Fitting Decision Bound Models to Identification or Categorization Data

Daniel M. Ennis

The Institute for Perception

F. Gregory Ashby

University of California at Santa Barbara

August 25, 2003

Correspondence: F. Gregory Ashby
Department of Psychology
University of California
Santa Barbara, CA 93106

Phone: 805-893-2130
Fax: 805-893-4303
Email: ashby@psych.ucsb.edu

Running Head: Fitting the Decision Bound Model

Abstract

Decision bound models are derived from general recognition theory, which is a multidimensional generalization of signal detection theory. This broad class of models has been remarkably successful at accounting for data from categorization and identification experiments. Fitting decision bound models however, requires the numerical evaluation of multiple integrals over the multidimensional normal distribution. This article describes an extremely general algorithm for fitting decision bound models to identification or categorization data that is both fast and straightforward to implement. The algorithm is based on a mathematical technique called Cholesky factorization.

Fitting Decision Bound Models to Identification or Categorization Data

On each trial of a categorization task, a stimulus is selected randomly from an ensemble of n stimuli and presented to the observer. The observer responds with one of m response alternatives preselected by the experimenter, where $m < n$. An identification task is identical, except $m = n$. Data from both types of task are typically summarized in an $n \times m$ confusion matrix containing in row i and column j the frequency of R_j responses on trials when stimulus S_i was presented. Among the most successful models at accounting for data from categorization and identification experiments are decision bound models (Ashby & Lee, 1991, 1992; Ashby, Lee, & Balakrishnan, 1992; Ashby & Perrin, 1988; Ashby & Townsend, 1986; Ashby & Waldron, 1999; Ashby, Waldron, Lee, & Berkman, 2001; Ennis & Ashby, 1993; Maddox, 2001, 2002; Maddox & Ashby, 1993, 1996; Maddox & Bogdanov, 2000; Maddox & Dodd, 2003; Maddox, Molis, & Diehl, 2002).

Decision bound models are derived from general recognition theory (Ashby & Townsend, 1986), which is a multidimensional generalization of signal detection theory (e.g., Green & Swets, 1966). Decision bound models assume the trial-by-trial perceptual effects of a stimulus can be represented by a multivariate normal distribution. The models also assume that the observer partitions the perceptual space into m regions (where m is the number of response alternatives) and associates a unique response alternative with each region. The partitions between these regions are called decision bounds. The percept on a trial when stimulus S_i is presented is represented as a random sample from the perceptual distribution associated with S_i. The observer is assumed to determine which region the percept is in and then to emit the associated response.

Current applications of the model place no constraints on any parameters of the perceptual distributions and a wide variety of different types of decision bounds have been tested. These include models in which 1) the decision bounds are arbitrary linear functions (the general linear classifier), 2) the decision bounds are arbitrary quadratic functions (the general quadratic classifier), 3) the observer is assumed to give the response associated with the nearest perceptual mean (the minimum distance classifier) or nearest striatal grid point (the striatal pattern classifier), and 4) the decision bounds maximize response accuracy (the optimal classifier).

In all versions of the model, however, each frequency in the predicted confusion matrix is obtained by integrating under some multivariate normal distribution throughout a specific response region. Except for a few restricted special cases, the integration must be performed numerically. For this reason, application of this class of models has been restricted to only a few laboratories. This article describes an extremely general algorithm for fitting decision bound models to identification or categorization data that is both fast and straightforward to implement.

EVALUATING INTEGRALS WHEN THERE IS A SINGLE DECISION BOUND

Suppose the perceptual space is r dimensional and let $\mathbf{x}_i$ denote the $r \times 1$ dimensional random vector describing the perceptual effect on trials when stimulus S_i is presented. The model assumes $\mathbf{x}_i$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}_i$ and variance-covariance matrix Σ_i. Figure 1 shows a contour of equal likelihood from a hypothetical 2-dimensional perceptual distribution in which the perceived values along the two perceptual dimensions are positively correlated. Figure 1 also shows a quadratic decision bound that divides the perceptual plane into two regions. Suppose the region associated with response R_j is above the bound. Then the probability of responding R_j on S_i trials is equal to the proportion of the bivariate normal distribution illustrated in Figure 1 that falls above the quadratic bound.

