

PROBABILISTIC MULTIDIMENSIONAL SCALING: COMPLETE AND INCOMPLETE DATA

JOSEPH L. ZINNES

UNIVERSITY OF ILLINOIS

DAVID B. MACKAY

INDIANA UNIVERSITY

Simple procedures are described for obtaining maximum likelihood estimates of the location and uncertainty parameters of the Hefner model. This model is a probabilistic, multidimensional scaling model, which assigns a multivariate normal distribution to each stimulus point. It is shown that for such a model, standard nonmetric and metric algorithms are not appropriate.

A procedure is also described for constructing incomplete data sets, by taking into consideration the degree of familiarity the subject has for each stimulus. Maximum likelihood estimates are developed both for complete and incomplete data sets.

Key words: multidimensional scaling, nonmetric scaling, maximum likelihood estimation, complete and incomplete data, noncentral chi-square approximations.

In this paper we develop a procedure for obtaining maximum likelihood (ML) estimates of the parameters of a probabilistic, multidimensional scaling model. The data to be analyzed are assumed to consist of the judgments of the dissimilarity between pairs of stimuli, or of equivalent responses.

The maximum likelihood estimator has been selected because of its optimal properties, at least for large samples. Furthermore, since for such samples the sampling distribution of the ML estimator and that of the likelihood ratio statistic are well known, it is a relatively straight-forward matter to carry out statistical tests of goodness-of-fit or those of parameter invariance and dimensionality. It is shown that standard nonmetric or metric estimation procedures are not appropriate for the multidimensional model treated in the paper, and in fact the use of these estimation procedures can lead to some highly degenerate or pathological solutions.

The ML procedure developed contains two basic components: (i) a simple approximation of the likelihood function, and (ii) simple expressions for the initial estimates of the unknown parameters. These estimates are used as the starting values (SV) of the iterative procedure that maximizes the likelihood function. The first part of the paper focuses on the problem of obtaining a mathematically tractable approximation to the likelihood function, and the latter part of the paper on the problem of obtaining simple, but effective, SV estimates of the unknown parameters.

The specific probabilistic, multidimensional model used is the one proposed by Hefner [1958]. Other probabilistic models have been developed [Richardson, 1938; Bechtel, 1976; Ramsay, 1977], but the Hefner model seems to be a more natural multidimen-

This research was supported by National Science Grant No. SOC76-20517. The first author would especially like to express his gratitude to the Netherlands Institute for Advanced Study for its very substantial help with this research.

Requests for reprints should be addressed to Joseph Zinnes, School of Social Science, 220 Lincoln Hall, University of Illinois, Urbana, Illinois 61801.

sional extension of the single-dimensional model of Thurstone [1927]. Each stimulus in the Hefner model is characterized, on each dimension, by a location parameter and a variability parameter, which in the one-dimensional case corresponds exactly to the Thurstone pair comparison model [Thurstone, 1927]. Each stimulus in the Hefner model is characterized, on each dimension, by a location parameter and a variability parameter, which in the one-dimensional case corresponds exactly to the Thurstone pair comparison model [Thurstone, 1927].

The Hefner model has actually been used a great deal since 1958, but it has often not been explicitly identified. It has most frequently been used in simulation studies [for example Young, 1970; Sherman, 1972; Graef & Spence, 1979], but it has also been investigated experimentally [for example Zinnes & Wolff, 1977; MacKay & Zinnes, 1981] and it has been discussed in various theoretical ways [Suppes & Zinnes, 1963; Ramsay, 1969; Zinnes & Griggs, 1974; Zinnes & MacKay, 1981].

The goal of this paper is to obtain ML estimates of both types of parameters, the location and the variability parameters, and to do this for individual subjects without requiring either replicated data or complete data sets. Complete data sets consist of all the $n(n - 1)/2$ pairwise responses that can be obtained from a given set of n stimuli, when the order or position of the stimuli in each pair is ignored. Incomplete sets have some number of responses less than this. In this paper we are primarily interested in analyzing incomplete data sets of a specific type—to be described later—although of necessity both complete and incomplete sets are treated.

1. The Hefner Model

In this model, each stimulus S_i , $i = 1, n$, is represented by an r dimensional random vector $(X_{i1}, X_{i2}, \dots, X_{ir})$, the components of which are normally and independently distributed with mean μ_{ik} and variance σ_i^2 . From this notation it should be clear that we are assuming that the variance of each point does not change from dimension to dimension, but we are not assuming that the variances of different points on any given dimension are necessarily equal. Thus each dimension looks exactly like a Thurstone Case 3 pair comparison model [Thurstone, 1927].

The assumption that each stimulus has the same variance on all dimensions is a severe one. It implies, in effect, that the stimulus space is isotropic, that there are no dominant directions. It would be desirable to treat the more general case, which would allow for nonisotropic spaces, but such a treatment would of necessity require more specialized assumptions. It is not possible merely to allow all the variances on all of the dimensions to be completely independent. This would increase enormously the number of parameters to be estimated and would thus drastically limit the power of the model. (One possible approach, which might be useful in certain contexts, is discussed in Section 8.)

There are two possible interpretations of the variance parameter, depending on whether one wishes to analyze individual or group data. For individual data, the variance parameter can be conceptualized as the level of unfamiliarity or uncertainty the subject has concerning the nature of the stimulus. Thus, for example, the Case 3 version of the Hefner model treated here would be applicable to stimulus sets consisting of cities in the United States, whose locations are not equally familiar to the subject. When the data to be analyzed are group data, the variance parameter is typically interpreted as an individual difference parameter. Since we are chiefly interested in analyzing individual data, we refer to the variance parameter in this paper as an "uncertainty" parameter.

