PSYCHOMETRIKA—VOL. 66, NO. 1, 25–43
MARCH 2001

RANDOM EFFECTS DIAGONAL METRIC MULTIDIMENSIONAL SCALING MODELS

DOUGLAS B. CLARKSON

DATA ANALYSIS PRODUCTS DIVISION, MATHSOFT, INC.

RICHARD GONZALEZ

UNIVERSITY OF MICHIGAN

By assuming a distribution for the subject weights in a diagonal metric (INDSCAL) multidimensional scaling model, the subject weights become random effects. Including random effects in multidimensional scaling models offers several advantages over traditional diagonal metric models such as those fitted by the INDSCAL, ALSCAL, and other multidimensional scaling programs. Unlike traditional models, the number of parameters does not increase with the number of subjects, and, because the distribution of the subject weights is modeled, the construction of linear models of the subject weights and the testing of those models is immediate. Here we define a random effects diagonal metric multidimensional scaling model, give computational algorithms, describe our experiences with these algorithms, and provide an example illustrating the use of the model and algorithms.

Key words: multidimensional scaling, random coefficients.

1. Random Effects Diagonal Metric Multidimensional Scaling Models

Multidimensional scaling analysis estimates the coordinates of a set of n objects in a τ dimensional space given only a measure (observed with error) of the distances between the objects. The dimensionality of the space, τ, is specified by the model. In the models considered here, the distance measurements, called *dissimilarities*, are obtained from a sample of m subjects. The observed data are denoted by y_{lj}, where $l = 1, \ldots, m$ indexes subjects, and $j = 1, \ldots, N_l$ indexes distances between pairs of stimuli. The y_{lj} are usually not distances in the strict sense, but are thought to satisfy the various properties required of distance measures. Supposing that there are n stimuli, a maximum of $N = n(n+1)/2$ dissimilarities on each subject are possible, but only a subset of these need to be measured for estimation, i.e., $N_l \leq N$.

Here we are concerned with the diagonal metric (also called the individual differences or INDSCAL) model first proposed by Caroll and Chang (1970). Let x_{ik} denote the coordinates for $i = 1, \ldots, n$ stimuli in a $k = 1, \ldots, \tau$ dimensional space, and call the matrix $\mathbf{X} = (x_{ik})$ the *configuration matrix*. The diagonal metric model assumes that the expected distance between stimulus i and stimulus j for individual l is given as

$$d_{lij} = \left(\sum_{k=1}^{\tau} w_{lk} (x_{ik} - x_{jk})^2 \right)^{1/2}$$

In this model the *subject weights*, w_{lk}, account for variability between individuals—each individual gives different weight to each of the τ dimensions.

We would like to thank J. Douglas Carroll for early consultation of this research, and Robert I. Jennrich for commenting on an earlier draft of this paper and for help on the computational algorithms. James O. Ramsay and Forrest W. Young were instrumental in providing the example data. This work was supported in part by National Institute of Mental Health grant 1 R43 MH57559-01. We would also like to thank the anonymous referees for comments that helped to clarify our work.

Requests for reprints should be sent to Douglas Clarkson, Data Analysis Products Division, MathSoft, Inc., 1700 Westlake Ave. N., Suite 500, Seattle, WA, 98109-3044. E-Mail: clarkson@statsci.com

0033-3123/2001-1/1998-0637-A $00.75/0

In many analyses, a random sample of subjects is obtained. In this case, *conditional* upon the observed subjects, a traditional analysis computes estimates using least squares or maximum likelihood methods. In contrast, in the random effects models we propose, the subject weights are considered to be a sample from a distribution of subject weights, and unconditional maximum likelihood estimates are computed. In other words, we treat the subject weights as random coefficients. The inclusion of these random coefficients in the diagonal metric model yields the same benefits as are obtained by incorporating random coefficients into linear (or nonlinear) regression models (see, e.g., Davidian & Giltinan, 1995), including:

- a single model that combines the distribution of the dissimilarities with the distribution of the subject weights,
- model-based estimates of the variance components—the parameters associated with the variances of the subject weights,
- the ability to make inferences regarding the subject weights in the sampled population,
- avoidance of overfitting of individual subject weights by smoothing of the estimates,
- the ability to "borrow strength" from all individuals and thus obtain better estimates for an individual even when only a few data points are observed on the individual, and
- the ability to linearly model and make inferences about the subject weights based upon predictors collected on each individual.

In the next section we describe the model. We then describe three algorithms that compute maximum likelihood estimates using one of two models for the distribution of the random effects. These two models are: (a) a "parametric" model in which the random effects are assumed to be multivariate normal, and (b) a "nonparametric" model in which the distribution of the random effects is left completely unspecified. The final three sections illustrate how these algorithms perform.

2. A Random Coefficient Diagonal Metric Model

To avoid subscripting problems, in the following we let $d_{lj} = \mathrm{E}(y_{lj})$ be the expected value of dissimilarity j measured on individual l. Then for functions $\nu_{l1}(j)$ and $\nu_{l2}(j)$ taking the index $j = 1, \ldots, N_l$ into the set of integers from 1 to n, the dissimilarity is a measure of the distance between objects $\nu_{l1}(j)$ and $\nu_{l2}(j)$ for the l-th subject. This allows us to use double subscripts d_{lj} (and y_{lj}, etc.) in place of more unwieldy triple subscripts such as $d_{l_{ik}} = d_{l,\nu_{l1}(j),\nu_{l2}(j)}$.

Conditional upon the subject weights for individual l (denoted by a vector, $\mathbf{w}_l = (w_{lk})$) and the configuration matrix $\mathbf{X}$, we utilize the same assumptions as in the conditional models and assume independent and identically distributed random measurement errors e_{lj} such that

$$y_{lj} = d_{lj} + e_{lj},$$

where $\mathrm{E}(e_{lj}|\mathbf{w}_l, \mathbf{X}) = 0$ and $\mathrm{Var}(e_{lj}|\mathbf{w}_l, \mathbf{X}) = \sigma^2$. Also as in a conditional analysis, to make the estimates identifiable, translation of the columns of $\mathbf{X}$ is fixed by assuming that the mean of each column of $\mathbf{X}$ is zero, and the scale of the columns of $\mathbf{X}$ and $\mathbf{W}$ is fixed by the restriction that the ℓ_2 norm of each column in $\mathbf{X}$ is constant, that is, $(\sum_{i=1}^{n} x_{ij}^2)^{1/2} = n$ for $j = 1, \ldots, \tau$. In our algorithms we sometimes use simpler (to implement) restrictions in the initial estimation, but the final estimates are always standardized as above.

In some diagonal metric model programs (e.g., ALSCAL, see Takane, Young, & de Leeuw, 1977) the scale restrictions are placed on the subject weights rather than on the configuration matrix. We specifically avoid restricting the subject weight estimates—later we will assume a distribution for the (log of the) subject weights, and restrictions on the subject weight estimates would severely limit the possible distributions we could consider (in particular, this would eliminate the multivariate normal distribution).

Let $\mathbf{y}_l = (y_{lj})$ denote the column vector of length N_l of dissimilarities on subject l, let $\mathbf{d}_l = (d_{lj})$ denote the corresponding vector of expected dissimilarities, and let $\mathbf{e}_l = (e_{lj})$ denote the vector of measurement errors for subject l. Notice that this notation allows a different number of observations on each subject.

Re-expressing our assumptions in vector notation, we have $\mathrm{E}(\mathbf{e}_l|\mathbf{X}, \mathbf{w}_l) = 0$, and $\mathrm{Cov}(\mathbf{e}_l|\mathbf{X}, \mathbf{w}_l) = \sigma^2\mathbf{I}$. Although, for simplicity, we have chosen a simple covariance structure $\sigma^2\mathbf{I}$ for the measurement errors, in principle this can be generalized to arbitrary covariance structures $\boldsymbol{\Sigma}_e(\boldsymbol{\eta})$ for parameter vector $\boldsymbol{\eta}$. For example, Ramsay (1982) considers more general covariance structures in which the measurement error depends upon the stimuli, and covariance structures that permit correlations between observations on a single subject (e.g., repeated measures) may be desirable.

Models that condition upon the observed subjects yield diagonal metric models such as the INDSCAL model. Extending the conditional model so that inferences regarding the population of subjects may be made, consider the vector of subject weights, $\mathbf{w}_l$. The subject weights are, in fact, random coefficients—each randomly sampled subject brings a new random vector of subject weights from some unknown distribution. Modeling this distribution, we assume a linear model for $\log \mathbf{w}_l$ such that

$$\log \mathbf{w}_l = \mathbf{U}_l\boldsymbol{\beta} + \mathbf{V}_l\boldsymbol{\gamma}_l,$$

where $\mathbf{U}_l$ and $\mathbf{V}_l$ are known design matrices, $\boldsymbol{\beta}$ is a vector of unknown parameters, and $\boldsymbol{\gamma}_l$ is a random vector with mean zero and covariance structure $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ that depends upon a vector of unknown parameters $\boldsymbol{\theta}$. Then $\mathrm{E}(\log \mathbf{w}_l) = \mathbf{U}_l\boldsymbol{\beta}$, while the covariance matrix for $\log \mathbf{w}_l$ is given by $\mathbf{D}(\boldsymbol{\theta}) = \mathbf{V}_l\boldsymbol{\Sigma}(\boldsymbol{\theta})\mathbf{V}_l^T$. Initially we assume a multivariate normal distribution for $\boldsymbol{\gamma}$, but later this assumption is extended to include other distributions. The log of the subject weights are modeled (rather than the subject weights) to avoid negative subject weights in the estimation. It is also possible to model the subject weights directly using $\mathbf{w}_l = \mathbf{U}_l\boldsymbol{\beta} + \mathbf{V}_l\boldsymbol{\gamma}_l$, but in this case it may be necessary to eliminate the possibility of negative subject weights.

