PSYCHOMETRIKA—VOL. 58, NO. 2, 315–330
JUNE 1993

A LATENT CLASS APPROACH TO FITTING THE WEIGHTED EUCLIDEAN MODEL, CLASCAL

SUZANNE WINSBERG

IRCAM, PARIS, FRANCE

GEERT DE SOETE

UNIVERSITY OF GHENT, BELGIUM

A weighted Euclidean distance model for analyzing three-way proximity data is proposed that incorporates a latent class approach. In this latent class weighted Euclidean model, the contribution to the distance function between two stimuli is per dimension weighted identically by all subjects in the same latent class. This model removes the rotational invariance of the classical multidimensional scaling model retaining psychologically meaningful dimensions, and drastically reduces the number of parameters in the traditional INDSCAL model. The probability density function for the data of a subject is posited to be a finite mixture of spherical multivariate normal densities. The maximum likelihood function is optimized by means of an EM algorithm; a modified Fisher scoring method is used to update the parameters in the M-step. A model selection strategy is proposed and illustrated on both real and artificial data.

Key words: weighted Euclidean distance model, INDSCAL, latent class analysis, mixture distribution model, EM algorithm.

Introduction

Multidimensional scaling is a procedure in which dissimilarity data arising from N sources each relating J objects pairwise, is modeled to fit distances in some type of space, generally Euclidean of low dimensionality R. The INDSCAL or weighted Euclidean distance model proposed by Carroll and Chang (1970) removes the rotational invariance existing in the classical Euclidean model proposed by Torgerson (1958) and Gower (1966), thus providing the user with dimensions that are potentially psychologically meaningful. Equation (1) represents the weighted Euclidean distance model and (2) the classical model:

$$y_{ijk} \approx d_{ijk} = \left[\sum_{r=1}^{R} w_{ir}(x_{jr} - x_{kr})^2 \right]^{1/2}, \tag{1}$$

$$y_{ijk} \approx d_{ijk} = \left[\sum_{r=1}^{R} (x_{jr} - x_{kr})^2 \right]^{1/2}, \tag{2}$$

where x_{jr} is the coordinate of the j-th stimulus on the r-th dimension ($j = 1, \ldots, J$); w_{ir} is the weight for the r-th dimension associated with the i-th source ($i = 1, \ldots, N$); d_{ijk} is the model distance between the j-th and the k-th stimulus from the i-th

The second author is supported as "Bevoegdverklaard Navorser" of the Belgian "Nationaal Fonds voor Wetenschappelijk Onderzoek".

Requests for reprints should be sent to Geert De Soete, Department of Data Analysis, University of Ghent, Henri Dunantlaan 2, B-9000 Ghent, Belgium.

0033-3123/93/0600-92430$00.75/0

© 1993 The Psychometric Society

source ($j, k = 1, \ldots, J; j \neq k$); and y_{ijk} is the observed dissimilarity between stimuli j and k from source i. The weights w_{ir} in (1) are required to be nonnegative:

$$w_{ir} \geq 0. \tag{3}$$

The explicit modeling of individual differences through the w_{ir} parameters and the rotational uniqueness of the object coordinates undoubtedly account for the popularity of the INDSCAL model among users. Most often the N different sources represent N subjects from whom dissimilarity data are obtained. In this case (or in any case where N is large), the cost of removing the rotational invariance thus obtaining ease of interpretation is the introduction of many nuisance parameters (the individual subject weights w_{ir}). These weights are rarely interpreted for individual subjects, and the improvement in goodness-of-fit measures seldom seems to justify so many additional parameters. Therefore, we propose a latent class approach to this problem, removing the rotational invariance, retaining psychologically meaningful dimensions, and drastically reducing the number of parameters in the INDSCAL model.

Latent class formulations, or more general mixture distribution approaches, have recently been explored in the context of various uni- and multidimensional scaling models for paired comparisons data (Böckenholt & Böckenholt, 1990; De Soete, 1990; De Soete & Winsberg, 1993; Formann, 1989), pick any/n data (Böckenholt & Böckenholt, 1990, 1991; De Soete & DeSarbo, 1991), and single stimulus preference data (DeSarbo, Howard, & Jedidi, 1991; De Soete & Winsberg, in press; De Soete & Heiser, 1993). In these applications, latent class modeling has proven to be a viable technique for capturing systematic group differences in a parsimonious way.

The CLASCAL Model

Begin with J stimuli, $M = J(J-1)/2$ is the number of stimulus pairs. Let the M-component column vector $\mathbf{y}_i = (y_{i21}, y_{i31}, y_{i32}, \ldots, y_{iJ(J-1)})'$ contain the M dissimilarity values for subject i ($i = 1, \ldots, N$). The total data set will be indicated by the $N \times M$ matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)'$.

In the latent class approach, we assume that each of the N subjects belongs to one and only one of a small number T ($T \ll N$) of latent classes or subpopulations. It is not known in advance to which latent class a particular subject i belongs. The (unconditional) probability that any subject belongs to latent class t will be denoted $\lambda_t (1 \leq t \leq T)$, with, of course,

$$\sum_{t=1}^{T} \lambda_t = 1. \tag{4}$$

The column vector $\boldsymbol{\lambda}$ is defined as $(\lambda_1, \ldots, \lambda_T)'$.

Latent class analysis was originally developed for categorical data and assumed that, within each latent class, the data are distributed according to a product of independent Bernoulli distributions (Lazarsfeld & Henry, 1968). Here we are dealing with continuous proximity data. Instead of assuming independent Bernoulli distributions, we will assume independent normal distributions that have a common variance. Hence, our model assumes a finite mixture of spherical multivariate normal distributions and is consequently a special case of the general mixture model of multivariate normal distributions (see McLachlan & Basford, 1988). However, since the present mixture model relies on the same local independence assumption as traditional latent class analysis, it can be considered as a latent class model for continuous rating data (see De