To obtain the ML estimates of μ_{ik} and σ_i , it will be necessary to specify the density function of the distance random variable d_{ij} , defined by:

$$d_{ij}^2 = \sum_{k=1}^r (X_{ik} - X_{jk})^2.$$

We start first with the square of the "standardized" distance d_{ij}^2/σ_{ij}^2 , where we define

$$\sigma_{ij}^2 = \sigma_i^2 + \sigma_j^2. \quad (1)$$

It will also be necessary to work with the "true" distance D_{ij} between points i and j , defined by

$$D_{ij}^2 = \sum_{k=1}^r (u_{ik} - u_{jk})^2. \quad (2)$$

Then, under the assumptions of this model, it can be shown [Hefner, 1958] that d_{ij}^2/σ_{ij}^2 has the noncentral chi-square distribution $\chi'^2(v, \lambda_{ij})$, where v , the degrees of freedom, equals r (the dimensionality of the space) and λ_{ij} , the noncentrality parameter, equals $\lambda_{ij} = D_{ij}^2/\sigma_{ij}^2$. It will simplify the discussion to let $D'_{ij} = D_{ij}/\sigma_{ij}$ and $d'_{ij} = d_{ij}/\sigma_{ij}$.

Before proceeding further, it will be useful to point out the differences between the random variable d_{ij} , the true distance D_{ij} , and the expected value $E(d_{ij})$. The expected value $E(d_{ij})$ can be approximated by [Patnaik, 1949]

$$E(d_{ij}) \cong \sigma_{ij} \left[\frac{2a - (1 + b)}{2} \right]^{1/2} \quad (3)$$

where $a = v + \lambda_{ij}$ and $b = \lambda_{ij}/(v + \lambda_{ij})$ and v and λ_{ij} are as defined previously. The accuracy of (3) improves as $(v + \lambda_{ij})$ increases. We can therefore use (3) to show that as λ_{ij} approaches infinity, $E(d_{ij})$ approaches D_{ij} . Thus, in the Hefner model the expected value of the random variable d_{ij} will correspond exactly to the true distance D_{ij} only when λ_{ij} or, equivalently when D'_{ij} is indefinitely large.

To show what happens when D'_{ij} is small, it is easier to work with the squared distance d_{ij}^2 . Since the mean of the noncentral chi-square distribution is exactly equal to $v + \lambda$, it follows that the expected value of the squared distance d_{ij}^2 equals

$$E(d_{ij}^2) = v\sigma_{ij}^2 + D_{ij}^2. \quad (4)$$

From (4) it can be seen that as the true distance D_{ij} decreases toward zero, $E(d_{ij}^2)$ approaches $v\sigma_{ij}^2$. This result shows that the expected value of d_{ij} will not equal zero when the true distance D_{ij} is zero, and in fact, can be appreciably larger than zero. In particular, if σ_{ij} approaches infinity, the expected value of the distance d_{ij} will also become indefinitely large, even when the true distance D_{ij} is zero.

These properties of the Hefner model indicate that the expected distance $E(d_{ij})$ and the true distance D_{ij} are not related to each other in a simple, monotonic way. Large values of the true distance D_{ij} do not necessarily result in large values of $E(d_{ij})$. As noted, $E(d_{ij})$ can be indefinitely large even when the true distance D_{ij} is negligible. In more empirical terms, these properties of the model imply that similarity judgments of a subject, who follows the Hefner model, will not necessarily be monotonically related to the true distance, even if those distances are replicated an infinite number of times. (Some of the practical consequences of this nonmonotonicity property of the Hefner model are explored in Section 7.)

The nonmonotonicity property is also applicable to the one-dimensional case and, in fact, is well known in the older literature. It is well known that the one-dimensional Thurstone Case 3 model does not imply that equally-often-noticed-differences are equal. It is also well known that this model does not imply strong stochastic transitivity, a property closely related to the nonmonotonicity property.

Even in the preference literature, this nonmonotonicity property is well known. It has been shown, for example, that if the Thurstonian assumptions are added to the unfolding model [Coombs et al., 1961], the choice probabilities will not be monotonically related to distances. The magnitudes of the choice probabilities will depend not only on the distance

between the ideal point and the stimulus points, but also on the position of the ideal point.

2. The Distance Density Function

In the previous section it was indicated that the random variable d_{ij}^2 has a noncentral chi-square distribution. This fact can be used to obtain the density function of d_{ij} , which is the basis of the likelihood function that is to be maximized.

Starting with the distribution function $F(d_{ij})$, it follows from what has been stated thus far that

$$F(d_{ij}) = G\left(\frac{d_{ij}^2}{\sigma_{ij}^2}\right), \quad (5)$$

where G is the noncentral chi-square distribution function. Differentiating (5) by d_{ij} , and using g for the noncentral chi-square density function gives

$$f(d_{ij}) = g\left(\frac{d_{ij}^2}{\sigma_{ij}^2}\right) \frac{2d_{ij}}{\sigma_{ij}^2}, \quad (6)$$

which shows that the desired density function $f(d_{ij})$ can be expressed quite simply in terms of $g(\cdot)$. Thus, to evaluate $f(d_{ij})$, it is sufficient to concentrate on procedures for evaluating g .

Although the noncentral chi-square density function is generally expressed in terms of an infinite series of beta functions [Kendall & Stuart, 1961, p. 228], for calculation purposes it is simpler to express the infinite series recursively, giving

$$g(z) = e^{-(1/2)(z+\lambda)} \left(\frac{z}{2}\right)^{v/2} z^{-1} \sum_{k=0}^{\infty} A_k \quad (7)$$

where

$$A_0 = \frac{1}{\Gamma} \frac{v}{2} \quad (8)$$

$$A_k = \frac{\lambda z}{4k\left(k + \frac{v}{2} - 1\right)} A_{k-1} \quad (9)$$

and z is the noncentral chi-square random variable. Equation (9) makes it clear that the infinite series will converge rapidly only when λz is small or v is large. For two dimensional configurations, $v = 2$, so rapid convergence depends entirely on λz being small.

From some simple calculations, some of which are shown in Table 1, it can be demonstrated that reasonably accurate results will be obtained in five terms or less if $\lambda z < 6.5$, when $v = 2$. More specifically, letting g_k be the value of $g(z)$ when only the first k terms of (7) are summed, then it will be the case, for some value of $k \leq 5$, that

$$\frac{g_k}{g_{k-1}} < 1.00001 \quad (10)$$

when $v = 2$ and $\lambda z < 6.5$, or equivalently when:

$$D'_{ij} d'_{ij} < 2.55 \quad (11)$$

If v is greater than 2, this λz criterion will produce even higher levels of accuracy in five terms or less.