As an example of where linear models predicting the subject weights might be useful, consider a wine tasting example discussed by Clarkson and Gentle (1986). In this example subjects were asked to give dissimilarities for nine types of Cabernet Sauvignon wine. On the day of measurement one individual had a cold, and in the subsequent diagonal metric analysis, this individual gave zero weight to one of the derived dimensions. Utilizing "cold" as a predictor, a linear model for the subject weight is possible, $\log \omega_{l1} = \beta_0 + u_l\beta_2 + \gamma_{l1}$, where u_l is zero or one depending upon whether or not the l-th subject has a cold, and where γ_{l1} is a random effect giving the derivation of the subject weight on the l-th subject from the linear model. Inclusion of the predictor permits likelihood ratio tests for its statistical importance, and also allows better prediction of the subject weights when considering new subjects. The impact of the presence of a cold on the subject weight for one dimension also suggests that the dimension in question may be related to a sense of smell—using linear model predictors for the subject weights may facilitate the interpretation of the derived dimensions.

Because the log of the subject weights are modeled, the linear model in $\log \mathbf{w}_l$ is interpreted as decomposing $\mathbf{w}_l$ into its multiplicative parts. In this case, the predictors have a multiplicative effect on the subjects' interpretation of any dimension. Thus, supposing that $\mathrm{E}\log w_{l1} = 0.36 - 0.47u_{l1}$, then the expected subject weights are given by $\mathrm{E}w_{l1} = \exp(0.36)\exp(-0.47u_{l1}) = 1.433\exp(-0.47u_{l1})$. This means that we expect a subjects with $u_{l1} = 1$ to experience distances along dimension 1 that are $\exp(-0.47) = 0.625$ the length of subject with $u_{l1} = 0$.

It is also possible to linearly model the elements in the coordinate matrix x_{ik} so that, for example, we might have $x_{ik} = \mathbf{Z}_{ik}\boldsymbol{\xi}$ for fixed parameters $\boldsymbol{\xi}$ and known predictor matrix $\mathbf{Z}_{ik}$. However, in this paper we will assume that the configuration matrix is constant over all individuals. Linear models on the configuration matrix parameters may be more relevant in multidimensional unfolding, where the primary source of variation between individuals is in the subject ideal points, which may also be modeled as random effects.

Linear models for predicting multidimensional scaling parameters such as $\log \mathbf{w}_i$ and x_{ik} are not new—Carroll, Pruzansky, & Kruskal (1980; see also Carroll, DeSoete, & Pruzansky 1989); de Leeuw and Heiser (1980); Bloxom (1978); Bentler and Weeks (1978) and others have proposed multidimensional scaling models with linearly constrained parameters. They do not include random effects in these models. Notice that linearly constrained parameters are qualitatively different from the linear models in our random effects models. Rather than restricting the subject weights, the linear models we propose restrict the expected values of the (log) subject weights (as in regression analysis). The subject weights themselves are free to vary about these expected values as specified by the distribution of the random effects. This use of predictors allows us to reduce the variances of the (log) subject weights (the variance components) given the predictor and allows us to better predict subject weights for new subjects. In a constrained model, if the subject weights do not fall in the constrained subspace, the model is not valid.

In multivariate normal models the covariance structure of the random effects, $\mathbf{\Sigma}(\boldsymbol{\theta})$, must be specified. In our examples, because of its simplicity and ease of implementation, we assume a diagonal covariance structure in which $\mathbf{\Sigma}(\boldsymbol{\theta})$ is a diagonal matrix with diagonal elements σ^2_{jj}. In principle other covariance structures are possible. For example, one might use $\mathbf{\Sigma}(\boldsymbol{\theta}) = \sigma^2_\gamma \mathbf{I}$ for scalar parameter σ^2_γ, or $\mathbf{\Sigma}(\boldsymbol{\theta}) = \mathbf{L}\mathbf{L}^T$ for arbitrary lower triangular matrix $\mathbf{L}$.

In a conditional model each subject weight is a parameter, and thus the number of parameters must increase as the number of subjects increases. An important improvement of the (parametric) random effects model is that it eliminates an undesirable characteristic: the number of parameters is fixed and does not increase with increasing sample size. Because the number of parameters is much smaller, we would expect fewer identifiability problems, though showing this is beyond the scope of this paper.

2.1. Summary

Summarizing, we model the intra-individual variation as

$$\begin{aligned} \mathbf{y}_l &= \mathbf{d}_l + \mathbf{e}_l \\ \mathrm{E}(\mathbf{e}_l|\mathbf{X}, \mathbf{w}_l) &= 0, \quad \text{and} \\ \mathrm{Cov}(\mathbf{e}_l|\mathbf{X}, \mathbf{w}_l) &= \sigma^2 \mathbf{I} \end{aligned}$$

just as in the conditional case, while the inter-individual variation has model

$$\log \mathbf{w}_l = \mathbf{U}_l \boldsymbol{\beta} + \mathbf{V}_l \boldsymbol{\gamma}_l,$$

where

$$\boldsymbol{\gamma}_l \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}(\boldsymbol{\theta})).$$

This model can be generalized by specifying a more general covariance structure for

$$\mathrm{Cov}(\mathbf{e}_l|\mathbf{X}, \mathbf{w}_l),$$

by specifying a different structure for $\mathbf{\Sigma}(\boldsymbol{\theta})$, by changing the distribution of $\boldsymbol{\gamma}_l$, or by changing the model for $\log \mathbf{w}_l = \mathbf{U}_l \boldsymbol{\beta} + \mathbf{V}_l \boldsymbol{\gamma}_l$ to include different predictors in $\mathbf{U}_l$, or alternative parameterizations for the random effects $\boldsymbol{\gamma}_l$.

By modeling the subject weights as random variables, we more accurately represent the data because: (a) the subject weights are, in fact, random effects if a random sample of subjects is obtained, (b) the use of random subject weights implies that the observations on each subject are correlated, a more reasonable assumption that the assumption of independence most often used in conditional models, (c) random effects models allow direct, maximum likelihood, estimation of the variances of the random effects, and (d) we obtain a better estimate of the expected subject

weight on each subject (given the model) because we can "borrow strength" from all subjects when estimating each subject weight, and because we can predict the subject weights in terms of a linear model of predictors. As in the linear mixed effects models, the random effects model estimates we propose shrink the subject specific estimates toward the expected mean subject weight. For linear mixed effects models, these estimates are the best linear unbiased estimates (BLUP, see, e.g., Davidian & Giltinan 1995).

2.2. *Interval Data*

Generalization of the diagonal metric model to dissimilarities that are "interval" (rather than "ratio") variables is straightforward. In this case a constant vector $\boldsymbol{\alpha}_l$ must be added to the dissimilarity so that $\mathbf{y}_l + \boldsymbol{\alpha}_l$ (rather than $\mathbf{y}_l$) is the distance measurement. We assume that the $\boldsymbol{\alpha}_l$ varies from subject to subject, and, because we are considering random effect models, we further assume that $\boldsymbol{\alpha}_l$ is composed of both a random and a fixed component, that is, that $\boldsymbol{\alpha}_l = \boldsymbol{\beta}^{\alpha} + \boldsymbol{\gamma}_l^{\alpha}$, where $\boldsymbol{\beta}^{\alpha}$ is fixed and where the population mean of the random effects $\boldsymbol{\gamma}_l^{\alpha}$ is zero.

Notice that our models have assumed that the random errors are associated with the distances (i.e., that $\mathbf{y}_l + \boldsymbol{\alpha}_l = \mathbf{d}_l + \mathbf{e}_l$). It is also common to assume that the random errors are added to the logarithm or to the square of the distances, that is, to assume that $\log(\mathbf{y}_l + \boldsymbol{\alpha}_l) = \log(\mathbf{d}_l) + \mathbf{e}_l$ or that $(\mathbf{y}_l + \boldsymbol{\alpha}_l)^2 = \mathbf{d}_l^2 + \mathbf{e}_l$. It is relatively easy to generalize the present model to other parameterizations. The transformation that is typically preferred is one that yields additive measurement errors.

3. Maximum Likelihood Estimation

The parameters to be estimated are given by $\boldsymbol{\Psi} = (\mathbf{X}, \boldsymbol{\beta}.\boldsymbol{\theta}, \sigma^2)$, where $\mathbf{X}$ is the configuration matrix, $\boldsymbol{\beta}$ is a vector of parameters that model the mean of the log of the subject weights such that $\mathrm{E}(\log \mathbf{w}_l) = \mathbf{U}_l\boldsymbol{\beta}$, $\boldsymbol{\theta}$ is a vector of parameters associated with the distribution of the random effects, $g(\cdot|\boldsymbol{\theta})$, and σ^2 is the variance of the measurement error. Depending on the distribution assumed for the random effects, several estimation algorithms are available. In each algorithm, initial estimates may be obtained as appropriately standardized conditional diagonal metric model estimates.

We begin by giving an expression for the likelihood. Let

$$f(\mathbf{y}_l|\mathbf{X}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma}_l) = \prod_{j=1}^{N_l} f(y_{lj}|\mathbf{X}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma}_l)$$

denote the conditional distribution for the vector of dissimilarities on individual l given the random effects $\boldsymbol{\gamma}_l$ and the parameters $(\mathbf{X}, \sigma^2, \boldsymbol{\beta})$, and let $g(\boldsymbol{\gamma}|\boldsymbol{\theta})$ denote the marginal distribution of the random effects. Then the marginal distribution of the dissimilarities is obtained as the integral of the joint distribution of $\mathbf{y}$ and $\boldsymbol{\gamma}$ with respect to $\boldsymbol{\gamma}$,

$$f(\mathbf{y}_l|\boldsymbol{\Psi}) = \int f(\mathbf{y}_l|\mathbf{X}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma})\, g(\boldsymbol{\gamma}|\boldsymbol{\theta})\, d\boldsymbol{\gamma} = \int f(\mathbf{y}_l, \boldsymbol{\gamma}|\boldsymbol{\Psi})\, d\boldsymbol{\gamma},$$

where $\boldsymbol{\gamma}$ is vector valued. The log-likelihood contribution for the l-th subject is given by $\ell(\boldsymbol{\gamma}_l|\boldsymbol{\Psi}) = \log f(\mathbf{y}_l|\boldsymbol{\Psi})$, and the log-likelihood is $\ell(\mathbf{y}|\boldsymbol{\Psi}) = \sum_{l=1}^{m} \ell(\mathbf{y}_l|\boldsymbol{\Psi})$.