Soete, in press, for a further discussion of this point). Note that the local independence assumption (i.e., independence between stimulus pairs *within* each latent class) does not imply global independence between the stimulus pairs. Rather it implies that any overall correlation between the stimulus pairs is due to the fact that the subjects belong to different latent classes. Hence, the local independence assumption on which CLASCAL model is based, is weaker than the global independence usually assumed in maximum likelihood multidimensional scaling procedures (Ramsay, 1977, 1982; Winsberg & Carroll, 1989a, 1989b). Global independence is also implicitly assumed in most least squares multidimensional scaling methods. As noted by Ramsay (1991, p. 65), with rating data, violations of the independence assumption are usually not serious enough to warrant concern. The assumption of a *common* variance for all M stimulus pairs is consistent with the common practice of fitting multidimensional scaling models by means of *unweighted* least squares methods. A stochastic model based on independent normal distributions with a common variance is one of the models incorporated in MULTISCALE (Ramsay, 1991) and has been assumed in other maximum likelihood multidimensional scaling methods (Winsberg & Carroll, 1989a, 1989b). In the final section, we discuss how the assumption of a common variance can be relaxed in a straightforward way. Note that while it might be tempting to assume a finite mixture of general multivariate normal densities instead of a finite mixture of spherical (or elliptical) normal densities, estimating the $M(M + 1)/2$ parameters of a covariance matrix between the $M = J(J - 1)/2$ stimulus pairs would be unwieldy, unless the number of subjects is extremely large. For instance, with 12 objects, 2211 additional parameters would have to be estimated!

Thus, it is assumed that for a particular subject i in latent class t, the data $\mathbf{y}_i$ are independently normally distributed with means $\boldsymbol{\delta}_t = (\delta_{t21}, \delta_{t31}, \delta_{t32}, \ldots, \delta_{tJ(J-1)})'$ and common variance σ^2:

$$\mathbf{y}_i \sim N(\boldsymbol{\delta}_t, \sigma^2\mathbf{I}) \text{ for subject } i \text{ in class } t. \tag{5}$$

In the latent class weighted Euclidean distance model, the elements δ_{tjk} of the M-component vector $\boldsymbol{\delta}_t$ are defined as

$$\delta_{tjk} = \left(\sum_{r=1}^{R} w_{tr}(x_{jr} - x_{kr})^2 \right)^{1/2}, \tag{6}$$

where $\mathbf{w}_t = (w_{t1}, \ldots, w_{tR})'$ denotes the INDSCAL-type weights for latent class t. For each latent class t, we estimate a separate set of weights $\mathbf{w}_t$. These weights are constrained to be nonnegative:

$$w_{tr} \geq 0. \tag{7}$$

The stimulus configuration $\mathbf{X}$ and the variance parameter σ^2, on the contrary, are assumed to be the same for all T latent classes. The R dimension weights for the T latent classes will be denoted as the $T \times R$ matrix $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_T)'$.

To fully identify the latent class weighted Euclidean distance model, the following constraints are imposed on the stimulus coordinates and the latent class weights:

$$\sum_{t=1}^{T} w_{tr} = T, \text{ for } r = 1, \ldots, R, \tag{8}$$

$$\sum_{j=1}^{J} x_{jr} = 0, \text{ for } r = 1, \ldots, R. \tag{9}$$

The latent class weighted Euclidean distance model has $T + 1 + J \cdot R + T \cdot R$ parameters corresponding to $\boldsymbol{\lambda}$, σ^2, $\mathbf{X}$, and $\mathbf{W}$, respectively. By subtracting from the number of model parameters the number of constraints that are imposed on these parameters through (4), (8), and (9), the degrees of freedom of the model are obtained

$$T + (T + J - 2)R. \tag{10}$$

When $T = 1$, it is necessary to subtract from (10) $T(T - 1)/2$ for the rotational indeterminacy that occurs in this case.

This model is denoted as CLASCAL to distinguish it from INDSCAL. When $T = N$, CLASCAL is equivalent to the INDSCAL model defined in (1) and when $T = 1$ it is equivalent to the classical Euclidean model defined in (2). When $1 < T \ll N$, CLASCAL is a generalization of both the weighted and the unweighted Euclidean distance model in that it more parsimonious that the former and more interpretable than the latter (due to the rotational invariance property).

Parameter Estimation

Likelihood Function

Because of (5), the probability density function (pdf) of the data of a subject i that belongs to latent class t can be written as

$$f(\mathbf{y}_i|\mathbf{X}, \mathbf{w}_t, \sigma^2) = (\sigma\sqrt{2\pi})^{-M} \exp\left[-\frac{(\mathbf{y}_i - \boldsymbol{\delta}_t)'(\mathbf{y}_i - \boldsymbol{\delta}_t)}{2\sigma^2}\right]. \tag{11}$$

Since it is not known in advance to which latent class a particular subject i belongs, the pdf of $\mathbf{y}_i$ becomes a finite mixture of multivariate normal densities:

$$g(\mathbf{y}_i|\mathbf{X}, \mathbf{W}, \sigma^2, \boldsymbol{\lambda}) = \sum_{t=1}^{T} \lambda_t f(\mathbf{y}_i|\mathbf{X}, \mathbf{w}_t, \sigma^2)$$

$$= (\sigma\sqrt{2\pi})^{-M} \sum_{t=1}^{T} \lambda_t \exp\left[-\frac{(\mathbf{y}_i - \boldsymbol{\delta}_t)'(\mathbf{y}_i - \boldsymbol{\delta}_t)}{2\sigma^2}\right]. \tag{12}$$

Maximum likelihood estimates of the parameters $\mathbf{X}$, $\mathbf{W}$, σ^2, and $\boldsymbol{\lambda}$ can be obtained by maximizing the likelihood function

$$L(\mathbf{X}, \mathbf{W}, \sigma^2, \boldsymbol{\lambda}|\mathbf{Y}) = \prod_{i=1}^{N} g(\mathbf{y}_i|\mathbf{X}, \mathbf{W}, \sigma^2, \boldsymbol{\lambda})$$

$$= (\sigma\sqrt{2\pi})^{-N \cdot M} \prod_{i=1}^{N} \left\{ \sum_{t=1}^{T} \lambda_t \exp\left[-\frac{(\mathbf{y}_i - \boldsymbol{\delta}_t)'(\mathbf{y}_i - \boldsymbol{\delta}_t)}{2\sigma^2}\right] \right\}, \tag{13}$$

subject to (4), (7), (8), and (9).