Table 1
Accuracy of the Approximation of the Density Function $f(d)$ for
Large Values of Dd ($\sigma_{ij}=1$)

D=1				D=2			
d	K	Exact	Percent Error	d	K	Exact	Percent Error
2.61	6	.1880	-1.498	1.345	6	.2808	-1.260
2.70	6	.1643	-2.028	1.660	6	.3595	.905
2.79	6	.1424	-2.509	1.975	7	.4115	1.737
2.88	6	.1224	-2.929	2.290	8	.4224	1.621
2.97	6	.1043	-3.274	2.605	8	.3895	.920
3.06	6	.0881	-3.533	2.920	9	.3233	-.043
3.15	6	.0739	-3.694	3.235	10	.2417	-.989
3.24	6	.0613	-3.747	3.550	10	.1630	-1.680
3.33	6	.0505	-3.680	3.865	11	.0991	-1.906
3.42	7	.0413	-3.481	4.180	11	.0544	-1.471
3.51	7	.0335	-3.142	4.495	12	.0270	-.187
D= $\sqrt{10}$ = 3.162				D=5			
.849	6	.0151	-5.747	1.500	10	.0005	-3.363
1.115	7	.0304	-4.396	2.187	13	.0051	-2.329
1.381	8	.0558	-2.988	2.760	16	.0243	-1.205
1.648	8	.0939	-1.733	3.561	19	.1204	-.150
1.914	9	.1456	-.725	4.133	21	.2507	.152
2.180	10	.2085	.004	4.591	23	.3536	.177
2.446	11	.2763	.458	5.049	24	.4024	.077
2.712	11	.3392	.666	5.393	25	.3853	.038
2.979	12	.3861	.671	5.900	29	.2903	-.202
3.245	13	.4080	.520	6.650	32	.1184	-.244
3.511	14	.4002	.266	7.850	36	.0086	.873

Note: Column "d" contains values of the distance random variable greater than $2.55/D$. Column "K" shows the number of terms that have to be summed if the exact expression (Eqs. 6 and 7) is used. The column "Percent Error" indicates the percent by which the approximation overestimates the correct value.

What is needed is an approximation to the density function when λz is large. For this purpose we use Sankaran's [1959] normal distribution approximation of the noncentral chi-square distribution. Of the four well known noncentral chi-square approximations, the Sankaran approximation seems best, even compared to those that use a central chi-square approximation [Zinnes & Wolff, 1977].

According to the Sankaran approximation, if

$$y = \left(\frac{z}{v + \lambda} \right)^h, \quad (12)$$

where z is the noncentral chi-square random variable, and

$$h = 1 - \frac{2}{3} \frac{(v + \lambda)(v + 3\lambda)}{(v + 2\lambda)^2},$$

then y will be approximately normally distributed with a mean of

$$\mu_y = 1 + h(h - 1) \frac{k_2}{2k_1^2} - h(h - 1)(2 - h)(1 - 3h) \frac{k_2^2}{8k_1^4} \quad (13)$$

and a standard deviation of

$$\sigma_y = \frac{h(k_2)^{1/2}}{k_1} \left[1 - \frac{(1 - h)(1 - 3h)k_2}{4k_1^2} \right] \quad (14)$$

where $k_1 = v + \lambda$ (15)

and $k_2 = 2(v + 2\lambda)$. (16)

Thus, from the Sankaran approximation, we obtain

$$G(z) \cong \Phi \left[\left(\frac{z}{v + \lambda} \right)^h \right] \quad (17)$$

where Φ is the normal distribution function, with μ and σ as defined in (13) and (14). Differentiating (17) with respect to z , and using ϕ for the normal density function gives:

$$g(z) = \phi \left[\left(\frac{z}{v + \lambda} \right)^h \right] (v + \lambda)^{-h} (hz^{h-1}) \quad (18)$$

which provides the desired approximation, the accuracy of which increases as λz increases.

Table 1 gives some idea of the accuracy that can be expected from (6) and (18) when $v = 2$, $D_{ij}d_{ij} \geq 2.55$ and $\sigma_{ij} = 1$. The column labeled "Exact" consists of values of the density function $f(d)$, using (6) and (7) and summing the number of terms shown in the previous column. The last column, "Percent Error", shows how well the approximation (18) works under these conditions. Positive percent errors indicate that the approximation overestimates the correct value, the value shown in the "Exact" column.

This table shows that the approximation of the density function $f(d)$ typically has less than 2 or 3 percent error, and frequently considerably less. The larger percent errors occur only in the extreme left hand tail of the distribution, and therefore are likely to be relatively rare.

It is also evident from Table 1 how impractical it would be to use the exact expression (7) for all values of d . For example, when $D_{ij} = 5$, $d_{ij} = 7.85$, and $\sigma_{ij} = 1$, it takes 36 terms to achieve the level of accuracy described previously in (10).

To summarize: the algorithm that we use to approximate the density function $f(d_{ij})$ has two distinct parts, depending on whether $D'd'$ is less than or greater than 2.55. If $D'd' < 2.55$, then the algorithm uses the exact expression given in (7). The number of terms summed in (7) is governed by the accuracy criterion defined in (10), but in no case will it exceed five terms. When $D'd' \geq 2.55$, then the approximation given in (18) is used. If $v = 2$, and the values of the random variable are not too far out in the left-hand tail of the distribution, this algorithm typically does not produce errors exceeding three percent. Furthermore, when v exceeds two, the accuracy of this algorithm improves considerably.

3. *Constructing Incomplete Data Sets*

Several procedures have been suggested for constructing incomplete data sets that have desirable properties. Spence and Domoney [1974] and Graef and Spence [1979] investigated cyclic designs, designs based on random deletions, and designs based on deletions of small, interpoint distances. Young and Cliff [1972], working with computer interactive procedures, have suggested a procedure which involves partitioning the stimulus set into two disjoint subsets. Complete data are obtained from one subset (the reference set) and incomplete data from the other. Young and Cliff's criterion for determining which stimuli to include in the reference set tends to select stimuli that are far apart and do not lie in a subspace of the complete space.

The procedure we have adopted for selecting incomplete data sets also partitions the stimuli into reference and nonreference sets, but does so on the basis of the magnitude of their uncertainty values. The stimuli having the lowest uncertainty values make up the reference set, and the remaining stimuli, the nonreference set. The exact size of the reference set needed to achieve a given degree of recovery will depend on how small the uncertainty values of the stimuli are. Procedures for obtaining preliminary estimates of the uncertainty values of the stimuli are described below.