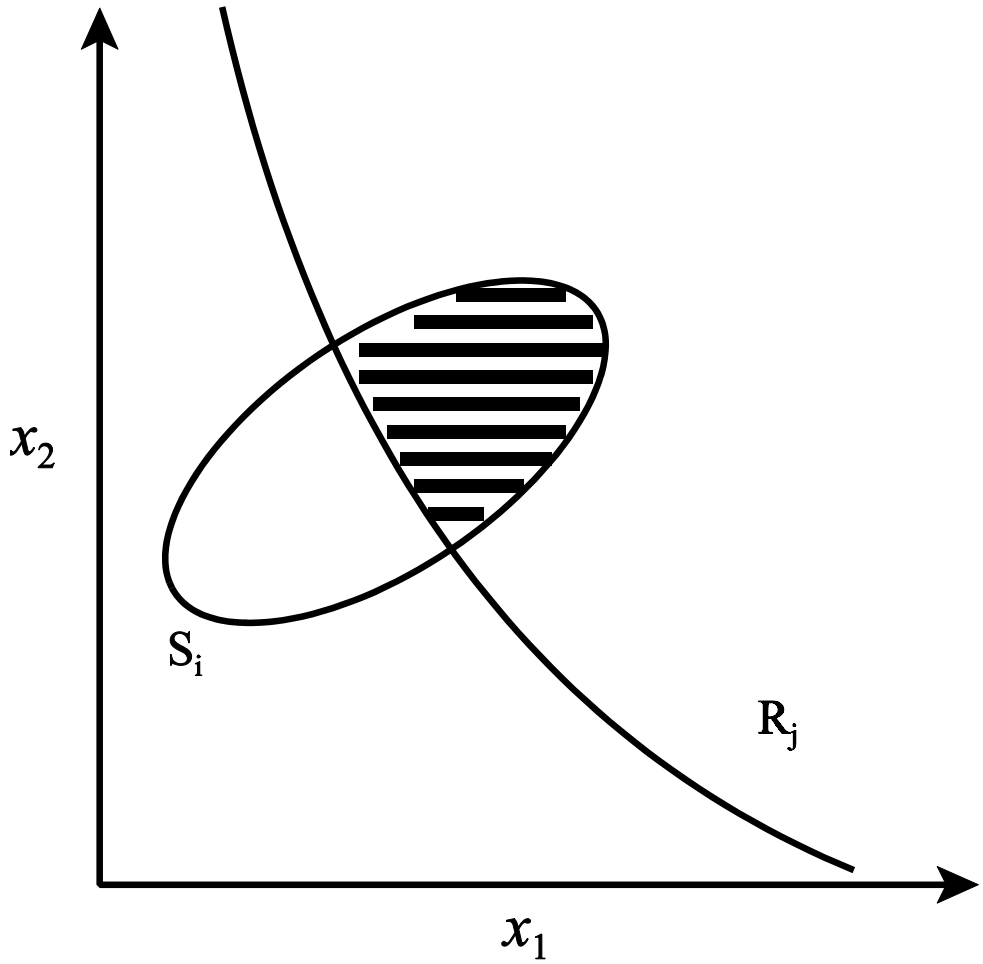


Figure 1. A contour of equal likelihood from a 2-dimensional perceptual distribution in which the perceptual variates x_1 and x_2 are positively correlated. Also shown is a quadratic decision bound that divides the perceptual plane into two response regions. The region associated with response R_j is above the bound.