Once parameter estimates $\hat{\mathbf{X}}$, $\hat{\mathbf{W}}$, $\hat{\sigma}^2$, and $\hat{\boldsymbol{\lambda}}$ are obtained, the a posteriori probability that a subject i belongs to latent class t can be computed by means of Bayes' theorem. This a posteriori probability, written as $h_{it}(\hat{\mathbf{X}}, \hat{\mathbf{W}}, \hat{\sigma}^2, \hat{\boldsymbol{\lambda}})$, equals

$$h_{it}(\hat{\mathbf{X}}, \hat{\mathbf{W}}, \hat{\sigma}^2, \hat{\boldsymbol{\lambda}}) = \frac{\hat{\lambda}_t f(\mathbf{y}_i|\hat{\mathbf{X}}, \hat{\mathbf{w}}_t, \hat{\sigma}^2)}{g(\mathbf{y}_i|\hat{\mathbf{X}}, \hat{\mathbf{W}}, \hat{\sigma}^2, \hat{\boldsymbol{\lambda}})}$$

$$= \frac{\hat{\lambda}_t \exp\left[-\dfrac{(\mathbf{y}_i - \hat{\boldsymbol{\delta}}_t)'(\mathbf{y}_i - \hat{\boldsymbol{\delta}}_t)}{2\hat{\sigma}^2}\right]}{\sum_{s=1}^{T} \hat{\lambda}_s \exp\left[-\dfrac{(\mathbf{y}_i - \hat{\boldsymbol{\delta}}_s)'(\mathbf{y}_i - \hat{\boldsymbol{\delta}}_s)}{2\hat{\sigma}^2}\right]}. \tag{14}$$

A subject can then be assigned to the class for which the a posteriori membership probability is the largest.

Estimation Algorithm

As in many mixture distribution problems (McLachlan & Basford, 1988), the likelihood function (13) is most easily optimized by means of an EM algorithm (Dempster, Laird, & Rubin, 1977). To enable an EM algorithm formulation, some non-observed data are introduced:

$$z_{it} = \begin{cases} 1 & \text{iff subject } i \text{ belongs to class } t, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

The column vector $\mathbf{z}_i$ is defined as $(z_{i1}, \ldots, z_{iT})'$ and the N by T matrix $\mathbf{Z}$ as $(\mathbf{z}_1, \ldots, \mathbf{z}_N)'$. It is assumed that the non-observed data $\mathbf{z}_i$ are independently and identically multinomially distributed with probabilities $\boldsymbol{\lambda}$, that is,

$$(\mathbf{z}_i|\boldsymbol{\lambda}) \sim \prod_{t=1}^{T} \lambda_t^{z_{it}}. \tag{16}$$

The distribution of $\mathbf{y}_i$ given $\mathbf{z}_i$ is

$$(\mathbf{y}_i|\mathbf{z}_i, \mathbf{X}, \mathbf{W}, \sigma^2, \boldsymbol{\lambda}) \sim \sum_{t=1}^{T} z_{it} f(\mathbf{y}_i|\mathbf{X}, \mathbf{w}_t, \sigma^2)$$

$$\sim \prod_{t=1}^{T} f(\mathbf{y}_i|\mathbf{X}, \mathbf{w}_t, \sigma^2)^{z_{it}}. \tag{17}$$

The log likelihood of the complete data $\mathbf{Y}$ and $\mathbf{Z}$ can now be written as

$$\log L_C(\mathbf{X}, \mathbf{W}, \sigma^2, \boldsymbol{\lambda}|\mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} \log f(\mathbf{y}_i|\mathbf{X}, \mathbf{w}_t, \sigma^2)$$

$$+ \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} \log \lambda_t. \tag{18}$$

The EM algorithm alternates between an E-step (expectation step) and an M-step (maximization step) in an iterative fashion. In the E-step, the expectation of $\log L_C$ needs to be calculated over the conditional distribution of the non-observed data $\mathbf{Z}$, given the observed data $\mathbf{Y}$ and provisional estimates $\mathbf{X}^{(0)}$, $\mathbf{W}^{(0)}$, $\sigma^{2(0)}$, and $\boldsymbol{\lambda}^{(0)}$ of the parameters. This expectation is

$$Q(\mathbf{X}, \mathbf{W}, \sigma^2, \boldsymbol{\lambda}, \mathbf{X}^{(0)}, \mathbf{W}^{(0)}, \sigma^{2(0)}, \underline{\boldsymbol{\lambda}^{(0)}}) = \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it}^{(0)} \log f(\mathbf{y}_i | \mathbf{X}, \mathbf{w}_t, \sigma^2)$$

$$+ \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it}^{(0)} \log \lambda_t, \tag{19}$$

with

$$z_{it}^{(0)} = h_{it}(\mathbf{X}^{(0)}, \mathbf{W}^{(0)}, \sigma^{2(0)}, \boldsymbol{\lambda}^{(0)}), \tag{20}$$

(see, e.g., McLachlan & Basford, 1988). Thus, in the E-step, the nonobserved $\mathbf{Z}$ are replaced by the a posteriori probabilities calculated on the basis of the provisional parameter estimates $\mathbf{X}^{(0)}$, $\mathbf{W}^{(0)}$, $\sigma^{2(0)}$, and $\boldsymbol{\lambda}^{(0)}$.

In the M-step, $Q(\mathbf{X}, \mathbf{W}, \sigma^2, \boldsymbol{\lambda}, \mathbf{X}^{(0)}, \mathbf{W}^{(0)}, \sigma^{2(0)}, \boldsymbol{\lambda}^{(0)})$ must be maximized with respect to $\mathbf{X}$, $\mathbf{W}$, σ^2, and $\boldsymbol{\lambda}$ to obtain new provisional parameter estimates. To maximize (19) with respect to $\mathbf{X}$ and $\mathbf{w}_1, \ldots, \mathbf{w}_T$ subject to (7), (8), and (9), it suffices to minimize q_1 with respect to these parameters (subject to the same constraints):

$$q_1(\mathbf{X}, \mathbf{W}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j<k}^{J} z_{it}^{(0)} (y_{ijk} - \delta_{tjk})^2$$

$$= \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j<k}^{J} z_{it}^{(0)} (y_{ijk} - \bar{y}_{tjk})^2 + \sum_{t=1}^{T} \sum_{j<k}^{J} F_t (\bar{y}_{tjk} - \delta_{tjk})^2, \tag{21}$$

with

$$\bar{y}_{tjk} = \frac{\sum_{i=1}^{N} z_{it}^{(0)} y_{ijk}}{\sum_{i=1}^{N} z_{it}^{(0)}}, \tag{22}$$

and

$$F_t = \sum_{i=1}^{N} z_{it}^{(0)}. \tag{23}$$

Because of the orthogonal decomposition in (21), q_1 is minimized with respect to $\mathbf{X}$ and $\mathbf{W}$ whenever q_1^* is minimal:

$$q_1^*(\mathbf{X}, \mathbf{W}) = \sum_{t=1}^{T} \sum_{j<k}^{J} F_t(\bar{y}_{tjk} - \delta_{tjk})^2. \tag{24}$$