The source of the difficulty in computing maximum likelihood estimates is in integrating the joint distribution to obtain $\ell(\mathbf{y}_l/\boldsymbol{\Psi})$. Analytic integration is not generally possible and numerical quadrature is not efficient because a multidimensional integral (often of dimension 5 or more) is involved. In the following we discuss three maximum likelihood algorithms, beginning with an

algorithm that utilizes a linear approximation to the log likelihood, then move to a Monte Carlo EM algorithm, and finally discuss a nonparametric algorithm.

3.1. The Linearization Algorithm

Assume that both $f(\mathbf{y}_l|\mathbf{X}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma})$ and $g(\boldsymbol{\gamma}|\boldsymbol{\theta})$ are multivariate normal. Then our model is a nonlinear mixed effects model, and a popular estimation algorithm uses a linear approximation to the likelihood that was first suggested by Sheiner, Rosenberg, and Melmon (1972) and later refined by Lindstrom and Bates (1990). Here we use algorithms developed by Pinheiro and Bates (1995) and available as the *nlme* function in S-PLUS. Expressing the distances $\mathbf{d}_l = h(\boldsymbol{\gamma}_l)$ as a function of the random effects $\boldsymbol{\gamma}_l$, and expanding around the mode of the posterior distribution of $\boldsymbol{\gamma}_l$ (denoted by $\boldsymbol{\gamma}_l^*$), we have

$$\begin{aligned}
\mathbf{y}_l &= \mathbf{d}_l + \mathbf{e}_l = h(\boldsymbol{\gamma}_l) + \mathbf{e}_l \\
&\approx h(\boldsymbol{\gamma}_l^*) + \left.\frac{\partial h(\boldsymbol{\gamma}_l)}{\partial \boldsymbol{\gamma}_l}\right|_{\boldsymbol{\gamma}_l=\boldsymbol{\gamma}_l^*} (\boldsymbol{\gamma}_l - \boldsymbol{\gamma}_l^*) + \mathbf{e}_l \\
&= h(\boldsymbol{\gamma}_l^*) + \mathbf{Z}_l(\boldsymbol{\gamma}_l - \boldsymbol{\gamma}_l^*) + \mathbf{e}_l \\
&= h(\boldsymbol{\gamma}_l^*) - \mathbf{Z}_l(\boldsymbol{\gamma}_l^* - \boldsymbol{\mu}_{\boldsymbol{\gamma}}) + \mathbf{Z}_l(\boldsymbol{\gamma}_l - \boldsymbol{\mu}_{\boldsymbol{\gamma}}) + \mathbf{e}_l,
\end{aligned}$$

where $\mathbf{Z}_l$ is the matrix of partial derivatives of $h(\boldsymbol{\gamma}_l)$ with respect to $\boldsymbol{\gamma}_l$ evaluated at $\boldsymbol{\gamma}_l^*$, $\mathbf{Z}_l = \partial h(\boldsymbol{\gamma}_l)/\partial \boldsymbol{\gamma}_l|_{\boldsymbol{\gamma}_l=\boldsymbol{\gamma}_l^*}$. Because, by assumption, $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \mathbf{0}$, the final equation above is identical to a linear components of variance model in which

$$\mathrm{E}(\mathbf{y}_l) \approx h(\boldsymbol{\gamma}_l^*) - \mathbf{Z}_l\boldsymbol{\gamma}_l^*,$$

and

$$\boldsymbol{\Gamma}_l = \mathrm{Cov}(\mathbf{y}_l) \approx \sigma^2\mathbf{I} + \mathbf{Z}_l\boldsymbol{\Sigma}(\boldsymbol{\theta})\mathbf{Z}_l^T.$$

Given initial estimates, Davidian and Giltinan (1995, p. 167) describe this algorithm as a generalized least squares procedure with the following two steps:

1. Estimate $(\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$ as the minimizing values of the approximate marginal log-likelihood for the random effects. For our models this log-likelihood is

$$\ell(\boldsymbol{\Psi}) = \sum_{l=1}^{m} \log|\boldsymbol{\Gamma}_l| + (\mathbf{y}_l - h(\boldsymbol{\gamma}_l^*) - \mathbf{Z}_l\boldsymbol{\gamma}_l^*)^T\boldsymbol{\Gamma}_l^{-1}(\mathbf{y}_l - h(\boldsymbol{\gamma}_l^*) - \mathbf{Z}_l\boldsymbol{\gamma}_l^*),$$

   where the $\boldsymbol{\gamma}_l^*$'s are fixed. The parameters $\boldsymbol{\beta}$ and $\mathbf{X}$ as well as the random effects $\boldsymbol{\gamma}_l^*$ are fixed at their current values when computing $\boldsymbol{\Gamma}_l$. This is a components of variance problem.
2. Obtain new estimates $\hat{\mathbf{X}}$, $\hat{\boldsymbol{\beta}}$, and the random effects $\boldsymbol{\gamma}_l^*$ by minimizing the generalized least squares criterion

$$\sum_{l=1}^{m} \log|\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})| + \boldsymbol{\gamma}_l^T\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\,\boldsymbol{\gamma}_l + N_l\log\hat{\sigma}^2 + \frac{|\mathbf{y}_l - h(\boldsymbol{\gamma}_l)|^2}{\hat{\sigma}^2}.$$

   Here the elements of $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})$ and $\hat{\sigma}^2$ are fixed. If a check for convergence indicates that convergence has been reached, stop. Otherwise, go to Step 1.

The configuration matrix $\hat{\mathbf{X}}$ estimated in steps one and two must satisfy location and scale restrictions to make the problem identifiable. When iterating, we fix the location by fixing the last

row in $\hat{\mathbf{X}}$, $\hat{x}_{nk}, k = 1, \ldots, \tau$. To fix the scale, we fix one additional element in each column of $\hat{\mathbf{X}}$. This second element is chosen such that the absolute difference between the two fixed elements in the column is maximum. At convergence, we standardize as described earlier.

This algorithm uses sequential optimization and separately optimizes over two sets of "parameters". Such algorithms can exhibit linear convergence properties unless the asymptotic correlation between the two sets of parameters is nearly zero. In our limited experience for multidimensional scaling problems, we have seen examples in which convergence was quite slow requiring fifty iterations or more, but on most problems the algorithm converged in ten or fewer iterations, and used (comparatively) little computer time.

A recent paper by Demidenko (1997) shows that the linearization algorithm estimates are not consistent as the number of subjects increases without bound when the number of observations per subject remains small and fixed. Lack of consistency makes these estimates less appealing. Even so, the algorithm produces reasonable estimates quickly, and thus will often be preferred. Many other approximate likelihood estimates are possible, including estimates using alternative linear expansions and estimates based upon the Laplace approximation (Vonesh & Chinchilli, 1997).

3.2. The MCEM Algorithm

Recently Geyer (1996; Geyer & Tompson, 1992; McCulloch, 1997; McLachlan & Krishnan, 1997; Tanner, 1996; Wei & Tanner, 1990) has provided a maximum likelihood algorithm based upon the Monte Carlo integration of the likelihood using a Monte Carlo EM algorithm. Though computationally intense, this Markov Chain Monte Carlo method generally converged in a reasonable length of computer time. In Monte Carlo integration the multidimensional integrals that are required in the marginal density for y and its derivatives are approximated as the average function value evaluated over a randomly generated sample of the integration variable. Here the Metropolis-Hastings algorithm was used in generating the Monte Carlo sample on each subject.

Rather than maximizing the log-likelihood directly, a ratio of probability densities is used. Let $f(\mathbf{y}|\boldsymbol{\Psi})$ denote the density for the l-th subject and let $\tilde{\boldsymbol{\Psi}}$ be a known vector of parameters chosen to be close to the maximum likelihood estimates $\hat{\boldsymbol{\Psi}}$. Then the derivation of the function to be optimized is given as follows:

$$\begin{aligned} f(\mathbf{y}_l|\boldsymbol{\Psi}) &= \int f(\mathbf{y}_l, \boldsymbol{\gamma}|\boldsymbol{\Psi})\, d\boldsymbol{\gamma} \\ &= f(\mathbf{y}_l|\tilde{\boldsymbol{\Psi}}) \int \frac{f(\mathbf{y}_l, \boldsymbol{\gamma}|\boldsymbol{\Psi})}{f(\mathbf{y}_l|\tilde{\boldsymbol{\Psi}}) f(\boldsymbol{\gamma}|\mathbf{y}_l, \tilde{\boldsymbol{\Psi}})} f(\boldsymbol{\gamma}|\mathbf{y}_l, \tilde{\boldsymbol{\Psi}})\, d\boldsymbol{\gamma} \\ &= f(\mathbf{y}_l|\tilde{\boldsymbol{\Psi}}) \int \frac{f(\mathbf{y}_l, \boldsymbol{\gamma}|\boldsymbol{\Psi})}{f(\mathbf{y}_l, \boldsymbol{\gamma}|\tilde{\boldsymbol{\Psi}})} f(\boldsymbol{\gamma}|\mathbf{y}_l, \tilde{\boldsymbol{\Psi}})\, d\boldsymbol{\gamma}, \end{aligned}$$

