if for all k(i.e., for $1\leqslant k\leqslant m$)

$$\lim_{N\to\infty}p_k=0 \tag{A-3}$$

then

$$\lim_{N\to\infty}\left[f(\mathbf{x})\prod_{k=1}^{m}(1-2p_k)+\theta_{\varnothing}\prod_{k=1}^{m}p_k + \cdots +\theta_{12\cdots(m-1)}p_m\prod_{k=1}^{m-1}(1-2p_k)\right]=f(\mathbf{x}),$$

so, under the conditions of Eq. (A-3), $\hat{f}_N$ is a consistent estimator of $f(\mathbf{x})$. From Theorem 8 it is clear that Eq. (A-3) holds if and only if

$$\lim_{N\to\infty}q_k=0.$$

Finally, if $\pi=1$ and all $q_k=0$, then all $p_k=0$ and N equals the total number of exemplars presented by the experimenter. In this case,

$$\hat{f}_N(\mathbf{x})=\frac{N_{12\cdots m}}{N},$$

which is just the proportion of stored exemplars exactly equal to the stimulus X. This is the minimum variance unbiased estimator of $f(\mathbf{x})$. ∎

Proof of Theorem 12. There are two parts to the proof. First, we must show that the four conditions of Theorem 11 are satisfied. Second, we must establish the implications on the parameters of the generalized context model of results (a) and (b) of Theorem 11.

Part 1. When all the $h_k=1$, the marginal kernels in the $(r=\alpha=1)$ version of the generalized context model have a Laplace distribution, whereas the marginal kernels in the $(r=\alpha=2)$ version have a normal distribution. In both cases, the mean is zero and the variance is 1. Thus, conditions (i), (ii), and (iii) of Theorem 11 are satisfied. Condition (iv) is also satisfied because all moments of the standard Laplace and normal distributions are finite (see, e.g., Johnson & Kotz, 1970a, 1970b).

Part 2. Condition (b) of Theorem 11 implies that no h_k may be zero. The parameter mappings given in Corollaries 1 and 2 indicate that one or more $w_k=0$ only if one or more of the $h_k=0$. Therefore, consistency requires all $w_k>0$. The parameter mappings also show that $C=0$, only if all h_k are infinite. Condition (a) of Theorem 11 implies that all h_k are finite-valued, so $C>0$. If condition (a) of Theorem 11 holds, then it is clear from the parameter mappings given in Corollaries 1 and 2 that condition (c) of Theorem 12 must hold. Condition (d) of Theorem 12 follows from condition (b) of Theorem 11 and the fact that when $r=\alpha=2$

$$C^{2m}\prod_{k=1}^{m}w_k=\frac{1}{2^m\prod_{k=1}^{m}h_k^2}.$$

This identity follows directly from results given in the proof of Corollary 1; When $r=\alpha=1$, then

$$C^{m}\prod_{k=1}^{m}w_k=\frac{2^{m/2}}{\prod_{k=1}^{m}h_k}.$$

This identity follows directly from results given in the proof of Corollary 2. ∎

ACKNOWLEDGMENTS

This research was supported in part by National Science Foundation Grant DBS92-09411 to the first author. Some of these results were presented at the Twenty-Sixth Annual Mathematical Psychology Meetings. We thank William Batchelder, Xiangen Hu, Michael Kalish, and In Jae Myung for their helpful comments and suggestions.

REFERENCES

Ashby, F. G. (1992a). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Erlbaum.

Ashby, F. G. (1992b). Multivariate probability distributions. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 1–34). Hillsdale, NJ: Erlbaum.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**, 33–53.

Ashby, F. G., & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General*, **120**, 150–172.

Ashby, F. G., & Lee, W. W. (1992). On the relationship between identification similarity, and categorization: Reply to Nosofsky and Smith (1992). *Journal of Experimental Psychology: General*, **3**, 385–393.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, **37**, 372–400.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, **93**, 154–179.

Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological Review*, **74**, 1–15.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization*. Hillsdale, NJ: Erlbaum.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, **14**, 153–158.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, **18**, 500-549.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review*, **83**, 37–64.

Estes, W. K. (1994). *Classification and cognition*. Oxford: Oxford Univ. Press.

Fix, E., & Hodges, J. L. (1951). *Discriminatory analysis. Nonparametric discrimination; consistency properties* (Rep. No. 4, Project No. 21-49-004), Randolph Field, TX: USAF School of Aviation Medicine.

Flannagan, M. J., Fried, L. S., & Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **12**, 241–256.

Fried, L. S., & Holyoak, F. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 234–257.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, **117**, 225–244.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Halmos, P. R. (1950). *Measure theory*. Princeton, NJ: Van Nostrand.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, **4**, 267–272.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, **13**, 243–266.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.

Hurwitz, J. B. (1990). *A hidden-pattern unit network model of category learning*. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Johnson, N. L., & Kotz, S. (1970a). *Continuous univariate distributions* (Vol. 1). New York: Wiley.

Johnson, N. L., & Kotz, S. (1970b). *Continuous univariate distributions* (Vol. 2). New York: Wiley.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22–44.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, **53**, 49–70.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207–238.

Myung, I. J. (1994). Maximum entropy interpretation of decision bound and context models of categorization. *Journal of Mathematical Psychology*, **38**, 335–365.

Neuts, M. R. (1973). *Probability*. Boston: Allyn & Bacon.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 104–114.

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

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, **34**, 393–418.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **35**, 1065–1076.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353–363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, **83**, 304–308.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, **3**, 382–407.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, **4**, 328–350.

Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Studies in cross-cultural psychology*. London: Academic Press.

Royden, H. L. (1968). *Real analysis*. New York: Macmillan.

Schaller, G. B. (1979). *Stones of silence*. New York: Viking.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. New York: Wiley.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, **22**, 325–345.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Silverman, B. W., & Jones, M. C. (1989). E. Fix and J. L. Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation. *International Statistical Review*, **57**, 233–247.

Slutsky, E. E. (1925). Über stochastische Asymptoten und Grenzwerte, *Metron*, **5**, 3–89.

Tribus, M. (1969). The principle of maximum entropy. In M. Tribus (Ed.), *Rational descriptions, decisions and designs* (Chap. 5). New York: Pergamon.

Received: February 11, 1994