Hence, in the M-step, it suffices to minimize q_1^* subject to (7), (8), and (9) to obtain new estimates of $\mathbf{X}$ and $\mathbf{W}$. The function q_1^* is minimized by alternating between optimizing the spatial model parameters $\mathbf{X}$ conditional on the weights $\mathbf{W}$, and optimizing the weights $\mathbf{W}$ conditional on the spatial parameters $\mathbf{X}$ until convergence occurs. In both the spatial parameter estimation substep and the weight estimation substep, we use a modified Fisher scoring method with $-\mathbf{H}^+\mathbf{g}$ as the search direction, where $\mathbf{H}^+$ is the Moore-Penrose inverse of the expected Hessian $E(\nabla^2 q_1^*)$ and $\mathbf{g}$ is the gradient ∇q_1^*. In the weight estimation substep, the weights are kept nonnegative by means of an active constraint algorithm described by Winsberg and Ramsay (1983). A safeguarded quadratic interpolation method is utilized to determine the stepsize. The algorithm used in the M-step is very similar to the numerical estimation method used in other maximum likelihood multidimensional scaling procedures (e.g., Winsberg & Carroll, 1989a).

Once new estimates of $\mathbf{X}$ and $\mathbf{W}$ are available, a new estimate of σ^2 can be computed as follows

$$\hat{\sigma}^2 = \frac{1}{NM} \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it}^{(0)} (\mathbf{y}_i - \hat{\boldsymbol{\delta}}_t)'(\mathbf{y}_i - \hat{\boldsymbol{\delta}}_t), \tag{25}$$

where $\hat{\boldsymbol{\delta}}_t$ denote the mean dissimilarities based on the new estimates of $\mathbf{X}$ and $\mathbf{W}$.

In the M-step, we also need to compute a new estimate of $\boldsymbol{\lambda}$. Maximizing (19) with respect to $\boldsymbol{\lambda}$ subject to (4) gives

$$\hat{\lambda}_t = \frac{F_t}{N}. \tag{26}$$

Schematically, the EM algorithm can be outlined as follows:

1. Initialize the iteration index α: $\alpha \leftarrow 0$.
   Obtain initial parameter estimates $\mathbf{X}^{(\alpha)}$, $\mathbf{W}^{(\alpha)}$, $\sigma^{2(\alpha)}$, and $\boldsymbol{\lambda}^{(\alpha)}$.
2. *E-step.* Compute $\mathbf{Z}^{(\alpha)} = ((z_{it}^{(\alpha)}))$ with

$$z_{it}^{(\alpha)} = \frac{\lambda_t^{(\alpha)} \exp\left[-\dfrac{(\mathbf{y}_i - \boldsymbol{\delta}_t^{(\alpha)})'(\mathbf{y}_i - \boldsymbol{\delta}_t^{(\alpha)})}{2\sigma^{2(\alpha)}}\right]}{\displaystyle\sum_{s=1}^{T} \lambda_s^{(\alpha)} \exp\left[-\frac{(\mathbf{y}_i - \boldsymbol{\delta}_s^{(\alpha)})'(\mathbf{y}_i - \boldsymbol{\delta}_s^{(\alpha)})}{2\sigma^{2(\alpha)}}\right]}.$$

3. *M-step.* Compute new estimates of $\mathbf{X}$ and $\mathbf{W}$ by minimizing the weighted least squares function (24) with

$$\bar{y}_{tjk} = \frac{\displaystyle\sum_{i=1}^{N} z_{it}^{(\alpha)} y_{ijk}}{F_t},$$

and

$$F_t = \sum_{i=1}^{N} z_{it}^{(\alpha)},$$

subject to (7), (8), and (9).
Compute a new estimate of σ^2 using (25), where $\hat{\delta}_t$ is based on the new estimates of $\mathbf{X}$ and $\mathbf{W}$.
Update the elements of $\boldsymbol{\lambda}$:

$$\lambda_t^{(\alpha+1)} = \frac{1}{N} \sum_{i=1}^{N} z_{it}^{(\alpha)}.$$

4. Test for convergence: if

$\log L(\mathbf{X}^{(\alpha+1)}, \mathbf{W}^{(\alpha+1)}, \sigma^{2(\alpha+1)}, \boldsymbol{\lambda}^{(\alpha+1)}|\mathbf{Y}) - \log L(\mathbf{X}^{(\alpha)}, \mathbf{W}^{(\alpha)}, \sigma^{2(\alpha)}, \boldsymbol{\lambda}^{(\alpha)}|\mathbf{Y}) < \varepsilon_1$, or
$\log L(\mathbf{X}^{(\alpha+1)}, \mathbf{W}^{(\alpha+1)}, \sigma^{2(\alpha+1)}, \boldsymbol{\lambda}^{(\alpha+1)}|\mathbf{Y}) -$
$\log L(\mathbf{X}^{(\alpha-4)}, \mathbf{W}^{(\alpha-4)}, \sigma^{2(\alpha-4)}, \boldsymbol{\lambda}^{(\alpha-4)}|\mathbf{Y}) < \varepsilon_2$,

terminate. (Currently, $\varepsilon_1 = 0$ and $\varepsilon_2 = 0.001$ is used.)
5. Increment α: $\alpha \leftarrow \alpha + 1$.
Go back to Step 2.