Since the bound is quadratic, all points on the bound must satisfy the equation

$$h(\mathbf{x}) = \mathbf{x}'\mathrm{A}\mathbf{x} + \mathbf{b}'\mathbf{x} + c = 0 \tag{1}$$

where A is a symmetric matrix of constants, **b** is a vector of constants, and c is a scalar. $h(\mathbf{x})$ can be viewed as a discriminant function because for all **x** below the bound, $h(\mathbf{x}) < 0$, and for all **x** above the bound, $h(\mathbf{x}) > 0$. Because of this property

$$P(\mathrm{R}_j{*}\mathrm{S}_i) = \int\int_{\mathbf{C}_x} \cdots \int \mathrm{mvn}(\boldsymbol{\mu}_i, \Sigma_i)\, d\mathbf{x} \tag{2}$$

where $\mathrm{mvn}(\boldsymbol{\mu}_i, \Sigma_i)$ is the probability density function of a multivariate normal distribution with mean vector $\boldsymbol{\mu}_i$ and variance-covariance matrix Σ_i and $\mathrm{C}_{\mathbf{x}} = \{\mathbf{x}^{\#}\, h(\mathbf{x}) > 0\}$.

To evaluate the Eq. 2 integral, one could 1) step through a rectangular grid of points in **x**-space, 2) at each point compute $h(\mathbf{x})$, 3) for each **x** for which $h(\mathbf{x}) > 0$, compute the height of the mvn($\boldsymbol{\mu}_i$, Σ_i) probability density function (pdf), and 4) increment the integral by an amount equal to the volume of the appropriate rectangular right prism. The main problem with this approach is that the grid in step 1 has to cover the perceptual distributions associated with all stimuli in the experiment, which means that in most applications it would be prohibitively large. The method we propose solves this problem and also eliminates step 3.

The first step is to standardize the Eq. 2 integral.

PROPOSITION 1. *Suppose the r-dimensional random vector* $\mathbf{x}_i$ *has a multivariate normal distribution with mean vector* $\boldsymbol{\mu}_i$ *and variance-covariance matrix* Σ_i. *Let* **z** *denote an r-dimensional multivariate normal vector with mean vector* **0** *and variance-covariance matrix* I *(where* I *is the identity matrix). Then, if there exists a matrix* P_i *such that* $\mathbf{x} = P_i\,\mathbf{z} + \boldsymbol{\mu}_i$, *then*

$$P(R_j^* S_i) = \iiint_{C_x} \text{mvn}(\boldsymbol{\mu}_i, \Sigma_i)\, d\mathbf{x}$$

$$= \iiint_{C_z} \text{mvn}(\mathbf{0}, I)\, d\mathbf{z}$$

where $C_z = \{\mathbf{z}^* h(P_i\,\mathbf{z} + \boldsymbol{\mu}_i) > 0\}$. ~

Thus, for any mean vector and variance-covariance matrix of the S_i perceptual distribution, $P(R_j \mid S_i)$ can always be computed by integrating over the multivariate z-distribution. Because of this, we can lay an initial grid over z-space that can be used for all stimuli and all parameter estimates. Note also, that if $h(\mathbf{x})$ is the quadratic function of Eq. 1, then $h(P_i\,\mathbf{z} + \boldsymbol{\mu}_i)$ is also quadratic. In fact,

$$h(P_i\,\mathbf{z} + \boldsymbol{\mu}_i) = h^*(\mathbf{z}) = \mathbf{z}'(P_i'AP_i)\mathbf{z} + (2\boldsymbol{\mu}_i' AP_i + \mathbf{b}'P_i)\mathbf{z} + (\boldsymbol{\mu}_i' A\boldsymbol{\mu}_i + \mathbf{b}'\boldsymbol{\mu}_i + c) \tag{3}$$

To use Proposition 1 we must find a matrix P_i such that

$$\mathbf{x}_i = P_i\,\mathbf{z} + \boldsymbol{\mu}_i\,. \tag{4}$$

Equation 4 holds if and only if

$$\Sigma_i = P_i\,\Sigma_{\mathbf{z}}\,P_i' = P_i\,P_i' \tag{5}$$

so the matrix P_i can be found from Eq. 5. There are many solutions to this problem. For example, let Q be the matrix whose columns are the eigenvectors of Σ_i, and let D be the diagonal matrix whose entries are the corresponding eigenvalues of Σ_i. Then,

$$\Sigma_i = Q\,D\,Q' = Q\,D^{1/2}\,D^{1/2}\,Q' = (Q\,D^{1/2})\,(Q\,D^{1/2})'$$

so one solution is $P_i = Q\,D^{1/2}$. Unfortunately, eigenvectors and eigenvalues are not quickly computed, so this solution is not ideal. The next result provides a much more efficient solution.