where $f(\mathbf{y}_l|\tilde{\boldsymbol{\Psi}})$ is an unknown constant, $\boldsymbol{\Psi}$ is the vector of all parameters, $f(\mathbf{y}_l, \boldsymbol{\gamma}|\boldsymbol{\Psi})$ is the joint distribution of the dissimilarities $\mathbf{y}_l$ and the random effects $\boldsymbol{\gamma}$ given the parameters, and $f(\boldsymbol{\gamma}|\mathbf{y}_l, \tilde{\boldsymbol{\Psi}})$ is the posterior distribution of the random effects $\boldsymbol{\gamma}$ given the parameters $\tilde{\boldsymbol{\Psi}}$ and the dissimilarities $\mathbf{y}_l$. Because closed form solutions for the integral are not usually known, the marginal density $f(\mathbf{y}_l|\tilde{\boldsymbol{\Psi}})$ cannot usually be computed. Rather, we use Monte Carlo integration in the E-step to evaluate the logarithm of the density ratio, $Q_l(\boldsymbol{\Psi}, \tilde{\boldsymbol{\Psi}}) = \log\{f(\mathbf{y}_l\, \boldsymbol{\Psi})/f(\mathbf{y}_l|\tilde{\boldsymbol{\Psi}})\}$, where the Monte Carlo sample, $\hat{\boldsymbol{\gamma}}_j, j = 1, \ldots, M$, is obtained from the distribution $f(\boldsymbol{\gamma}|\mathbf{y}_l, \tilde{\boldsymbol{\Psi}})$ using a Metropolis–Hastings algorithm. Then

$$Q_l(\boldsymbol{\Psi}, \tilde{\boldsymbol{\Psi}}) = \log\left\{\frac{f(\mathbf{y}_l|\boldsymbol{\Psi})}{f(\mathbf{y}_l|\tilde{\boldsymbol{\Psi}})}\right\} \approx \left\{\frac{1}{M}\sum_{j=1}^{M} \frac{f(\mathbf{y}, \boldsymbol{\gamma}_{lj}|\boldsymbol{\Psi})}{f(\mathbf{y}_l, \boldsymbol{\gamma}_{lj}|\tilde{\boldsymbol{\Psi}})}\right\}.$$

The approximation can be made as accurate as desired by increasing M (though very large M may be required). Differentiating under the integral sign, the gradient and Hessian for $\log f(\mathbf{y}_l|\boldsymbol{\Psi})$ can also be closely approximated. Letting $\dot{\ell}(\mathbf{y}_l|\boldsymbol{\Psi})$ denote the first partial derivative of $\log f(\mathbf{y}_l|\boldsymbol{\Psi})$ with respect to $\boldsymbol{\Psi}$, and letting $\ddot{\ell}(\mathbf{y}_l + \boldsymbol{\Psi})$ denote the second partial derivatives (and letting $\dot{\ell}(\mathbf{y}_l, \boldsymbol{\gamma}|\boldsymbol{\Psi})$ and $\ddot{\ell}(\mathbf{y}_l, \boldsymbol{\gamma}|\boldsymbol{\Psi})$ denote similar partial derivatives for the joint density), then the gradient is given by

$$\mathbf{g}_l = \dot{\ell}(\mathbf{y}_l|\boldsymbol{\Psi}) = \frac{\int \frac{f(\mathbf{y}_l,\boldsymbol{\gamma}|\boldsymbol{\Psi})}{f(\mathbf{y}_l,\boldsymbol{\gamma}|\tilde{\boldsymbol{\Psi}})} \dot{\ell}(\mathbf{y}_l, \boldsymbol{\gamma}|\boldsymbol{\Psi}) f(\boldsymbol{\gamma}|\mathbf{y}, \tilde{\boldsymbol{\Psi}})\, d\boldsymbol{\gamma}}{\int \frac{f(\mathbf{y}_l,\boldsymbol{\gamma}|\boldsymbol{\Psi})}{f(\mathbf{y}_l,\boldsymbol{\gamma}|\tilde{\boldsymbol{\Psi}})} f(\boldsymbol{\gamma}|\mathbf{y}_l, \tilde{\boldsymbol{\Psi}})\, d\boldsymbol{\gamma}}$$

$$\approx \frac{\sum_{j=1}^{M} \frac{f(\mathbf{y}_l,\boldsymbol{\gamma}_{lj}|\boldsymbol{\Psi})}{f(\mathbf{y}_l,\boldsymbol{\gamma}_{lj}|\tilde{\boldsymbol{\Psi}})} \dot{\ell}(\mathbf{y}_l, \boldsymbol{\gamma}_{lj}|\boldsymbol{\Psi})}{\sum_{j=1}^{M} \frac{f(\mathbf{y}_l,\boldsymbol{\gamma}_{lj}|\boldsymbol{\Psi})}{f(\mathbf{y}_l,\boldsymbol{\gamma}_{lj}|\tilde{\boldsymbol{\Psi}})}}$$

while the Hessian is given by

$$\mathbf{H}_l = \ddot{\ell}(\mathbf{y}_l|\boldsymbol{\Psi}) = \frac{\int \frac{f(\mathbf{y}_l,\boldsymbol{\gamma}|\boldsymbol{\Psi})}{f(\mathbf{y}_l,\boldsymbol{\gamma}|\tilde{\boldsymbol{\Psi}})} \left\{\ddot{\ell}(\mathbf{y}_l, \boldsymbol{\gamma}|\boldsymbol{\Psi}) + (\dot{\ell}(\mathbf{y}_l, \boldsymbol{\gamma}|\boldsymbol{\Psi}))^2\right\} f(\boldsymbol{\gamma}|\mathbf{y}, \tilde{\boldsymbol{\Psi}})\, d\boldsymbol{\gamma}}{\int \frac{f(\mathbf{y}_l,\boldsymbol{\gamma}|\boldsymbol{\Psi})}{f(\mathbf{y}_l,\boldsymbol{\gamma}|\tilde{\boldsymbol{\Psi}})} f(\boldsymbol{\gamma}|\mathbf{y}_l, \tilde{\boldsymbol{\Psi}})\, d\boldsymbol{\gamma}} - (\dot{\ell}(\mathbf{y}_l|\boldsymbol{\Psi}))^2$$

$$\approx \frac{\sum_{j=1}^{M} \frac{f(\mathbf{y}_l,\boldsymbol{\gamma}_{lj}|\boldsymbol{\Psi})}{f(\mathbf{y}_l,\boldsymbol{\gamma}_{lj}|\tilde{\boldsymbol{\Psi}})} \left\{\ddot{\ell}(\mathbf{y}_l, \boldsymbol{\gamma}_{lj}|\boldsymbol{\Psi}) + (\dot{\ell}(\mathbf{y}_l, \boldsymbol{\gamma}_{lj}|\boldsymbol{\Psi}))^2\right\}}{\sum_{j=1}^{M} \frac{f(\mathbf{y}_l,\boldsymbol{\gamma}_{lj}|\boldsymbol{\Psi})}{f(\mathbf{y}_l,\boldsymbol{\gamma}_{lj}|\tilde{\boldsymbol{\Psi}})}} - (\dot{\ell}(\mathbf{y}_l|\boldsymbol{\Psi}))^2.$$

Given $Q(\boldsymbol{\Psi}, \tilde{\boldsymbol{\Psi}}) = \sum_l Q_l(\boldsymbol{\Psi}, \tilde{\boldsymbol{\Psi}})$, the gradient, $\mathbf{g} = \sum_l \mathbf{g}_l$ and the Hessian, $\mathbf{H} = \sum_l \mathbf{H}_l$, a Newton–Raphson algorithm with step-halving is used to compute the maximum likelihood estimates. Alternatively, an empirical Hessian can be computed from the individual gradients, $\hat{\mathbf{H}} = \sum_l \mathbf{g}_l \mathbf{g}_l^T$, and a quasi-Newton method based upon this (or another) Hessian estimate can be used. Notice that when an empirical Hessian is used, the number of subjects must be greater than the number of parameters in the model.

Because the distribution $f(\boldsymbol{\gamma}|\mathbf{y}_l, \tilde{\boldsymbol{\Psi}})$ is complicated, the Metropolis–Hastings algorithm (see Gilks, Richardson, & Spiegelhalter, 1996) is used to generate a Markov sample of random deviates $\boldsymbol{\gamma}_{lj}$. In this algorithm, the sequence of deviates $\hat{\boldsymbol{\gamma}}_{lj}$ form a Markov chain (rather than a simple random sample), and we must rely upon the Ergodic theorem (rather than the strong law of large numbers) to ensure convergence of our approximate integrals, gradient, and Hessian. In our implementation we use a multivariate normal proposal distribution centered at $\boldsymbol{\gamma}_l^*$ and with variance–covariance matrix given by the Hessian of the logarithm of the posterior distribution, $\log f(\boldsymbol{\gamma}|\mathbf{y}_l, \tilde{\boldsymbol{\Psi}})$.

3.2.1. Computational Algorithm

Given the initial estimates, the MCEM algorithm proceeds as follows:

1. Given the current estimates $\boldsymbol{\Psi}^k$, use a quasi-Newton algorithm (Gay, 1983) to estimate the modes, $\boldsymbol{\gamma}_l^*$, and Hessians, $\mathbf{H}(\boldsymbol{\gamma}_l^*)$, of the log posterior distributions, $\log(\boldsymbol{\gamma}|\mathbf{y}_l, \boldsymbol{\Psi}^k)$.
2. Set $\tilde{\boldsymbol{\Psi}} = \boldsymbol{\Psi}^k$. Generate a Metropolis sample of $\boldsymbol{\gamma}_{lj}$'s for the parameters $\tilde{\boldsymbol{\Psi}}$ for each subject l. Set $Q(\boldsymbol{\Psi}^k, \tilde{\boldsymbol{\Psi}}) = 0$.
3. Set $k = k + 1$. Use the Metropolis sample to compute the gradient and Hessian of $Q(\boldsymbol{\Psi}, \tilde{\boldsymbol{\Psi}})$, and compute the Newton–Raphson direction, $\mathbf{d}_k = -\mathbf{H}_k^{-1}\mathbf{g}_k$. Use step-halving to find the maximum $\alpha \leq 1$ such that $Q(\boldsymbol{\Psi}^k + \alpha\mathbf{d}_k, \tilde{\boldsymbol{\Psi}}) \geq Q(\boldsymbol{\Psi}^{k-1}, \tilde{\boldsymbol{\Psi}})$, and take a step $\boldsymbol{\Psi}^{k+1} = \boldsymbol{\Psi}^k + \alpha\mathbf{d}_k$.