Once the m reference stimuli have been selected from among the n stimuli under investigation, the incomplete data set is constructed in the following way. All the $m(m-1)/2$ interpoint distance judgments between pairs of reference stimuli are obtained, as well as all the interpoint distance judgments between each of the m reference stimuli and the $n-m$ nonreference stimuli. Thus, the judgments that are specifically excluded from the data set are those involving judgments between pairs of nonreference stimuli. The total size of the incomplete data set is then equal to $m(m-1)/2 + (n-m)m$.

As an example, consider a set of 10 reference stimuli and 20 nonreference stimuli. The complete data set for this example contains 435 judgments of interpoint distances, compared to 245 for the incomplete set. This is a reduction of about 44 percent. The percent of reduction in any particular case will depend on the relative magnitudes of m and n .

This method of selecting incomplete data sets can have excellent recovery properties [Zinnes, MacKay, & Williams, Note 1]. A smaller amount of data containing less variability can give better results than a larger amount, containing more variability.

This procedure for constructing incomplete data sets assumes that the experimenter has prior knowledge of the stimuli which are familiar to the subject and those which are unfamiliar. In most practical applications, we have found it simplest to ask subjects, prior to the multidimensional phase of the experiment, to select the m stimuli with which they are most familiar. There seems to be a reasonable relationship between these familiarity ratings and estimates of the uncertainty parameter, when the degree of unfamiliarity is not too high [MacKay & Zinnes, 1981]. When data are collected interactively at computer terminals, it is possible to improve on these initial estimates of the uncertainty parameters, by calculating new estimates after some data have been obtained.

4. *Starting Values: Complete Data*

To estimate the coordinates and variances of a given set of stimuli, it turns out to be necessary to treat the reference and nonreference stimuli separately. If this is not done, the greater variability of the nonreference stimuli will perturb the solution of the reference stimuli, degrading unnecessarily the parameter estimates of these stimuli. Since the subset of reference stimuli form a complete set of data, it is therefore necessary to describe estimation procedures for dealing both with complete and incomplete sets of data, even though we are primarily interested in the latter. We start with the complete case and focus

on the problem of obtaining good starting values (SV) for these data. The incomplete case is taken up in the next section.

There are several plausible procedures for obtaining SV estimates for the complete data case. However, the more complicated procedures that we studied did not produce appreciably better estimates of either the location or the variance parameters. Consequently we describe next just the simplest one.

Estimates of the location parameters can be easily obtained by the Young-Householder procedure [Young & Householder, 1938], which involves converting inter-point distances into scalar products and calculating the eigenvectors and eigenroots of the scalar product matrix. The accuracy of these estimates of the location parameters, obtained by ignoring the variance parameters, will depend on the magnitude of these variances. (This point is further amplified in Section 7.)

To estimate the variance parameters, we make use of the approximation [Abramowitz & Stegun, 1967, p. 943]

$$\sigma_{d_{ij}}^2 \cong \frac{\sigma_{ij}^2}{2} \left\{ 1 + b - \frac{1}{4a} [8b + (1 + b)(1 - 7b)] \right\}, \quad (19)$$

which gives the variance of the distance d_{ij} when D_{ij} (or $v + \lambda_{ij}$) is large. (The definitions of a and b in this equation are the same as those given following (3).) When λ_{ij} becomes indefinitely large, (19) shows that

$$\sigma_{d_{ij}}^2 \cong \sigma_{ij}^2 = \sigma_i^2 + \sigma_j^2. \quad (20)$$

As an estimate of $\sigma_{d_{ij}}^2$, we use

$$\hat{\sigma}_{d_{ij}}^2 = (d_{ij} - \hat{D}_{ij})^2 \quad (21)$$

where d_{ij} is the observed dissimilarity judgment of the stimuli S_i and S_j and \hat{D}_{ij} is the euclidean distance, calculated from the coordinates that have been estimated from the Young-Householder procedure. From this equation and (20) we obtain

$$(d_{ij} - \hat{D}_{ij})^2 = \hat{\sigma}_i^2 + \hat{\sigma}_j^2.$$

The right-hand side of this equation is unknown, but since it generates a simple system of $m(m - 1)/2$ linear equations, least squares estimates of the variance parameters σ_i^2 , $i = 1, m$, can be obtained readily by standard methods. In fact, the solution can be expressed explicitly as

$$\hat{\sigma}_i^2 = \frac{R_i - \frac{T_u}{m-1}}{m-2}, \quad (22)$$

where we have let R_i denote the i -th row sum of the error matrix $(d_{ij} - \hat{D}_{ij})^2$ and T_u denote the sum of the upper half of this matrix. The diagonal terms of this error matrix are assumed to be zero.

The least square estimates given in (22) will be identified as the SV estimate of the variance parameter for the complete data case.

To determine the accuracy of these SV estimates for the complete data case, a simulation was carried out consisting of 30 points, randomly located in a 2 dimensional unit circle. The value of uncertainty assigned to each point was between 0 and .3. Ten independent random samples of the distance random variable d were obtained under these conditions.

The SV estimates of the variances, based on (22), are shown in Figure 1. This figure