PROPOSITION 2. *For any variance-covariance matrix* Σ_i, *there exists a unique lower triangular matrix* P_i, *such that* $\Sigma_i = P_i P_i'$. *This is called the Cholesky factorization of* Σ_i.

PROOF. *See Graybill (1976, pp. 231) or Johnson and Kotz (1970, chapter 29).*

The advantage of Cholesky factorization is that the Cholesky matrix P_i is easily obtained. To see this, note that $\Sigma_i = P_i P_i'$ implies

$$\begin{bmatrix} \sigma_1^2 & \text{cov}_{12} & \dots & \text{cov}_{1r} \\ \text{cov}_{21} & \sigma_2^2 & \dots & \text{cov}_{2r} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ \text{cov}_{r1} & \text{cov}_{r2} & \dots & \sigma_r^2 \end{bmatrix} = \begin{bmatrix} p_{11} & 0 & \dots & 0 \\ p_{21} & p_{22} & \dots & 0 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ p_{r1} & p_{r2} & \dots & p_{rr} \end{bmatrix} \begin{bmatrix} p_{11} & p_{21} & \dots & p_{r1} \\ 0 & p_{22} & \dots & p_{r2} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 0 & 0 & \dots & p_{rr} \end{bmatrix}$$

Solving for the p_{ij} leads to:

$$1)\ p_{11} = \sigma_1$$

$$2)\ p_{j1} = \frac{\text{cov}_{j1}}{\sigma_1}, \quad for \quad j = 2,3,\dots,r$$

$$3)\ p_{ii} = \left(\sigma_i^2 - \sum_{k=1}^{i-1} p_{ik}^2 \right)^{1/2}, \quad for \quad i = 2,3,\dots,r$$

$$4)\ p_{ij} = \frac{1}{p_{jj}} \left(\text{cov}_{ij} - \sum_{k=1}^{j-1} p_{ik} p_{jk} \right), \quad for \quad j < i \quad and \quad i = 2,\dots,r-1$$

$$5)\ p_{ij} = 0, \quad for \quad j > i \quad and \quad i = 2,3,\dots,r$$

Using these results, evaluation of the Eq. 2 integral is accomplished via the following steps.

1) Create a z-grid. The z-grid is created by selecting n z-values, each of which is the midpoint of an equal area interval. For an integral over the bivariate normal, we typically select 100 z-values on each dimension. These values are chosen before the integral is evaluated and stored in an array. Call the i<u>th</u> value in this array z_i. An r-dimensional grid is created by taking the Cartesian product of the z- array r times. For example, with a bivariate integral and 100 values in the z-array, the z-grid contains 10,000 points (i.e., 100^2). Each point is the center of a rectangle whose volume under the bivariate z-distribution equals .0001 (i.e., 1/10,000).

2) Use Σ_i to find the Cholesky matrix P_i (i.e., using the solution provided above).

3) Find the image of the Eq. 2 quadratic bound in z-space (i.e., using the solution provided in Eq. 3). Call this function $h^*(\mathbf{z})$.

4) Compute the integral. This is done by stepping through the n^r points in the z-grid. Call the kth of these points $\mathbf{z}_k$. First, compute the numerical value of $h^*(\mathbf{z}_k)$. If $h^*(\mathbf{z}_k) < 0$, then go to the next point in the z-grid. If $h^*(\mathbf{z}_k) > 0$, then increment the value of the integral by $1/n^r$.

The only time-limiting operation in this procedure is the computation of $h^*(\mathbf{z}_k)$. By

selecting the z-grid beforehand with the property that each point is the center of an equal volume region, much computation is avoided.