4. Test to see if a new Metropolis sample is necessary. If a new sample is necessary, go back to Step 1, otherwise, test for convergence. if convergence has been reached, stop, otherwise, go to Step 3.

To lessen the computational burden, we begin the iterations using only twenty-five Monte Carlo replicates for each subject in Step 1. When "convergence" (the change in the value $Q(\cdot, \cdot)$ is small) with twenty-five replicates is reached in Step 4, we continue the algorithm with four times the number of replicates and a tighter convergence criterion. Proceeding in this manner, we stop the algorithm when convergence at a given number, M_0, of replicates has been obtained. Here M_0 is chosen because of computational cost or because it yields a reasonably good approximation of the likelihood.

It is theoretically possible to compute maximum likelihood estimates based upon a single, very large Metropolis sample and thus never return to Step 1. There are two reasons to avoid this: (a) it is computationally more efficient to begin with small sample sizes, and then increase the sample size as the precision in the estimates increases, as was discussed above, and (b) as the iterations proceed, the initial sample (based upon $\tilde{\boldsymbol{\Psi}}^k$) becomes further removed from the current estimates, $\tilde{\boldsymbol{\Psi}}$. As this occurs, the variability in the quadrature approximations increases, so that larger sample sizes are required. To avoid unnecessarily large sample sizes, whenever the function $Q(\boldsymbol{\Psi}, \tilde{\boldsymbol{\Psi}})$ is greater than a constant value η_0, we restart in Step 1 with a new Metropolis sample. Because $Q(\cdot, \cdot)$ is the log of a likelihood ratio, we use the number of subjects as our initial value for η_0. Values that depend upon percentiles of the Chi-squared distribution might be preferred.

Because the Metropolis algorithm is used to generate the Markov sample, it is relatively easy to replace the multivariate distribution assumed for the random effects with a more general distribution. In particular, we have utilized a multivariate T distribution in which the degrees of freedom, ν, is also estimated.

3.3. The Nonparametric Algorithm

Instead of a multivariate normal distribution, in the nonparametric model the distribution of the random effects is left completely unspecified. In this case, Mallet (1986) shows that the maximum likelihood estimates of the random effects distribution, $g(\boldsymbol{\gamma}|\boldsymbol{\theta})$, is discrete with at most r points (where $r \leq m$ and m is the number of subjects). That is,

$$g(\boldsymbol{\gamma}|\boldsymbol{\theta}) = \sum_{k=1}^{r} \tau_k \delta(\boldsymbol{\gamma}_k),$$

where $\delta(\boldsymbol{\gamma}_k)$ is the Dirac delta function placing point mass at $\boldsymbol{\gamma}_k$ with probability τ_k, and where $\boldsymbol{\theta} = (\boldsymbol{\gamma}_k, \tau_k)$. Given this form for the random effects distribution, computing exact integrals to obtain the marginal likelihood is straightforward since the likelihood contribution becomes

$$\ell(\mathbf{y}_l|\boldsymbol{\Psi}) = \log f(\mathbf{y}_l|\boldsymbol{\Psi}) = \log \int f(\mathbf{y}_l|\mathbf{X}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma})\, g(\boldsymbol{\gamma}|\boldsymbol{\theta})\, d\boldsymbol{\gamma} = \int f(\mathbf{y}_l, \boldsymbol{\gamma}|\boldsymbol{\Psi})\, d\boldsymbol{\gamma},$$

$$= \log \sum_{k=1}^{r} \tau_k f(\mathbf{y}_l|\mathbf{X}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma}_k).$$

Notice that the random effect distribution introduces parameters τ_k and $\boldsymbol{\gamma}_k$, so that the number of these parameters may increase with the number of observed subjects. Also notice the similarities of the parameters in this likelihood with the conditional model likelihood.

Here the τ_k provide the *prior* probabilities for each of the points $\boldsymbol{\gamma}_k$. The *posterior* probabilities are given as

$$\kappa_{lk} = \frac{\tau_k f(\mathbf{y}_l, |\mathbf{X}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma}_k)}{\sum_{i=1}^{r} \tau_i f(\mathbf{y}_l|\mathbf{X}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma}_i)}.$$

Here κ_{lk} is the probability that the l-th subject has random effects $\boldsymbol{\gamma}_k$ given the vector $\mathbf{y}_l$. In our experience, different subjects tend to have posterior probability near 1.0 on one of the κ_{lk}'s, with the remaining κ_{lk}'s near zero. Observations placing significant weights on the same $\boldsymbol{\gamma}_k$ can be thought of as a cluster of observations.

An EM-like algorithm can be used to optimize the likelihood. The parameters τ_k are estimated during the "E" step, while the remaining parameters in the model $(\mathbf{X}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma}_k)$ are estimated during the "M" step. The algorithm we use is adapted from algorithms discussed by Aitkin and Aitkin (1996), Aitkin (1996), Davidian and Giltinan (1995), Mallet (1986), Schumitzky (1991), Laird (1978), and in other references. To speed convergence of the EM algorithm, we use conjugant gradient acceleration methods discussed by Jamshidian and Jennrich (1993, 1997).

3.3.1. Computational Algorithm

Given the initial estimates for the $\hat{\boldsymbol{\gamma}}_k$, the algorithm proceeds as follows:

1. Standardize the current estimates of $\hat{\boldsymbol{\gamma}}_k$ to a mean of zero while simultaneously translating the model intercepts. This has no effect on the log-likelihood, but it helps to uniquely identify the estimates.
2. Compute the prior probabilities $\hat{\tau}_k^{c+1}$. Let $\boldsymbol{\Psi} = (\mathbf{X}, \sigma^2, \boldsymbol{\beta})$. Then these are computed as

$$\hat{\tau}_k^{c+1} = \frac{1}{m} \sum_{l=1}^{m} \frac{\hat{\tau}_k^c f(\mathbf{y}_l | \hat{\boldsymbol{\Psi}}^c, \boldsymbol{\gamma}_k^c)}{\sum_{s=1}^{r} \hat{\tau}_s^c f(\mathbf{y}_l | \hat{\boldsymbol{\Psi}}^c, \boldsymbol{\gamma}_s^c)},$$

   where c indicates the estimates on iteration c.
3. For each k, compute $\boldsymbol{\gamma}_k^{c+1}$ by maximizing the log-likelihood with respect to $\boldsymbol{\gamma}_k$. We use a quasi-Newton algorithm (Gay, 1983).
4. Check to see if any $\max |\boldsymbol{\gamma}_k^{c+1} - \boldsymbol{\gamma}_j^{c+1}| < \epsilon$ for $k \neq j$. Merge any vectors $\boldsymbol{\gamma}_k^{c+1}$ satisfying this criterion.
5. Optimize the likelihood with respect to the remaining model parameters $\boldsymbol{\Psi} = (\mathbf{X}, \boldsymbol{\beta}, \sigma^2)$. We use a quasi-Newton algorithm (Gay, 1983).
6. Check for convergence (maximum change in all parameters is sufficiently small). If convergence has not been reached, go to Step 2, otherwise go on to the next step.
7. Verify that no additional vectors $\boldsymbol{\gamma}_k$ can be added to the random effects distribution. If r equals m, no more vectors can be added. Otherwise, maximize with respect to $\boldsymbol{\gamma}$ the function

$$\phi_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = \sum_{l=1}^{m} \frac{f(\mathbf{y}_l | \hat{\boldsymbol{\Psi}}, \boldsymbol{\gamma})}{\sum_{s=1}^{r} \hat{\tau}_s f(\mathbf{y}_l | \hat{\boldsymbol{\Psi}}, \boldsymbol{\gamma}_s)}.$$

   Let $\hat{\boldsymbol{\gamma}}$ denote this maximum. If $\phi_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\gamma}})$ is within ϵ_2 of m, no further $\boldsymbol{\gamma}_k$ can be added. In this case, proceed to the next step. Otherwise, add the point $\hat{\boldsymbol{\gamma}}$ to the set of $\boldsymbol{\gamma}_k$'s, and go back to Step 1.
8. Standardize the estimated $\hat{\boldsymbol{\gamma}}_k$'s to a mean of zero while simultaneously translating the model intercepts.

Other similar EM based algorithms omit Step 7 (discussed in Davidian & Giltinan, 1995), but we find that we occasionally add random effect vectors in this step, and thus find it to be important.

Steps 3 and 5 could be combined, but this would result in an optimization problem with a very large number of parameters. Separate optimization seems to work as well, though its impact on the conjugant gradient acceleration is not known.

It is usually the case that $r < m$ vectors $\boldsymbol{\gamma}_k$ are required to optimize the likelihood, especially when the number of observations on each subject is small. This can yield a model similar to the CLASCAL model of Winsberg and De Soete (1993) in that the usual effect is to combine subjects into a smaller set of classes, where the number of classes is determined by the data. Indeed, for fixed r, the likelihoods are identical. Notice that in the random effects model r is not fixed and will, in fact, increase with the number of subjects—the random effects model makes no assumptions about "latent classes". However, if the random subject weights do indeed belong to one of a few latent classes, then the nonparametric algorithm has a good chance of finding them.