Initial parameter estimates can be computed by performing a K-means clustering on the dissimilarity data $\mathbf{Y}$ (with $K \equiv T$). The means of the M variables for cluster t can be used as $\bar{y}_{tjk}$ in (24) to arrive at initial estimates of $\mathbf{X}$ and $\mathbf{W}$. These initial estimates can be used to compute an initial estimate of σ^2 using (25). An alternative method for computing initial parameter estimates is to perform an INDSCAL (Carroll & Chang, 1970) analysis on $\mathbf{Y}$ and to group the subjects into latent classes on the basis of the derived individual subject weights. Alternatively, the subjects may be grouped randomly to start. With each procedure, the relative cluster sizes can be used as an initial estimate of $\boldsymbol{\lambda}$.

The algorithm described in this section has been implemented in a transportable Fortran program which is available upon request.

Choosing the Appropriate Model

In most cases one does not know the appropriate number of classes T nor the appropriate number of dimensions R in advance. The usual procedure in maximum likelihood multidimensional scaling for choosing the number of dimensions for a spatial model with no weights (or alternatively put, one class) is to compare the AIC (Akaike, 1977) and BIC (Schwarz, 1978) statistics and choose the number of dimensions corresponding to the minimum value of these statistics. Here, AIC and BIC statistics may not be used to select among solutions with differing numbers of latent classes, because in such case the regularity conditions are not satisfied (see McLachlan & Basford, 1988). However, conditional on a given number of latent classes, the regularity conditions obtain, and one may select the appropriate spatial model or dimensionality for one class, two classes, etcetera, using a likelihood ratio test or information criteria. Furthermore, the usual procedure for deciding on the appropriate number of latent classes would require testing whether a solution for $T + 1$ latent classes gives a significantly better fit than a solution for the same model (that is the same dimensionality) with T classes. Unfortunately, because in such case the regularity conditions do not hold, the relevant likelihood-ratio statistic for testing T versus $T + 1$ latent classes is not as-

TABLE 1

Goodness-of-fit Results for the Artificial Data Example

No. of Classes	No. of Dimensions	Model Degrees of Freedom	Log Likelihood	AIC	BIC
1	2	16	300.5	-568.9	-495.7
1	3	22	301.0	-558.0	-457.3
1	4	27	301.1	-548.2	-424.6
2	2	20	670.1	-1300.1	-1208.5
2	3	29	672.6	-1287.2	-1154.4

ymptotically distributed as a chi-square with known degrees of freedom and neither likelihood-ratio tests nor information criteria such as AIC and BIC can be used. However, conditional on a given dimensionality, one may use a Monte Carlo significance testing procedure proposed by Hope (1968) and first applied in the context of latent class analysis by Aitkin, Anderson, and Hinde (1981). The procedure is as follows: (a) determine the parameter estimates $\hat{\mathbf{X}}$, $\hat{\mathbf{W}}$, $\hat{\sigma}^2$, $\hat{\boldsymbol{\lambda}}$ for a T-class model; (b) draw $S - 1$ random samples $\bar{\mathbf{Y}}$ of size N from the T-class population with parameters $\hat{\mathbf{X}}$, $\hat{\mathbf{W}}$, $\hat{\sigma}^2$, $\hat{\boldsymbol{\lambda}}$; (c) fit the R-dimensional CLASCAL model with T and $T + 1$ latent classes to each of the generated samples $\bar{\mathbf{Y}}$; (d) compute the relevant likelihood statistic for comparing the T-class and $(T + 1)$-class solution; (e) reject the T-class solution at significance level α in favor of the $(T + 1)$-class solution if the value of the likelihood-ratio exceeds $S(1 - \alpha)$ of the values of the statistic obtained for the Monte Carlo samples $\bar{\mathbf{Y}}$. A minimal value of S for a significance level $\alpha = 0.05$ is 20. The power of the test increases as S becomes larger.

In practice, we have found that for a given number of latent classes, the lowest values of both AIC and BIC occur for the same spatial model (i.e., for a model with the same number of dimensions). This selected spatial model is then used in the Monte Carlo procedure described above to determine the appropriate number of classes. In the next section, we illustrate this model selection procedure with the latent class weighted Euclidean model, CLASCAL, on both real and artificial data.

Applications

Artificial Data Example

Two examples will be presented in this section. The first is an artificial example. A random configuration was seeded in two dimensions for nine stimuli. The following weight vectors were chosen for the two latent classes: $(0.66, 1.0)'$ and $(1.34, 1.0)'$. The model distances for each of two latent classes were calculated. Twenty subjects were then randomly assigned to one of the two classes and then Gaussian error with a standard deviation equal to 15 percent of the standard deviation of the model values was added to the model distances for each subject to constitute the data. AIC and BIC statistics were computed for the solutions obtained for two, three, and four dimensions