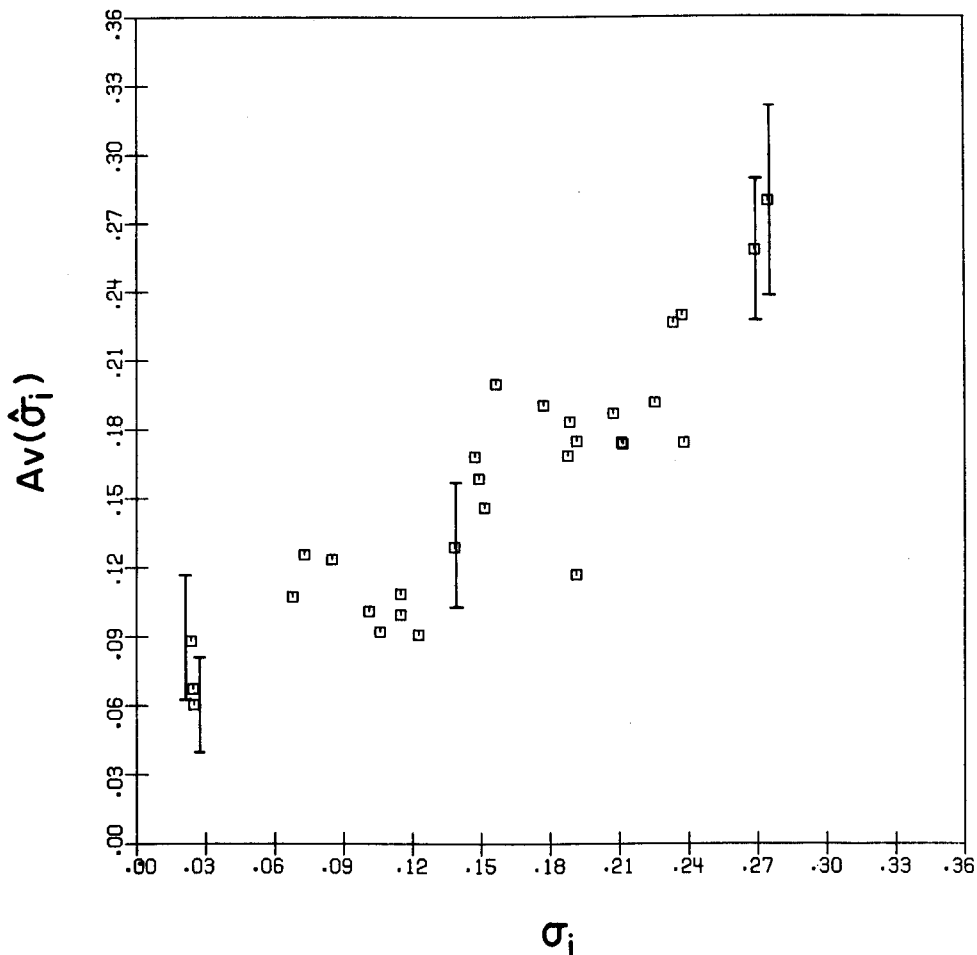


FIGURE 1.

SV estimates of the uncertainty parameter of the reference points. The standard error of estimate of 5 of the points is indicated by the vertical lines.

indicates that there is relatively little bias in the estimates of the uncertainty parameter, but there is also a fair amount of variability. This latter property is suggested by the variability of the points around the 45 degree line, and also by the sizes of the estimates of the standard error. However, these standard errors do not seem to be affected by the magnitude of the true uncertainty values.

No doubt the accuracy of these estimates of the variances would improve drastically if the interpoint distances were replicated, even to a small degree. However, it appears remarkable to us how well the variance parameters are estimated without replication and without using more complicated estimation procedures.

5. Starting Values: Incomplete Data

We assume now that the coordinates and the variances of the m reference stimuli have been estimated using the procedures described in the previous section, and we turn next to the $n-m$ nonreference stimuli.

Estimates of the coordinates of the nonreference stimuli S_f , $f = m + 1, n$, can be

obtained by minimizing

$$E_f = \sum_{i=1}^m \frac{(d_{fi} - \hat{D}_{fi})^2}{m}, \quad f = m + 1, n \quad (23)$$

where the summation on the right side is over the m reference stimuli. Since the coordinates of the reference stimuli are assumed to be known, this minimization process involves determining just the r coordinates u_{fk} , $k = 1, r$, of each of the nonreference stimuli S_f , $f = m + 1, n$; r is the dimensionality of the space.

There are several well known optimization algorithms that can be used to carry out the minimization. For the examples described later on we have used, among others, algorithms based on quasi-Newton methods (ZXMIN from the IMSL Library, 1979), steepest descent methods [Kruskal, Young, Seery, Note 2.] and stepwise search methods [Chandler, 1969]. Generally the results obtained by these different approaches have been approximately equivalent, both as to time and accuracy.

There is an alternative approach, one that is simpler computationally, but is considerably less accurate, as indicated by the figures obtained. This approach minimizes errors using squared distances, rather than the distances themselves. Thus, instead of minimizing E_f , as defined in (23), the expression

$$E'_f = \sum_{i=1}^m \frac{(d_{fi}^2 - \hat{D}_{fi}^2)^2}{m} \quad (24)$$

is minimized. The simplicity of this approach derives from the fact that it can be reduced to a standard linear multiple regression problem [Schönemann, 1970; Carroll, 1972; Bechtel, 1976]. In the present context the linearity of the problem can be seen by expanding

$$D_{fi}^2 = \sum_{t=1}^r (u_{ft} - u_{it})^2,$$

and then letting

$$y_{fi} = D_{fi}^2 - \sum_{t=1}^r u_{it}^2$$

$$a_f = \sum_{t=1}^r u_{ft}^2$$

$$b_{ft} = -2u_{ft},$$

to obtain the standard multiple regression equation

$$y_{fi} = a_f + \sum_{t=1}^r b_{ft} u_{it}. \quad (25)$$

Assuming that the origin has been placed at the centroid of the reference stimuli, the least squares solution to the regression weights b_{ft} , $t = 1, r$ is just

$$\mathbf{b}_f = (\mathbf{U}' \mathbf{U})^{-1} \mathbf{U}' \mathbf{Y}_f, \quad (26)$$

where \mathbf{b}_f is the r dimensional vector (b_{ft}), \mathbf{U} the m by r matrix (u_{it}) and \mathbf{Y}_f the m dimensional vector (y_{fi}). The SV estimates of the coordinates u_{ft} , $t = 1, r$ of nonreference stimulus S_f would then be equal to

$$\hat{u}_{ft} = \frac{-b_{ft}}{2}. \quad (27)$$

In the following discussion we pursue further the first method, based on minimizing E_f , although it may very well be the case that the computational simplicity of the alternative approach more than compensates for its lower level of accuracy.