MULTIPLE DECISION BOUNDS

When there are multiple decision bounds, the algorithm described in the last section can be used in a slightly generalized form. Consider the case when there are m different decision bounds. Each of these can be described by a discriminant function. For example, the kth bound satisfies $h_k(\mathbf{x}) = 0$. The discriminant functions can always be defined in such a way that $h_k(\mathbf{0}) < 0$. Every point in the x-space will be associated with either a positive or negative discriminant value for each discriminant function[1]. Thus, every point can be characterized by a string of m pluses and minuses. The sign of the kth of these depends on whether the point is on the same side of the kth bound as the origin, or on the opposite side. Note also that every point in the response region is associated with the same pattern of pluses and minuses, because every point in the region has the same relation to the origin with respect to each bound as every other point.

As an example, consider the model illustrated in Figure 2. The four decision bounds define four discriminant functions h_1, h_2, h_3, and h_4. Every point in the region associated with response R_1 generates a negative value on each discriminant function. Thus, this region is associated with the pattern (-,-,-,-). Similarly, every point in the R_5 region is associated with the pattern (+,-,+,-). The nine conditional response probabilities $P(R_j|S_i)$, for j = 1, 2, ..., 9, can be computed simultaneously in one pass through the z-grid. During step 4 of the algorithm, at each new z-point all four discriminant functions are computed and a match is performed between the obtained pattern of pluses and minuses and the patterns associated with each of the nine response regions. The integral associated with the matching region is then incremented by $1/n^r$. Thus, computation time is only weakly related to the number of response alternatives. Instead, the more important factor is the number of stimuli, since each new stimulus requires a new pass through the z-grid.

The pattern matching can be performed quickly using integer arithmetic, which is typically much faster than if the same operations were performed in real or double-precision mode. We illustrate the method using the example of Figure 2. To begin, for each of the four discriminant functions, define a new variable δ_i with the property that $\delta_i = 1$ if $h_i(\mathbf{x}) > 0$ and $\delta_i = 0$ if $h_i(\mathbf{x}) < 0$. Next, create the variable

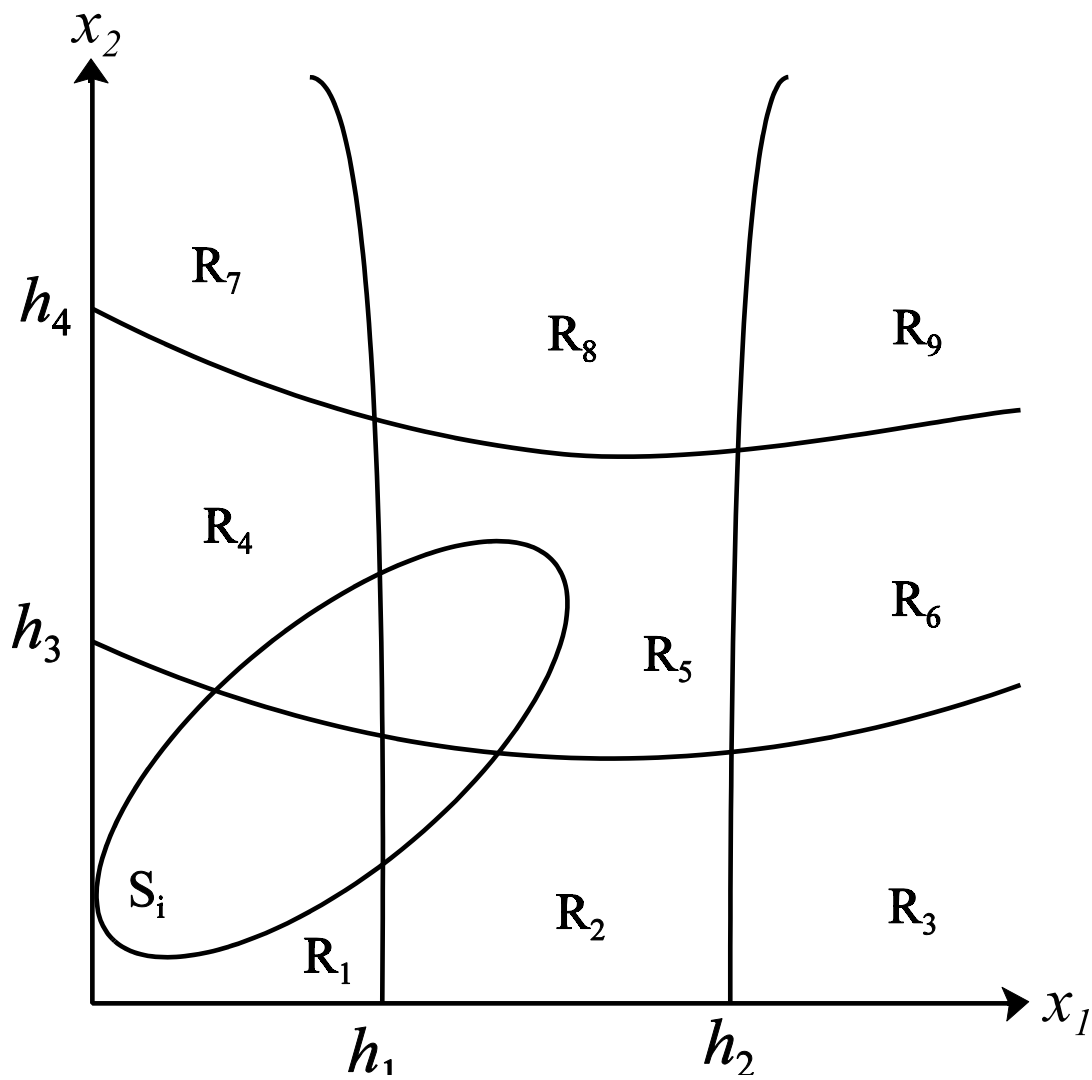
$$T = 1 + \delta_1 + 2\,\delta_2 + 4\,\delta_3 + 8\,\delta_4 \tag{6}$$