3.4. Relationship With Other Models

Aside from the CLASCAL model and the constrained models mentioned above, DeSarbo, Howard, and Jedidi (1991) propose a mixture model for two-way dominance data, called the MULTICLUS model, which clusters the subjects into groups while simultaneously performing a multidimensional scaling analysis. This model is not so closely associated with the models proposed here, but, as with the CLASCAL model, a mixture of distributions is used. Notice that the random effects distribution used in our models can be generalized to include mixture distributions. A mixture of normals might be important, for example, when the subject weights depend upon an unobserved predictor or when outliers are present in the data.

Mixture distributions are also common in multidimensional unfolding. In particular, the GENFOLD2 model (see DeSarbo & Rao, 1984) is a multidimensional unfolding model that includes linear restrictions on the configuration and ideal point matrices. Moreover, Wedel and DeSarbo (1996) include latent classes in the GENFOLD2 model. As in the CLASCAL model, the main difference of the Wedel and DeSarbo model and the random effects model we propose is that we are dealing with random effects, not latent classes. Moreover, we allow linear models for predicting the (expected log) subject weights. The restrictions imposed by the GENFOLD2 models are qualitatively different from the linear models we use, as is discussed above.

4. Algorithm Verification

Before considering an example involving real data, we first verify that the algorithms are working correctly, and also illustrate the estimates that might be expected, by fitting a Monte Carlo data set generated from known population values. For this example we consider the population configuration matrix given in Table 1.

The model for the subject weights includes one predictor, $u_{l1} = x_l$, for each dimension. This predictor is 1.0 for five of the subjects, while for the remaining five subjects the predictor is 2.0. The subject weights are generated according to models $\log \mathbf{w}_l = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{x}_l + \boldsymbol{\gamma}$ where the coefficients $\boldsymbol{\beta}$ and the variances of the $\boldsymbol{\gamma}$'s are given in Table 2 on lines labeled "Population". Here the $\boldsymbol{\gamma}_{li}$ are independently generated according to a normal distribution with a mean of zero and with equal standard deviations of 0.25. The measurement error standard deviations are also 0.25.

Using this model, forty-five measurements comparing all pairs of the ten stimuli were obtained on each of 10 subjects. In generating these measurements, first the subject weights for each subject were computed using the S-Plus (StatSci, 1993) pseudo-random normal generator. Given these subject weights, the model distances between the pairs of stimuli were computed, and pseudo-random normal measurement errors were added to the computed distances to obtain the subject observation.

Maximum likelihood estimates for these data were computed using the three algorithms. Differences between the population values and the Monte Carlo algorithm estimates of the configuration are indicated by the letter "m" in Figure 1. Points indicated by the letter "l" in Figure 1 refer to differences between the linearization algorithm and the population values, while points indicated by the letter "n" refer to the corresponding differences for the nonparametric

TABLE 1.
Artificial data configuration matrix

$$\mathbf{X}$$

	1	2	3
1	−2.24	0	0
2	0	0	0
3	2.24	0	0
4	0	−2.54	0
5	0	0.85	0
6	0	1.69	0
7	0	0	−2
8	0	0	−1
9	0	0	2
10	0	0	1

algorithm. The linearization algorithm estimates are closest to the population values, followed by the Metropolis algorithm, with the nonparametric algorithm giving estimates that differed most from the population values in this example. Clearly all three algorithms did a good job of recovering the configuration matrix, however, since the maximum of all of the deviations is small (0.2).

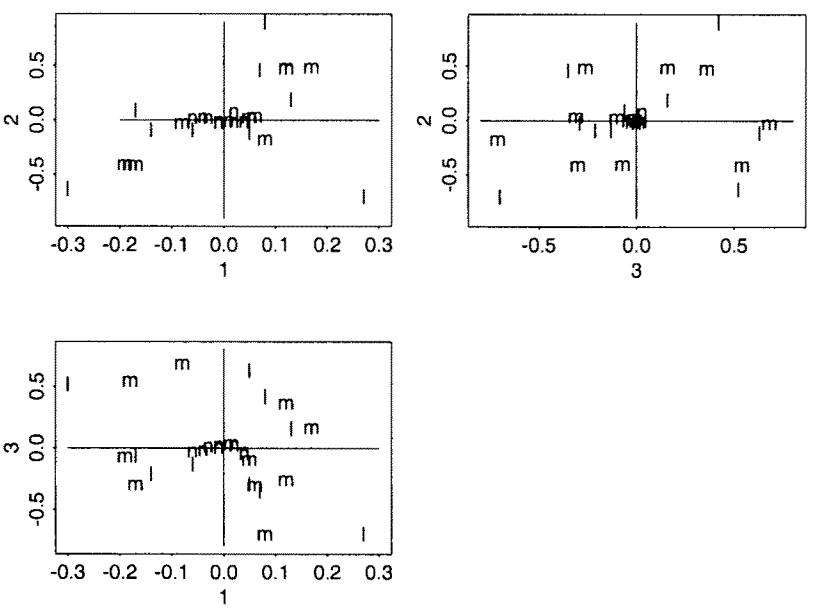

FIGURE 1.
Algorithm verification. Plots of the differences between the configuration matrix estimates and the population values; m–MCEM algorithm, n–Nonparametric Algorithm; l–linearization algorithm.

TABLE 2.
Linear model and variance component parameters

		1			2			3		
Estimates	σ	β_{01}	β_{11}	σ_{11}	β_{02}	β_{12}	σ_{22}	β_{03}	β_{13}	σ_{33}
Population	0.250	3.00	0.250	0.250	2.00	0.500	0.250	1.00	0.750	0.250
Linear	0.246	3.44	0.016	0.025	2.30	0.245	0.033	1.34	0.570	0.027
MCEM	0.246	3.45	0.022	0.255	2.23	0.244	0.243	1.34	0.571	0.246
NP	0.241	3.44	0.022		2.31	0.239		1.35	0.562	

Comparing the sums of squared differences between the various configuration matrix estimates with the population values, the smallest sums of squared differences, 0.00976, is obtained by the linearization algorithm, while the MCEM algorithm yielded 0.0282, and the nonparametric algorithm yielded 0.0815. It is, perhaps, not surprising that the nonparametric algorithm yielded the largest differences— it is most able to adapt to the individual subject weights. All three sums of squared differences are acceptably small, however.

The estimates for the linear model parameters, the standard deviation of the measurement error, and the estimates for the variance components are given in Table 2. While these estimates are not as close to their true values as the configuration matrix estimates, note that there are only ten subjects, not a large number of degrees of freedom for each linear regression. Standard errors for these coefficients were not computed, but in a simple linear regression with the same design matrix, coefficients, and residual standard error as are used here, the standard errors are 0.250 and 0.158 for the slope and intercept, respectively. The estimates here are well within what would be expected in a simple linear regression analysis. The variance components from the linearization algorithm are all too small. In general, the variance component estimates obtained from the linearization algorithm will be poor. The MCEM algorithm clearly does much better at estimating these variance components.

In this example we have a single predictor, x. Often, one will not know which predictors to include in the linear model for the subject weights. Because these are all maximum likelihood procedures, likelihood ratio tests and AIC/BIC statistics can be used to compare models, and thus to select an appropriate model. A potential problem is the correlation between the variance components and the linear models—model selection is complicated by the fact that random effects models can compensate for deletion of predictors by increasing the variances of the random effects. One thus needs to consider both the reduction in the likelihood and the reduction in the variances of the random effects, when deciding whether or not a predictor belongs in the model. This is discussed further below. A further complication in the MCEM algorithm is that the likelihood is not computed directly, so that likelihood ratios must be used.

5. Cola Example

We now turn to an example with real data. Consider data collected on ten students on the distances between ten cola drinks. These data were originally collected by Schiffman in unpublished work at Duke University in 1977, and appear in Schiffman, Reynolds, & Young (1981), whose analyses were based upon many conditional models. The ten subjects were asked to provide all 45 possible dissimilarities comparing the taste of ten brands of decarbonated, warm soda. Using the conditional model estimates as the initial estimates, we analyzed this data set using the three random effects algorithms described above, and assuming a diagonal covariance structure for the random effects. The data has a single predictor, *PTC*, that is zero or one depending upon whether or not the subject can taste the chemical "PTC". PTC tasters respond to diet drinks as being bitter, non-tasters do not. Consistent with the number of dimensions used by other authors, we used a three-dimensional solution for all models. In practice, the number of dimensions in

TABLE 3.
SAS estimates for the conditional model

X

	1	2	3
1	−0.99	0.65	−1.28
2	0.81	0.74	−0.93
3	0.54	−1.30	1.18
4	0.94	−1.43	−1.31
5	1.25	0.46	0.39
6	0.54	0.31	1.50
7	−0.94	−1.58	−0.77
8	−1.41	0.01	1.17
9	0.69	1.38	0.11
10	−1.41	0.75	−0.05

W

	1	2	3
1	11.15	7.95	8.12
2	4.45	12.04	8.94
3	5.95	13.54	9.24
4	12.46	6.15	8.23
5	14.36	2.13	6.15
6	13.76	2.72	6.30
7	8.24	12.25	11.49
8	7.08	8.12	7.18
9	12.25	6.55	7.73
10	6.40	11.16	8.76

a model may not be known. In this case AIC and BIC statistics may be used to determine the number of dimensions, or the number of dimensions can be determined from conditional model fits of the data.

For comparison, we fit a conditional model using the SAS MDS procedure (SAS, 1988). The SAS procedure MDS options were chosen for a metric model that, except for the random effects, was otherwise identical to the models fitted by our algorithms. This required us to set SAS option "Level" to "ABSOLUTE", set the "Condition" option to "MATRIX", set the "Formula" to "1", and set the "Fit" option to 1. The fitted configuration matrix and subject weights from this model are given in Table 3. Notice that the SAS model makes no use of the predictor.