To obtain the SV estimates of the variance parameter σ_f^2 of the nonreference stimuli $S_f, f = m + 1, n$, we make use of the approximation given in (20) to obtain

$$\sigma_{d_{fi}}^2 = \sigma_f^2 + \sigma_i^2, \quad i = 1, m, \quad f = m + 1, n. \quad (28)$$

Summing (28) over the m reference stimuli gives

$$\sum_{i=1}^m \sigma_{d_{fi}}^2 = m \sigma_f^2 + \sum_{i=1}^m \sigma_i^2, \quad f = m + 1, n, \quad (29)$$

and solving for σ_f^2 , we obtain the estimator

$$\hat{\sigma}_f^2 = \frac{1}{m} \left[\sum_{i=1}^m \sigma_{d_{fi}}^2 - \sum_{i=1}^m \sigma_i^2 \right], \quad f = m + 1, n. \quad (30)$$

The first summation on the right side of (30) can be evaluated by using (21), which in

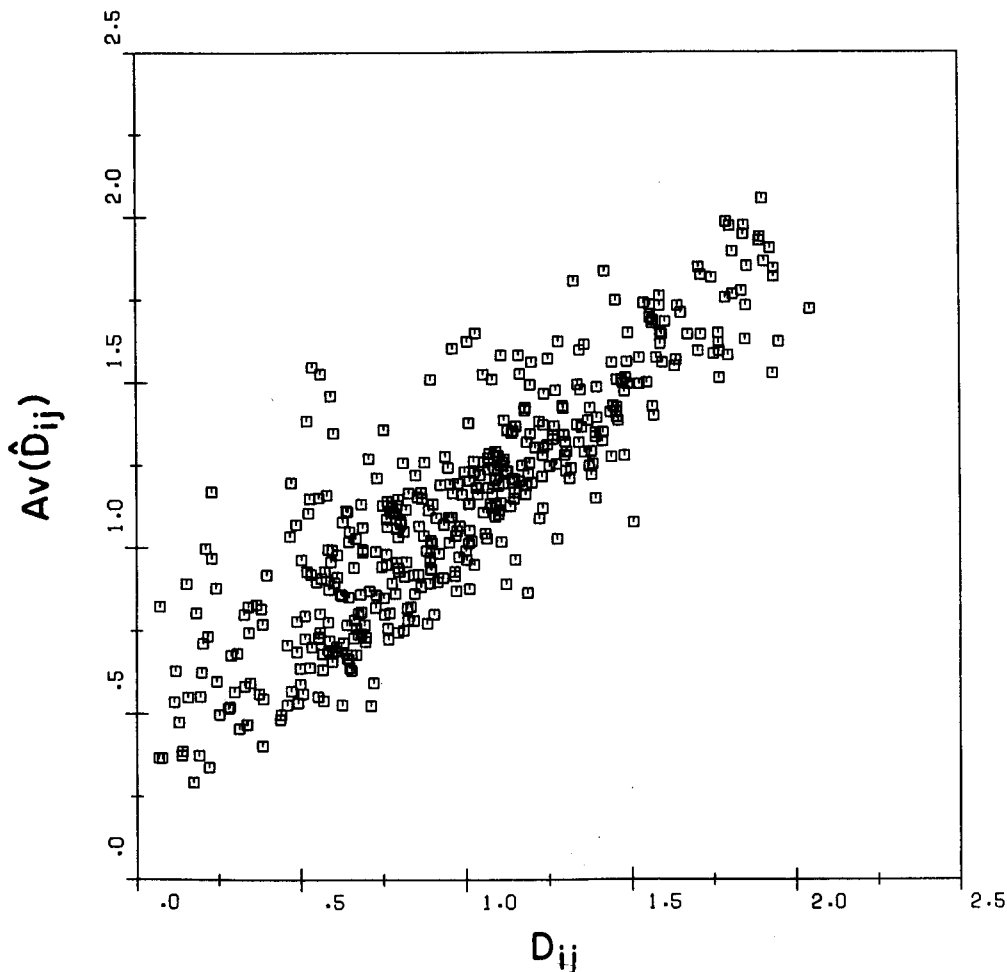


FIGURE 2.

SV estimates of the coordinates of the reference and nonreference points. The horizontal axis shows the correct values of the interpoint distances and the vertical axis the average of the estimated values.

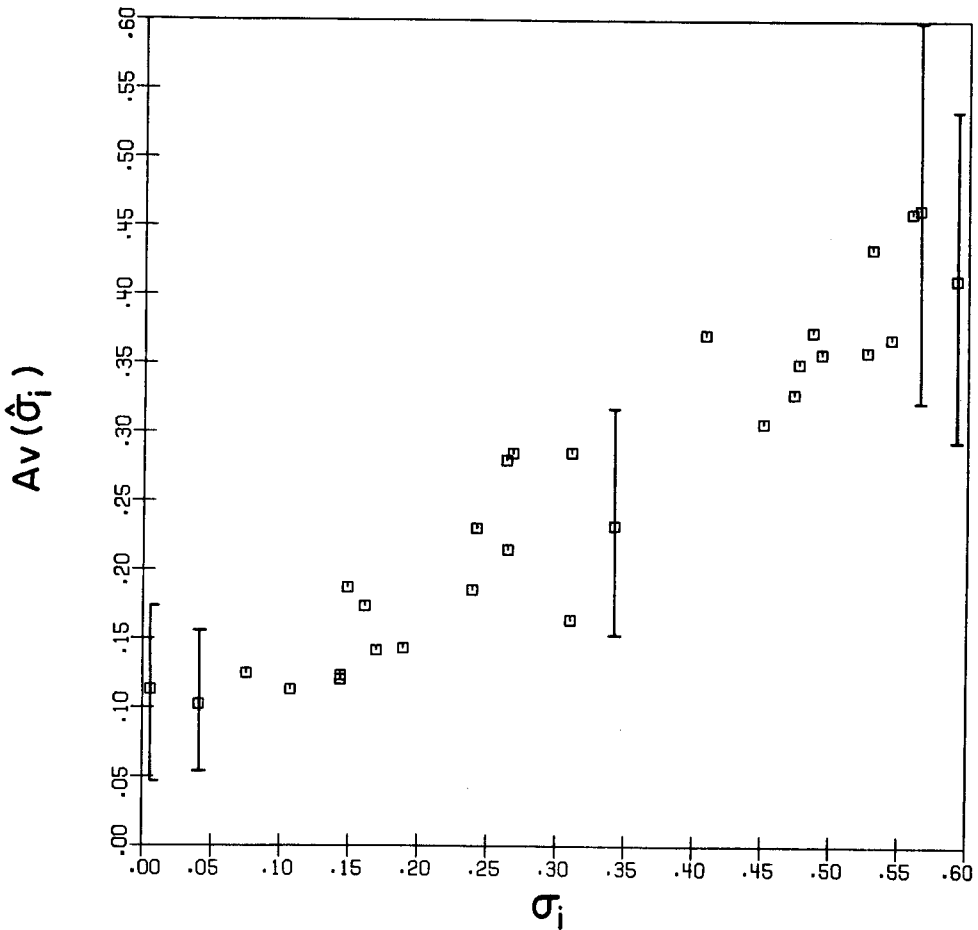


FIGURE 3.
SV estimates of the uncertainty parameter of reference and nonreference points.