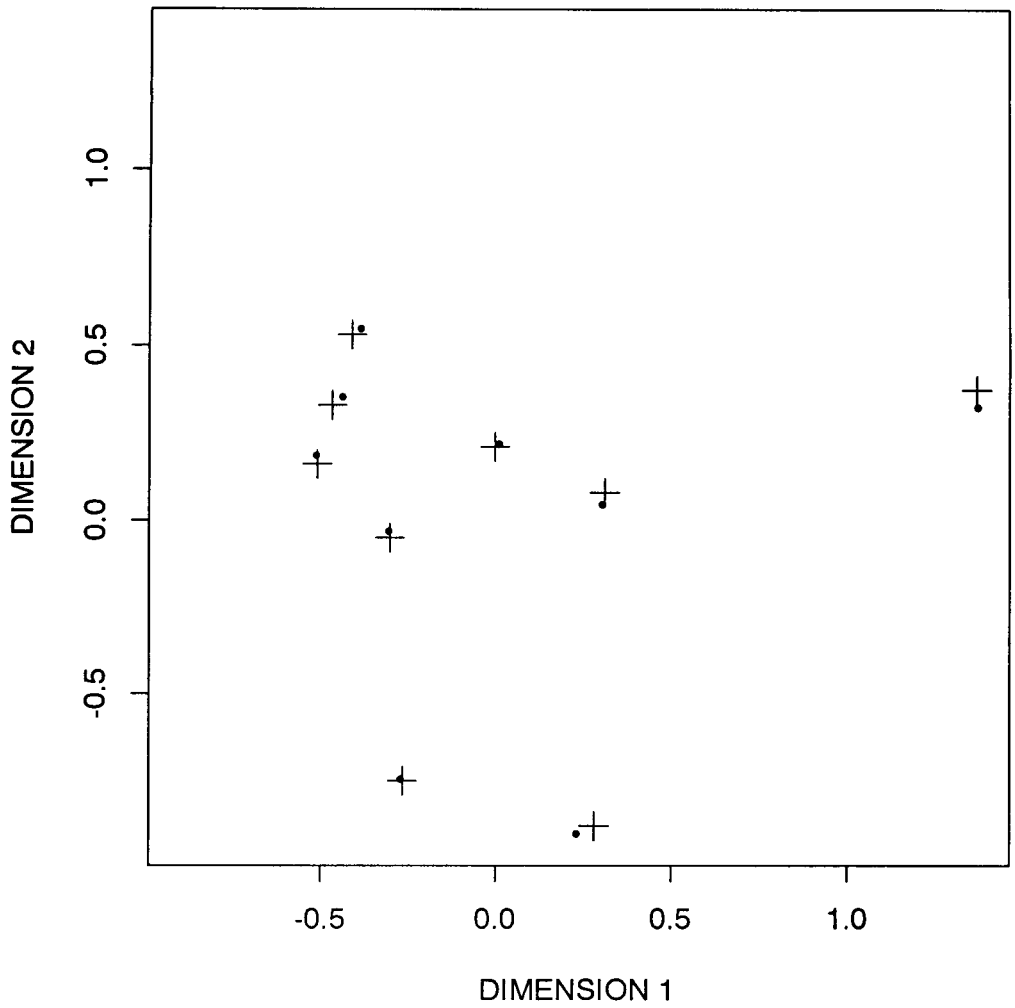

FIGURE 1.
True and recovered common space for the artificial data example. The crosses indicate the true locations, while the dots indicate the recovered locations.

for both the one-class and the two-class models. These goodness-of-fit statistics are summarized in Table 1. Both AIC and BIC select the appropriate number of dimensions (i.e., two) for both the one-class and the two-class case. On the basis of the Monte Carlo significance test with $S = 20$, we select two as the appropriate number of classes. As can be seen from Figure 1, the true configuration is recovered well. The posterior probabilities $h_{it}(\hat{\mathbf{X}}, \hat{\mathbf{W}}, \hat{\sigma}^2, \hat{\boldsymbol{\lambda}})$ listed in Table 2, classify each subject correctly. The true latent class weights are well recovered also. The final estimates are: $\mathbf{w}_1 = (0.652, 1.012)'$ and $\mathbf{w}_2 = (1.348, 0.988)'$.

The Monte Carlo significance tests for comparing one versus two classes and two versus three classes took respectively 212 and 417 CPU seconds on a DEC DS5820 computer to complete.

Real Data Example

The second set of data concerns nine stimuli each of which corresponds to the same short musical selection (a piece of Debussy) played in a particular simulated

TABLE 2

Posterior Probabilities for the Artificial Data Example

Subject	Class 1	Class 2
1	0.000	1.000
2	0.000	1.000
3	0.000	1.000
4	0.000	1.000
5	0.000	1.000
6	1.000	0.000
7	1.000	0.000
8	0.000	1.000
9	0.000	1.000
10	0.000	1.000
11	0.000	1.000
12	1.000	0.000
13	1.000	0.000
14	0.000	1.000
15	0.000	1.000
16	0.000	1.000
17	1.000	0.000
18	1.000	0.000
19	1.000	0.000
20	0.000	1.000

concert hall. The nine simulated halls differ on two physical dimensions: clarity-80 and reverberation time. A hall has the effect of transforming sound. To characterize a hall, the effect of the hall on an acoustic impulse is measured. Clarity-80 and reverberation time are typical measures of the way an impulse is transformed by a hall. More specifically, clarity-80 is the ratio of the response energy arising in the first 80 milliseconds to the total response energy, while the reverberation time is a measure of the decay time

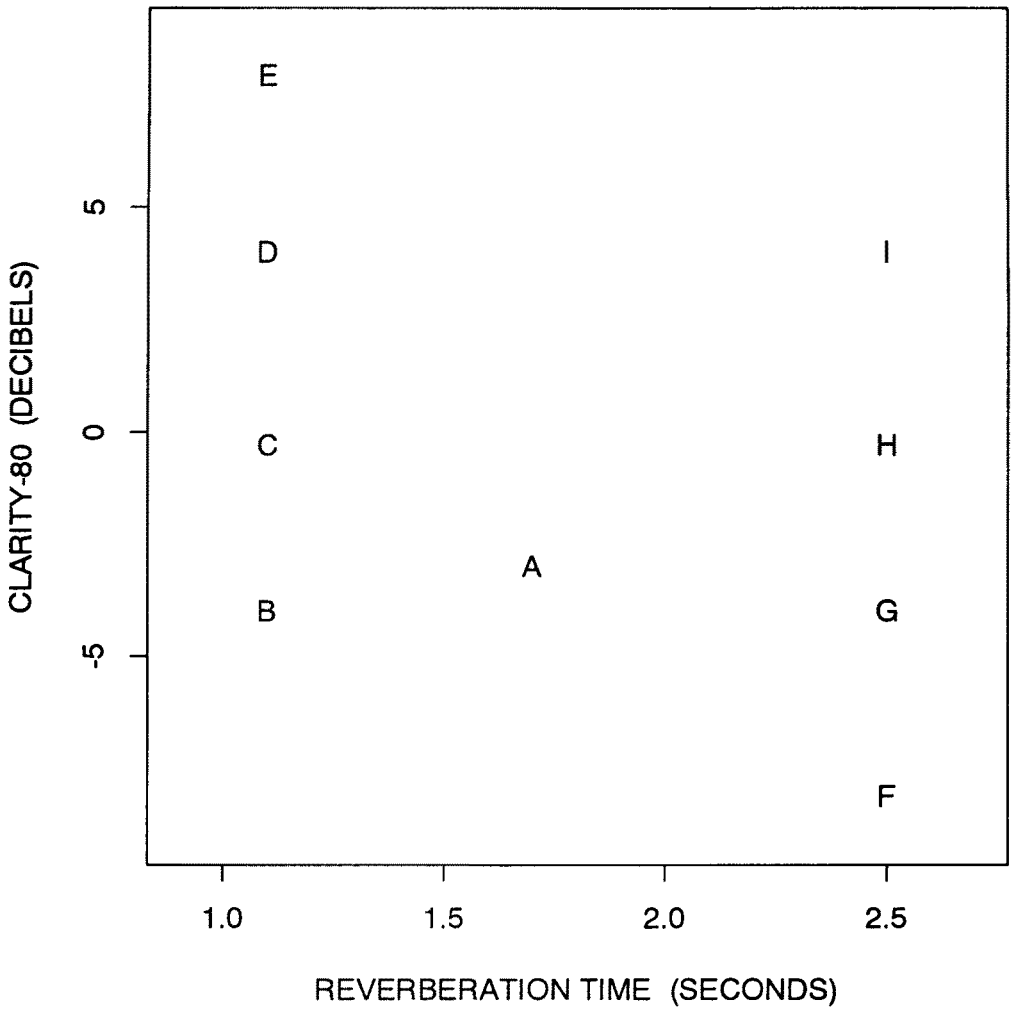

FIGURE 2.
The nine concert halls plotted in the plane defined by the two physical variables.