The variable T associates a unique integer value (from 1 to 16) with each of the nine Figure 2 response regions. The intermediate values of the integrals associated with the nine conditional response probabilities can be stored in an array indexed by the variable T. For example, for every point in region R_4, $T = 5$, so when the evaluation of a z-point leads to $T = 5$, the value in the fifth position of the integral array can be incremented by $1/n^2$.

Figure 2. A contour of equal likelihood from the perceptual distribution associated with stimulus S_i and 4 decision bounds that divide the perceptual plane into 9 response regions.

---

[1] The discriminant value equals zero with probability zero.



FITTING THE OPTIMAL CLASSIFIER, THE MINIMUM DISTANCE CLASSIFIER, AND THE STRIATAL PATTERN CLASSIFIER

The method described above requires that each decision bound is specified by a discriminant function. The discriminant functions are transformed to z-space and the integral is then computed. In some models, however, the decision rule is not easily described in terms of discriminant functions. There are three obvious examples: the optimal classifier, the minimum distance classifier, and a model introduced by Ashby and Waldron (1999) called the striatal pattern classifier (SPC). We will begin by considering the optimal classifier, and then since the minimum distance classifier and the SPC are computationally similar, we will consider these two models together.

When there are many stimuli, the simplest way to describe the rule used by the optimal classifier is that at each point **x**, the likelihoods are all computed (e.g., $f_i(\mathbf{x})$, for $i = 1, 2, ..., 9$),

and then the response associated with the largest of these is given. When the stimulus presentation probabilities are all equal this decision rule maximizes response accuracy. To fit this model, the algorithm described above needs only to be altered slightly. The linear transformation, $\mathbf{z} = P_i^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i)$, carries optimal bounds to optimal bounds, so the identity of the stimulus associated with the maximum likelihood in x-space is preserved by the Cholesky transformation. As a result, the conditional response probabilities associated with stimulus S_i can be computed using a modified version of the Cholesky algorithm in which steps 3 and 4 are replaced with:

3) Transform all perceptual distributions to z-space. To compute the conditional response probabilities for stimulus S_i, the S_i perceptual distribution is standardized via the transformation $\mathbf{z} = P_i^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i)$. After applying this same transformation to the other perceptual distributions, the S_k distribution is multivariate normal with mean vector $P_i^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)$ and variance-covariance matrix $P_i^{-1} \Sigma_k P_i^{-1\prime}$.

4) Compute the integral. The integral can now be computed by stepping through z-space. At each point in the z-grid, the likelihood is computed with respect to each of the transformed perceptual distributions. The integral associated with the largest of these is incremented by $1/n^r$.