We fit linear random effects models $\log w_{lk} = \beta_{0k} + \beta_{lk} u_l + \gamma_{lk}$, where β_{0k} is an intercept for the k-th log subject weight, β_{lk} is the slope, u_l is the predictor "PTC", and γ_{lk} is the random effect for the k-th log subject weight. The linearization (*nlme*) algorithm required only two (outer) iterations (using the default convergence criteria), the MCEM algorithm required 28 iterations, and the nonparametric algorithm required 41 iterations. The EM algorithm used in the nonparametric estimates can be quite slow to converge, though generally the total cpu time is much less than for the MCEM algorithm.

To conserve space we only give the configuration matrix and subject weight estimates for the linearization algorithm in Table 4. The estimates and standard errors in this table reflect relationships that are also observed in the conditional solution, and in the other algorithms as well. The standard errors are computed using standard asymptotic methods based upon the inverse of the Hessian matrix. Table 5 gives the estimates for the other model parameters, while Figure 2 plots the differences between the estimated configuration matrices and the SAS model estimates, and Figure 3 plots these same differences for the subject weight matrices.

Table 6 gives the sums of squared differences between the configuration matrix estimates from the various algorithms (the configuration matrix estimates are normalized to be as similar as possible). Sums of squared differences between the subject weights "estimates" are given in parenthesis (because the nonparametric model does not associate subject weights with individuals, the sums of squared differences are not given). For the MCEM and linearization algorithm, the subject weight "estimates" are taken as the modes of the posterior distribution of the random effects given the data. In this example, the SAS and the nonparametric configuration matrix estimates are most similar (sums of squared differences of 0.025), with the MCEM estimates

TABLE 4.
Estimates from the parametric model, linearization algorithm. Standard errors are displayed in parentheses

X

	1	2	3
1	−1.16 (.12)	0.74 (.33)	−1.34 (.15)
2	0.86 (.13)	0.72 (.29)	−1.22 (.17)
3	0.62 (.14)	−0.40 (.34)	1.60 (.09)
4	0.64 (.15)	−2.06 (.16)	−0.79 (.31)
5	1.11 (.08)	0.37 (.19)	0.18 (.15)
6	0.61 (.13)	0.77 (.27)	1.15 (.16)
7	−0.89 (.11)	−1.69 (.11)	−0.14 (.31)
8	−1.28 (.11)	0.20 (.32)	1.33 (.12)
9	0.96 (.10)	0.68 (.21)	−0.59 (.17)
10	−1.47 (.08)	0.66 (.20)	−0.18 (.14)

W

	1	2	3
1	12.09	6.02	7.73
2	5.03	7.58	9.83
3	6.95	9.06	9.40
4	12.87	4.64	8.29
5	14.56	2.68	5.33
6	14.03	2.87	5.97
7	8.96	8.85	11.03
8	7.21	7.31	6.57
9	12.43	4.28	8.75
10	7.31	8.33	8.18

and the linearization algorithm configuration matrix estimates the least similar to the SAS model estimates, and also somewhat unsimilar with one another. These results can be explained by the fact that the nonparametric and the conditional SAS estimates place the fewest restrictions on the subject weights while the linearization model and MCEM algorithm estimates shrink the random effect estimates toward zero, the mean of the multivariate normal distribution.

Standard asymptotic normal theory tests and confidence intervals can be computed for the coefficients estimated in the various algorithms. For example, consider β_{l1}. In the linearization algorithm, the asymptotic standard error estimate of this slope, computed from the inverse of the Hessian of the log-likelihood, is 10.8. This gives an asymptotic standard normal score of $z = 0.64/10.8 = 0.059$ for the test that the slope is zero, clearly not a significant value. Likelihood ratio tests on the coefficients are also possible: performing the same test using likelihood ratios, we fit the model in which the "PTC" predictor for $\log w_{l1}$ is omitted. This gives log-likelihood -904.6249, slightly larger than the log-likelihood when the "PTC" term is included (this is possible only because the linearization algorithm uses an approximation to the likelihood). This test also reveals that the "PTC" effect is not significant in the linearization algorithm.

It is also possible to test the hypothesis that the variance of a random effect is zero. This is accomplished by fitting a model in which the random effect is not included, and then performing the usual likelihood ratio test. For example, to test that $H_0: \sigma_{11}^2 = 0$ versus the alternative that $H_1: \sigma_{11}^2 > 0$, we first fit a model in which $\log w_{l1} = \beta_{01} + \beta_{1l} u_l$ (and all $\gamma_{l1} = 0$). This gives log-likelihood of -907.18. Then the likelihood ratio test statistic is computed as $2 \times (-904.72 +$

TABLE 5.
Linear model and variance component parameters

			1			2			3		
Estimates	σ	$\ell(\hat{\Psi})$	β_{01}	β_{11}	σ_{11}	β_{02}	β_{12}	σ_{22}	β_{03}	β_{13}	σ_{33}
Linear	1.81	904.72	3.61	0.64	0.13	5.13	−0.74	0.10	4.72	−0.22	0.11
MCEM	1.87		2.37	0.69	0.00	4.91	−1.10	0.00	4.49	−0.04	0.00
NP	1.77	−480.77	3.02	0.87		6.11	−1.60		4.45	−0.33	

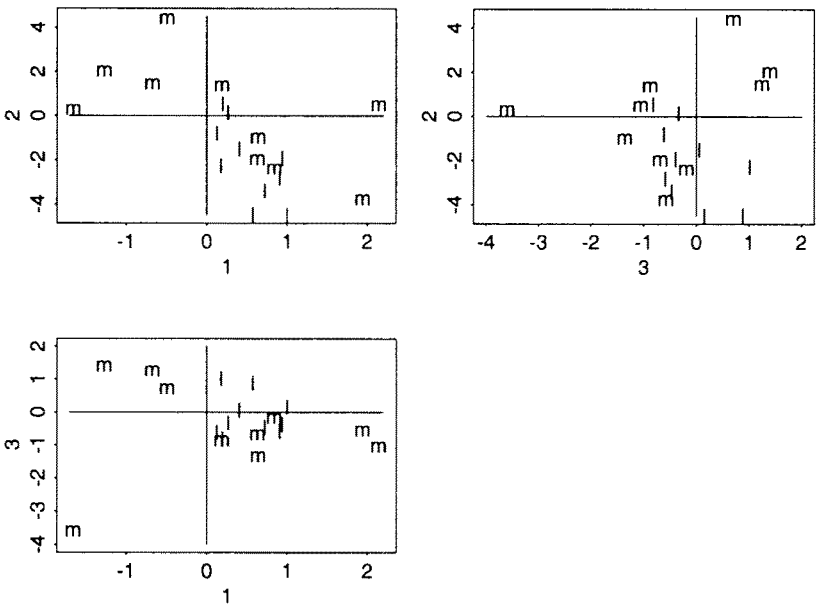

FIGURE 2.
Cola data. Plots of the differences between the configuration matrix estimates and the conditional model (SAS) estimates; m–MCEM algorithm, n–Nonparametric Algorithm; 1–linearization algorithm.

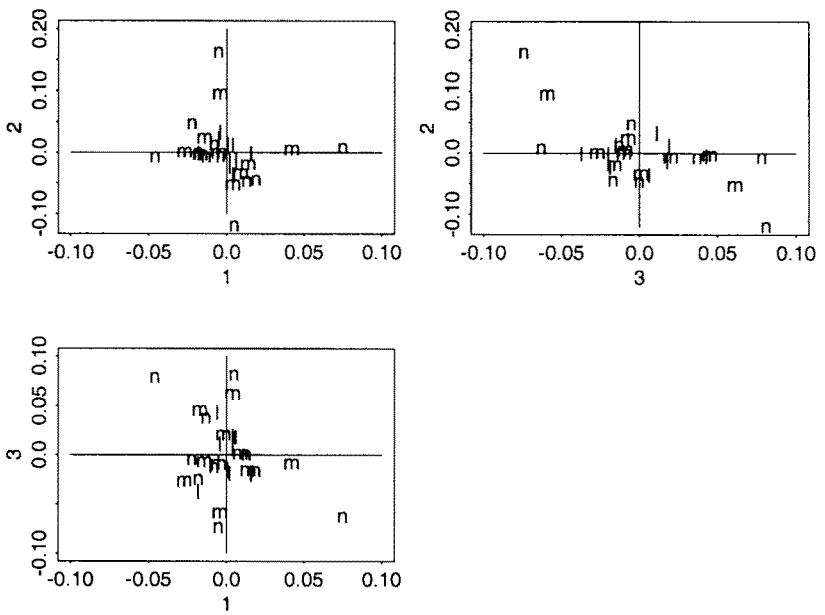

FIGURE 3.
Cola data. Plots of the differences between the subject weight matrix estimates and the conditional model (SAS) estimates; m–Metropolis algorithm, 1–linearization algorithm.

TABLE 6.
Sums of squared differences between the configuration (and subject weights) matrices for the estimates

	SAS	Linear	MCEM	Nonparametric
SAS		3.8601 (79.2819)	3.0931 (89.1881)	0.0253
Linear			1.2266 (142.8892)	3.9556
MCEM				3.1444

907.18) = 4.92. This statistic, statistically significant, is asymptotically distributed according to the chi-squared distribution with one degree of freedom.

A test that the β_{11} coefficient is zero for the MCEM algorithm is given as $z = 0.69/0.114 = 6.05$. Unlike the linearization algorithm, this value is clearly significant. On the other hand, after accounting for differences in the subject weights due to the predictor, the estimate of the variance component σ_{11} is zero. This results in a model in which the subject weights on dimension 1 are entirely predicted by the ability to taste "PTC". Differences between the linearization and MCEM algorithms can be explained, in part, by the approximate likelihood used in the linearization algorithm. In addition, one must also consider collinearity between the variance component σ_{11} and the parameter β_{11}. The approximation used in the linearization algorithm places less emphasis on the slope β_{11}, which adds variability in the random effects γ_{11}; thus, while $\hat{\beta}_{11}$ is not significantly different from zero, the variance component $\hat{\sigma}_{11}$ is significant. The MCEM likelihood, on the other hand, emphasizes the effect of the predictor so that $\hat{\beta}_{11}$ is significantly different from zero, while the $\hat{\sigma}_{11} = 0$. This illustrates an important aspect of including predictors in a model—they can reduce the variance of the random subject weights. Continuing, we fit an MCEM model in which the PTC predictors was omitted. In this new model, $\hat{\sigma}_{11} = 0.24$ with standard error estimated (obtained from a generalized inverse of the Hessian) of 0.066. This yields $Z = 0.24/0.066 = 3.61$ in a test that $H_0: \sigma_{11} = 0$, clearly significant at the 5 percent level.