fact results in E_f , as given in (23). The second summation on the right side of (30) is just the average variance of the reference stimuli, which can be calculated by the method described in the previous section.

In the examples that follow, we take (30) to be the SV estimates of the variance parameter of the nonreference stimuli.

6. Properties of the SV and ML Estimates: Incomplete Data

The simulation described next is intended to illustrate the accuracy of the SV and ML estimates for the incomplete case.

The ML estimates are obtained using the procedures described in Sections 2, 4, and 5. The approximation of the density function of the dissimilarities comes from Section 2. The sum of these density functions, one for each of the observed dissimilarities, is just equal to the log of the likelihood function. The SV estimates for the reference stimuli are obtained by the method given in Section 4 and those for the nonreference stimuli by the method given in Section 5.

A number of different iterative algorithms were used to carry out the maximization process, although primary use was made of the ZXMIN algorithm from the IMSL library [IMSL, 1979]. However, it did turn out to be useful to alternate between estimating the coordinates and estimating the variances. There was some reduction in computing time

when each one of these types of parameters was held fixed, while the other was being manipulated. (The computer program that implements all this is called PROSCAL, for Probabilistic Scaling.)

The simulated data were constructed using 30 stimuli, randomly located in a two dimensional unit circle. Each stimulus was randomly assigned an uncertainty value between 0 and .6; the ten stimuli having the smallest uncertainty values became the "reference" stimuli and the remaining stimuli the "nonreference" stimuli. The set of simulated distances consisted of the 45 interpoint distances between reference stimuli and the 200 interpoint distances between reference and nonreference stimuli.

Ten independent simulations satisfying these conditions were performed. Each simulation involved independent random samples of the relevant interpoint distances. The true location of the 30 points was the same over all ten simulations, as was the value of uncertainty assigned to each point.

The ML estimates of the coordinates and variances were obtained by maximizing the likelihood function, using the approximations described in Section 2. The SV estimates for the reference and nonreference stimuli were obtained by the procedures described in the previous two sections.

The SV estimates are shown in Figures 2 and 3, and those for the ML estimates in

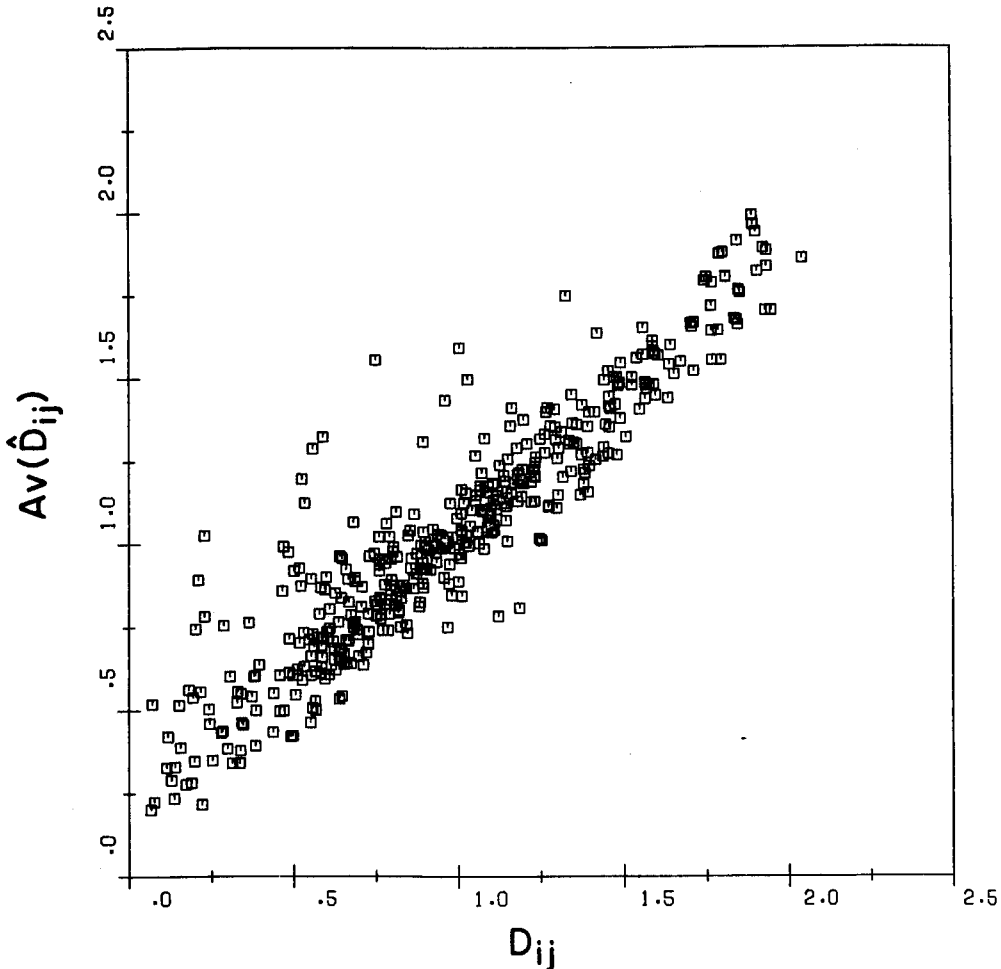


FIGURE 4.

ML estimates of the coordinates of reference and nonreference points.