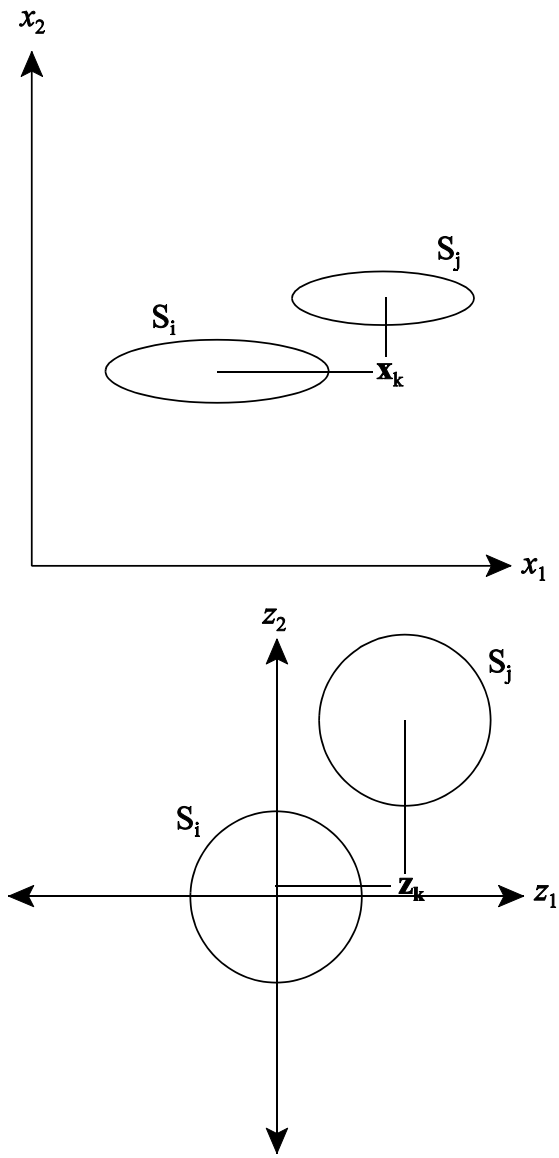
The minimum distance classifier and the SPC are similar to the optimal classifier, except that instead of computing likelihoods, some distances are computed, and the response associated with the smallest of these is given. In the minimum distance classifier, the distances are from the percept to each perceptual mean, whereas in the SPC the distances are from the percept to arbitrary points that mark the centroid of each response region. Unfortunately, the analogous algorithm does not work for these models (i.e., where distance is substituted for likelihood in step 4). This is because distance relations are not preserved by the Cholesky transformation. This is easily seen in Figure 3. In the x-space the point $\mathbf{x}_k$ is closest to the S_j perceptual mean, so the minimum distance classifier responds R_j to this percept. However, after the Cholesky transformation $\mathbf{z} = P_i^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_i)$, the point in z-space that corresponds to $\mathbf{x}_k$ (i.e., $\mathbf{z}_k$) is closest to the S_i perceptual mean. As a consequence, the conditional response probabilities predicted by the minimum distance classifier are most easily computed in x-space. Even so, for both the minimum distance classifier, steps 1 and 2 of the Cholesky algorithm are still used. However, steps 3 and 4 must be replaced by:

3) Transform the z-grid to x-space. Step through the z-grid. At each z-point, compute the corresponding x-point via the transformation $\mathbf{x}_i = P_i \mathbf{z} + \boldsymbol{\mu}_i$.

4) Compute the integral. For each point in x-space, compute the distance to every perceptual mean (in the case of the minimum distance classifier) or to every striatal grid point (in the case of the SPC) and then increment the integral associated with the smallest of these by $1/n^r$.

This algorithm will be considerably slower than an analogous algorithm based on discriminant functions because of the linear transformation that is required on every point in the z-grid. On the other hand, the minimum distance/SPC algorithm is exceedingly general. With appropriate modification, it can be used to fit *any* model that assumes multivariate normal perceptual distributions and that specifies a well-defined decision rule.