Likelihood ratio tests are also possible when the MCEM algorithm is used, though the log-likelihood cannot be computed directly. Rather, likelihood ratio tests are computed through the function $Q(\boldsymbol{\Psi}, \tilde{\boldsymbol{\Psi}})$. For example, to test the hypothesis $H_0: \boldsymbol{\Psi} = \boldsymbol{\Psi}_0$ against the alternative that $H_1: \boldsymbol{\Psi} \neq \boldsymbol{\Psi}_0$, the statistic $\chi^2 = 2Q(\hat{\boldsymbol{\Psi}}, \boldsymbol{\Psi}_0)$ can be used. This is a likelihood ratio statistic, and has degrees of freedom equal to the number of parameters fixed by the null hypothesis.

In the nonparametric algorithm, only 4 (out of a possible 10) points $\boldsymbol{\gamma}_k$ were utilized in the estimated distribution of the random effects. These are given in Table 7, along with the estimated posterior probabilities. The posterior probabilities in this table indicate four clusters with observation numbers as follows: {1, 4, 9}, {2, 3, 5, 6, 10}, {8}, {7}. It is important to note that these clusters are obtained from the random effects—the effect of the predictor PTC has already been accounted for in the linear model.

TABLE 7.
Random effect and posterior probability estimates from the nonparametric model

	(γ)		
	1	2	3
1	−0.30	1.00	0.22
2	−0.15	−0.06	−0.01
3	0.20	−0.03	0.24
4	−0.01	−0.45	−0.17

	Posterior Probabilities									
	1	2	3	4	5	6	7	8	9	10
1	1.00	0.0	0.0	1.00	0.0	0.0	0.0	0.0	1.00	0.0
2	0.0	1.00	0.98	0.0	0.89	0.96	0.04	0.01	0.0	0.99
3	0.0	0.0	0.0	0.0	0.11	0.04	0.0	0.99	0.00	0.01
4	0.0	0.0	0.02	0.0	0.0	0.0	0.96	0.0	0.0	0.0

6. Discussion

We proposed a mixed effects multidimensional scaling model with parametric and nonparametric distributions for the random effects, provided three fitting algorithms, and gave an example analysis. We believe that in many situations the subject weights in diagonal metric models are most appropriately modeled as random effects. As discussed in the Introduction, there are many benefits to the random effects approach. Perhaps the main benefit is that the model corresponds to what has been measured—the subject weights are random effects. Another significant benefit is that the (logarithm of the) subject weights can be modeled as a linear function of predictors, providing the possibility for hypothesis testing on the subject weights simultaneous to the estimation of the configuration and subject weight matrices. Up until now, the standard approach to hypothesis testing has been to estimate the subject weight matrix using a conditional technique and then treating the subject weights as a dependent variable in a subsequent analysis. This "two-step" procedure, though not always inappropriate, especially when the number of observation per subject is large, is inefficient, and leads to questionable significance levels (since the distributions of the subject weights is not known), allows overfitting of individual subject weights, and is known to be biased (Vonesh & Chinchilli, 1997). The ability offered by the proposed random effects models to test hypotheses on subject weights in a unified and direct manner will be useful to researchers. We believe that the benefits of random effects multidimensional scaling models outweigh the extra computational burden involved in fitting them.

Because random effects models require one to model the distribution of the subject weights, misspecification of this distribution becomes possible, as do "outliers" in the random effects. Outliers, for example, the individual with a cold in the wine study discussed earlier in the paper, are not possible in conditional models, because the subject weights in these models are fixed parameters to be estimated separately for each subject. Outlying observations are often the most important observations, so the introduction of the possibility of outliers should not necessarily detract from random effects models. One way to model the random effects that adapts to this new class of "outliers" is to fit a model for the mean that is able to adjust appropriately to each subject, for example, B-spline predictors might be used. Another possibility is to assume a mixture distribution for the random effects. One such mixture is discussed by Davidian and Giltinan (1995). These, and other alternatives, will be considered in more detail in a future paper. Finally, an alternative that also eliminates the possibility of misspecification of the distribution of the random effects, is to use the nonparametric model.

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, *6*, 251–262.

Aitkin, M., & Aitkin, I. (1996). A hybrid EM/Gauss–Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing*, *6*, 127–130.

Bentler, P.M., & Weeks, D.G. (1978). Restricted multidimennsional scaling models. *Journal of Mathematical Psychology*, *17*, 138–151.

Bloxom, B. (1978). Constrained multidimensional scaling in *N* spaces. *Psychometrika*, *43*, 397–408.

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

Carroll, J.D., De Soete, G., & Pruzansky, S. (1989). An evaluation of five algorithms for generating an initial estimate for SINDSCAL. *Journal of Classification*, *6*, 105–119.

Carroll, J.D., Pruzansky, S., & Kruskal, J.B. (1980). CANDELINC: A general approach to multidimensional analysis of many-way arrays with linear constraints on the parameters. *Psychometrika*, *45*, 3–21.

Clarkson, D.B., & Gentle, J.E. (1986). Methods for multidimensional scaling. In David M. Allen (Ed.), *Proceedings of the 17th Symposium on the Interface* (pp. 185–192). Amsterdam: Elsevier.

Davidian, M., & Giltinan, D. (1995). *Nonlinear models for repeated measures data*. New York: Chapman and Hall.

de Leeuw, J., & Heiser, W.J. (1980). Multidimensional scaling with restrictions on the configuration. In P.R. Krishnaiah (Ed.), *Multivariate analysis* (Vol. V, pp. 501–522). Amsterdam, The Netherlands: North-Holland.

Demidenko, E. (1997). Asymptotic properties of nonlinear mixed effects models in large samples. In T.G. Gregoire, D.R. Brillinger, P.J. Diggle, E. Russek-Cohen, W.G. Warren, & R.D. Wolfinger (Eds.), *Modelling longitudinal and spatially correlated data: Methods, applications, and future directions* (pp. 49–62). New York, NY: Springer-Verlag.

DeSarbo, W.S., Howard, D.J., & Jedidi, K. (1991). MULTICLUS: A new method for simultaneously performing multidimensional scaling and cluster analysis. *Psychometrika, 78*, 121–136.

DeSarbo, W.S., & Rao, V.R. (1984). Genfold2: A set of models and algorithms for the general unfolding analysis of reference/dominance data. *Journal of Classification, 1*, 147–186.

Gay, D.M. (1983). Algorithm 611. Subroutines for unconstrained minimization using a model/trust-region approach. *ACM Transactions on Mathematical Software, 9*, 503–524.

Geyer, C.J. (1996). Estimation and optimization of functions. In W.R. Gilks, S. Richardson, & D.J. Splegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 241–258). New York: Chapman and Hall.

Geyer, C.J., & Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B, 54*, 657–699.

Gilks, W.R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Marov Chain Monte Carlo. In W.R. Gilks, S. Richardson, & D.J. Splegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 1–19). New York: Chapman and Hall.

Jamshidian, M., & Jennrich, R.I. (1993). Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association, 88*, 221–228.

Jamshidian, M., & Jennrich, R.I. (1997). Acceleration of the EM algorithm using quasi-Newton methods. Manuscript submitted for publication.

Laird, N.M. (1978). Nonparametric maximum likelihood estimates of a mixing distribution. *Journal of the American Statistical Association, 73*, 805–811.

Lindstrom, M.J., & Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics, 46*, 673–687.

Mallet, A. (1986). A maximum likelihood estimation method for random coefficient regression models. *Biometrika, 73*, 645–656.

McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear models. *Journal of the American Statistical Association, 92*, 162–170.

McLachlan, G.J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: John Wiley & Sons.

Pinheiro, J.C., & Bates, D.M. (1995). *Mixed effects models, methods, and classes for S and S-Plus* (Report No. 89). Madison, Wisconsin: University of Wisconsin, Madison.

Ramsay, J.O. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society, Series B, 45*, 285–312.

SAS Institute. (1988). *SAS language guide*. Cary, North Carolina: Author.

Schiffman, S.S., Reynolds, M.L., & Young, F.W. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*. New York: John Wiley and Sons.

Schumitzky, A. (1991). Nonparametric EM algoritms for estimating prior distributions. *Applied Mathematics and Computations, 45*, 143–157.

Sheiner, L.B., Rosenberg, B., & Melmon, K.L. (1972). Modeling of individual pharmacokinetics for computer aided drug dosing. Computers and Biomedical Research, 5, 441–459.

StatSci, A Division of MathSoft. (1993). S-Plus reference manual (Vol. I). Seattle, WA: Author.

Takane, Y., Young, F.W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling properties. *Psychometrika, 42*, 7–67.

Tanner, M.A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*. New York: Springer-Verlag.

Vonesh, E.F., & Chinchilli, V.M. (1997). *Linear and nonlinear models for the analysis of repeated measurements*. New York: Marcel Dekker.

Wedel, M., & DeSarbo, W.S. (1996). An exponential-family multidimensional scaling mixture methodology. *Journal of Business and Economic Statistics, 14*, 447–459.

Wei, G.C.G., & Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association, 85*, 699–704.

Winsberg, S., & De Soete, G. (1993). A latent class approach to fitting the weigthed Euclidean model, CLASCAL. *Psychometrika, 58*, 315–330.

Manuscript received 2 MAR 1998
Final version received 14 JUL 1999