of the response energy. The nine concert halls are plotted in the plane defined by the two physical variables in Figure 2.

Dissimilarity data were collected from fifteen subjects for the nine concert halls. Both the AIC and BIC criteria select a three-dimensional representation for both the one-class and the two-class cases (see Table 3). The Monte Carlo likelihood-ratio test described in the previous section yields two as the appropriate number of classes. Incidentally, the INDSCAL solution for these data was jackknifed using the special jackknife for multidimensional scaling devised by de Leeuw and Meulman (1986), showing that the three-dimensional solution was most stable. Also, these data were analyzed by one of the authors using the extended INDSCAL model proposed by Carroll and Winsberg (1991) and a three-dimensional solution with no specificities and no spline transformation gave the best fit to the data. However, the CLASCAL solution appears to offer the clearest interpretation of the resulting dimensions and weights. The posterior probabilities listed in Table 4, reveal that the subjects are well-classified. The

TABLE 3

Goodness-of-fit Results for the Real Data Example

No. of Classes	No. of Dimensions	Model Degrees of Freedom	Log Likelihood	AIC	BIC
1	2	16	-1055.3	2142.6	2211.2
1	3	22	-1018.7	2081.3	2175.7
1	4	27	-1014.6	2083.3	2199.2
1	5	31	-1014.3	2090.7	2223.7
2	2	20	-1037.3	2114.5	2200.3
2	3	29	-988.2	2034.5	2158.9
2	4	38	-982.6	2041.2	2204.3

estimates of the latent class weights for two classes are $\mathbf{w}_1 = (1.047, 0.865, 0.431)'$ and $\mathbf{w}_2 = (0.952, 1.134, 1.569)'$.

The private spaces obtained for the two latent classes with the two-class three-dimensional CLASCAL model are displayed in Figure 3. One of the challenges of interpreting the solution is to see why two physical dimensions give rise to three psychological dimensions. Upon examining the upper panels of Figure 3 (for Dimension 1 versus Dimension 2) it is noted that the first two dimensions are the same for both classes. Indeed, the two classes are distinguished by the weights on the third dimension. Comparing Figure 2 with the upper panels of Figure 3 reveals that in the solution the two physical variables are rotated and that the reverberation time variable is compressed for smaller clarity-80. Moreover, stimuli F and G are seen as having greater reverberation time instead of lower clarity-80; that is, they are correctly perceived as having a smaller ratio of early energy to late energy, but this is incorrectly attributed to an increase in the late energy rather than a decrease in the early energy. The third (one might say extra) dimension is important for only one of the two latent classes (see lower panels of Figure 3). This third dimension separates F and to a smaller extent G from the rest of the stimuli. Interestingly enough when stimulus F was constructed, the clarity-80 was reduced naturally by decreasing both the direct sound and first reflections as compared to the total energy, but for this particular stimulus the ratio of direct sound to first reflection was considerably smaller than was the case for the other stimuli and this reduction in direct sound was apparently picked up by the second latent class.

Conclusion

The latent class approach to fitting the weighted Euclidean distance model elaborated in this paper yields parsimonious interpretable solutions for dissimilarity data. A Monte Carlo significance test coupled with AIC and BIC statistics provides a rationale for model selection.

TABLE 4

Posterior Probabilities for the Real Data Example

Subject	Class 1	Class 2
1	1.000	0.000
2	0.990	0.010
3	0.000	1.000
4	0.996	0.004
5	0.000	1.000
6	1.000	0.000
7	1.000	0.000
8	0.000	1.000
9	0.000	1.000
10	0.000	1.000
11	0.981	0.019
12	0.000	1.000
13	0.000	1.000
14	0.000	1.000
15	1.000	0.000

The CLASCAL model assumes a common variance for all stimulus pairs. Sometimes, dissimilarity data exhibit a typical mean-variance relationship (e.g., Ramsay, 1977). In such a case, the assumption of independent normal distributions with a common variance can be easily replaced by the assumption of independent lognormal distributions or independent normal distributions with standard deviations proportional to the means. In the paper we advocated the use of a Monte Carlo significance test for deciding on the appropriate number of classes. This procedure usually works quite well although with large data sets it can be very computationally intensive. It would be interesting to investigate to what extent recently proposed modified information criteria (see e.g., Bozdogan, 1987, 1992) can be used to choose the correct number of latent classes. Finally, the CLASCAL model can be extended to included specific dimensions as in the extended INDSCAL model (Winsberg & Carroll, 1989b) and nonmetric or quasi nonmetric (spline) transformations (Winsberg & Carroll, 1989a).