Figure 3. An illustration that distance relations are not preserved by the linear transformation from the x-space to the z-space. In the top panel, the point $\mathbf{x}_k$ is closer to the mean of the S_j perceptual distribution than to the mean of the S_i distribution, but the bottom panel shows that this distance ordering is reversed after the transformation to z-space.



CONCLUSIONS

For many years, the most successful model of stimulus identification was the so-called similarity-choice model (Luce, 1963; Shepard, 1957). For example, in 1992, J. E. K. Smith summarized its performance by concluding that the similarity-choice model "has never had a serious competitor as a model of identification data" (p. 199). However, in recent comparisons to decision bound models, whose fits were made possible by the algorithm described in this article, the similarity-choice model has fared poorly. For example, Ashby et al. (2001) fit the similarity-choice model and a variety of decision bound models to 25 separate data sets collected in a variety of different experiments. In 20 of these 25 cases the best account of the data was provided by a decision bound model, and in many of these instances the improvement provided by the decision bound model over the similarity-choice model was substantial.

Despite these successes, applications of decision bound models have been confined to just a few laboratories, at least partly because of the time-consuming and seemingly sophisticated computation needed to fit the model. The algorithm described in this article is conceptually straightforward and runs quickly on modern computers, and we believe it largely solves this problem.

REFERENCES

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 among identification, similarity, and categorization: Reply to Nosofsky and Smith (1992). *Journal of Experimental Psychology: General*, *121*, 385-393.

Ashby, F. G., Lee, W. W., & Balakrishnan, J. D. (1992). Comparing the biased choice model and multidimensional decision bound models of identification. *Mathematical Social Sciences*, *23*, 175-197.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review, 95*, 124-15.

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

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*, 363-378.

Ashby, F. G., Waldron, E. M., Lee, W. W., & Berkman, A. (2001). Suboptimality in human categorization and identification. *Journal of Experimental Psychology: General*, *130*, 77-96.

Ennis, D. M., & Ashby, F. G. (1993). The relative sensitivities of same-different and identification judgement models to perceptual dependence. *Psychometrika*, *58*, 257-279.

Graybill, F. A. (1976). *Theory and application of the linear model*. North Scituate, MA: Duxbury Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Huntington, NY: Krieger.

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

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103-190). New York: Wiley.

Maddox, W. T. (2001). Separating perceptual processes from decisional processes in identification and categorization. *Perception & Psychophysics*, *63*, 1183-1200.

Maddox, W. T. (2002). Learning and attention in multidimensional identification, and categorization: Separating low-level perceptual processes and high level decisional processes. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *28*, 99-115.

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

Maddox, W. T., & Ashby, F. G. (1996). Perceptual separability, decisional separability, and the identification- speeded classification relationship. *Journal of Experimental Psychology: Human Perception & Performance*, *22*, 795-817.

Maddox, W. T., & Bogdanov, S. V. (2000). On the relation between decision rules and perceptual representation in multidimensional perceptual categorization. *Perception & Psychophysics*, *62*, 984-997.

Maddox, W. T., & Dodd, J. L. (2003). Separating perceptual and decisional attention processes in the identification and categorization of integral-dimension stimuli. *Journal of Experimental Psychology: Learning , Memory, and Cognition*, *29*, 467-480.

Maddox, W. T., Molis, M. R, & Diehl, R. L. (2002). Generalizing a neuropsychological model of visual categorization to auditory categorization of vowels. *Perception & Psychophysics*,

64, 584-597.
Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325-345.
Smith, J. E. K. (1992). Alternative biased choice models. *Mathematical Social Sciences*, *23*, 199-219.

Author Notes

This research was supported in part by Public Health Service Grant MH3760. We thank Kelle Karp for her help in manuscript preparation. Correspondence concerning this article should be addressed to F. Gregory Ashby, Department of Psychology, University of California, Santa Barbara, CA 93106 (e-mail: ashby@psych.ucsb.edu), or Daniel M. Ennis, The Institute for Perception, 7629 Hull Street Road, Ste 200, Richmond, VA 23235 (email: ifpress@cs.com).