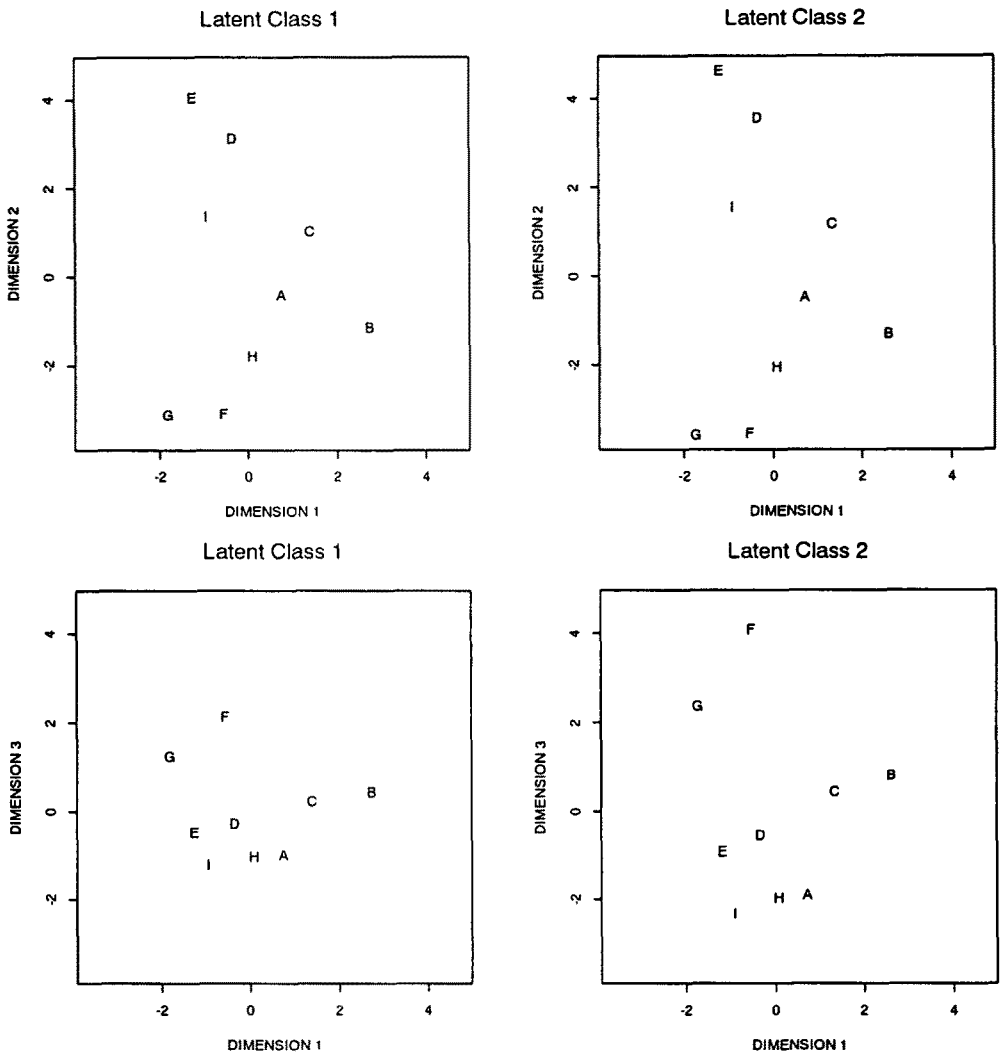

FIGURE 3.
The two private spaces obtained with the two-class three-dimensional CLASCAL model.

References

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, Series A, 144*, 419–461.

Akaike, H. (1977). On entropy maximization. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 27–41). Amsterdam: North-Holland.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370.

Bozdogan, H. (1992). *Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix.* Paper presented at the 16th Annual Meeting of the German Classification Society, Dortmund, Germany.

Böckenholt, U., & Böckenholt, I. (1990). Modeling individual differences in unfolding preference data: A restricted latent class approach. *Applied Psychological Measurement, 14*, 257–269.

Böckenholt, U., & Böckenholt, I. (1991). Constrained latent class analysis: Simultaneous classification and scaling of discrete choice data. *Psychometrika, 56*, 699–716.

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

Carroll, J. D., & Winsberg, S. (1991). *Fitting an extended INDSCAL model to three-way proximity data.* Unpublished manuscript, Rutgers University, Newark.

de Leeuw, J., & Meulman, J. (1986). A special jackknife for multidimensional scaling. *Journal of Classification, 3,* 97–112.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1–38.

DeSarbo, W. S., Howard, D. J., Jedidi, K. (1991). MULTICLUS: A new method for simultaneously performing multidimensional scaling and cluster analysis. *Psychometrika, 56,* 121–136.

De Soete, G. (1990). A latent class approach to modeling pairwise preferential choice data. In M. Schader & W. Gaul (Eds.), *Knowledge, data and computer-assisted decisions* (pp. 103–113). Berlin: Springer-Verlag.

De Soete, G. (1993). Using latent class analysis in categorization research. In I. Van Mechelen, J. Hampton, R. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 309–330). London: Academic Press.

De Soete, G., & DeSarbo, W. S. (1991). A latent class probit model for analyzing pick any/*N* data. *Journal of Classification, 8,* 45–63.

De Soete, G., & Heiser, W. J. (1992). *A latent class unfolding model for analyzing single stimulus preference ratings.* Unpublished manuscript, University of Ghent, Belgium.

De Soete, G., & Winsberg, S. (1993). A Thurstonian pairwise choice model with univariate and multivariate spline transformations. *Psychometrika, 58,* 233–256.

De Soete, G., & Winsberg, S. (in press). *A latent class vector model for analyzing preference ratings. Journal of Classification.*

Formann, A. K. (1989). Constrained latent class models: Some further applications. *British Journal of Mathematical and Statistical Psychology, 42,* 37–54.

Gower, J. C. (1966). Some distance properties of latent root and vector methods using multivariate analysis. *Biometrika, 53,* 325–338.

Hope, A. C. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society, Series B, 30,* 582–598.

Lazarsfeld, P. F., & Henry, R. W. (1968). *Latent structure analysis.* New York: Houghton Mifflin.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models.* New York: Marcel Dekker.

Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika, 42,* 241–266.

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society, Series A, 145,* 285–312.

Ramsay, J. O. (1991). *MULTISCALE manual (Extended version).* Montreal: McGill University.

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

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: Wiley.

Winsberg, S., & Carroll, J. D. (1989a). A quasi-nonmetric method for multidimensional scaling via an extended Euclidean model. *Psychometrika, 54,* 217–229.

Winsberg, S., & Carroll, J. D. (1989b). A quasi-nonmetric method for multidimensional scaling of multiway data via a restricted case of an extended INDSCAL model. In R. Coppi & S. Bolasco (Eds.), *Multiway data analysis* (pp. 405–414). Amsterdam: North-Holland.

Winsberg, S., & Ramsay, J. O. (1983). Monotone spline transformations for dimension reduction. *Psychometrika, 48,* 575–595.

Manuscript received 2/20/92
Final version received 5/23/92