Figures 4 and 5. Comparing Figures 2 and 4 suggests that the ML estimates of the location parameters are substantially better than the corresponding SV estimates, although even for the SV estimates the accuracy is amazingly good. The improvement obtained from the ML estimates appears to be mainly in the variability of the estimates, rather than in less bias.

Comparing Figures 3 and 5 suggests a somewhat different conclusion. The ML estimates of the uncertainty parameter, shown in Figure 5, do not look substantially better than the SV estimates shown in Figure 3. In fact, the ML estimates may even be worse.

These conclusions can be strengthened by using a quantitative measure of recovery. For a measure of error in recovery we use

$$ER = \frac{\sum_{i>j} (\hat{D}_{ij} - D_{ij})^2}{\sum_{i>j} D_{ij}^2} \quad (31)$$

A comparable measure of recovery can also be defined for the uncertainty parameter.

Table 2 shows the value of ER for both the SV and the ML estimates. On the basis of

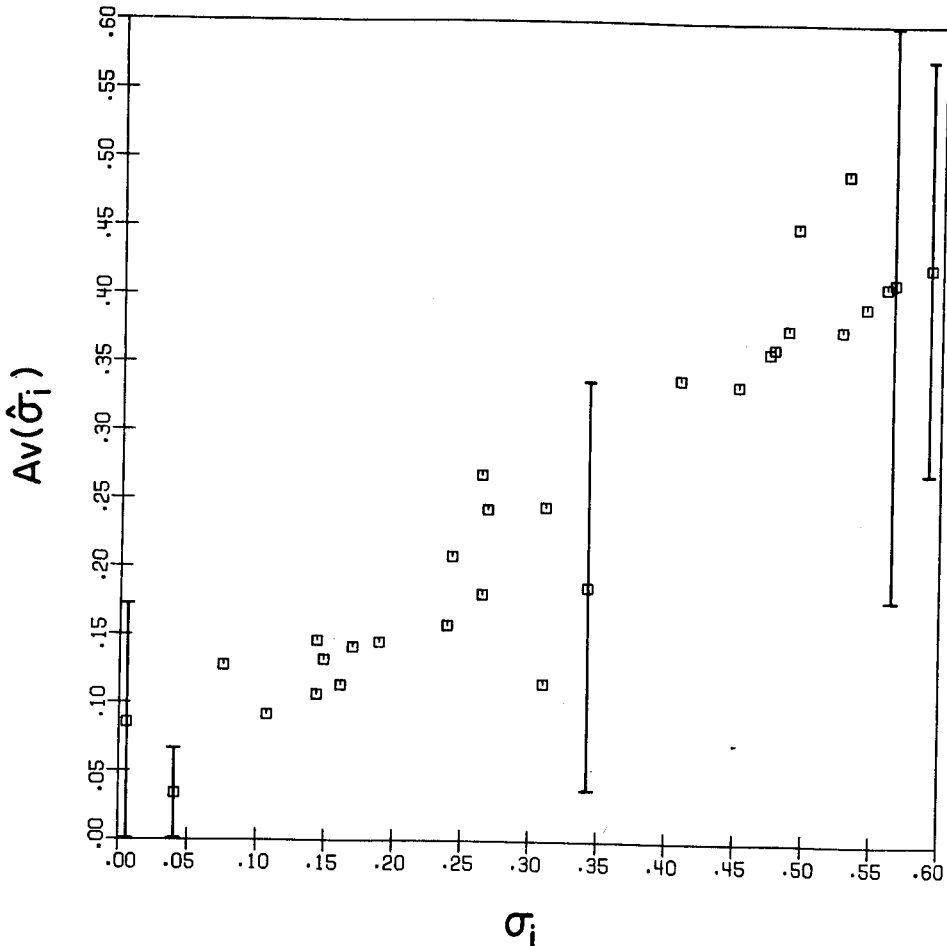


FIGURE 5.

ML estimates of the uncertainty parameter of the reference and nonreference points.

Table 2
Errors in Recovery of Distances and Uncertainty Values

Type of Estimate	Distances ER	Uncertainty Values ER
SV	.186	.136
ML	.110	.183

the ER values for the interpoint distances, the ML estimates of the location parameters (ER = .110) are definitely superior to the SV estimates (ER = .186). This is exactly what was concluded from visual inspection of the figures.

The values of ER for the variances also confirm what was suggested by the figures. According to the ER measure, the ML estimates of the variances are actually worse (ER = .183 versus .136 for the SV estimates).

We find these results surprising and disappointing. We had assumed that the ML estimates of both the coordinates and variances would be a substantial improvement over the simpler SV estimates. These results evidently reflect the small sample properties of the ML estimator. At present, it is an open question as to how large the data set would have to be for the large sample ML properties to come into play.

7. Nonmetric Solutions

In Section 1 it was pointed out that the Hefner model does not have the monotonicity property which underlies most, if not all, of the nonmetric multidimensional algorithms. The expected value of the distance in the Hefner model need not be monotonically related to the true distance. Does this property of the Hefner model have practical consequences? To answer this question, it may be useful to illustrate what happens when a typical nonmetric algorithm is applied to data generated by a subject who follows the Hefner model.

In the following three examples, simulated data were generated from three different two dimensional configurations. In each case, the data set consisted of a complete set of "observations" of all the interpoint distances in the configurations. Furthermore, in Examples 1 and 2 all the interpoint distances were replicated a fixed number of times: 30 in Example 1 and 10 in Example 2. In Example 3 the simulated data were equal to the expected values of the interpoint distances and therefore can be conceptualized as the average of an infinite number of replicated judgments of a Hefner type subject.

The same nonmetric program, KYST [Kruskal et al., Note 2], was used to analyze the simulated data in all three examples. This program was applied using both stress formulas 1 and 2, although the figures that follow show only the results for one of these formulas (stress 1 in Examples 1 and 2, and stress 2 in Example 3). Table 3 gives some of the results for both stress formulas. In fact, there were no detectable differences in the accuracy of the solutions obtained from either stress formula.

To handle the replicated data, it was assumed, in carrying out the nonmetric analyses, that a single configuration and a single regression were appropriate for all replicates. All the other KYST options used were those obtained under the default conditions.

To minimize possible problems arising from local minima, several different starting configurations were used, both for the KYST analyses and for the ML analyses. In all cases the true configuration was one of the starting configurations used, but it did not

always produce the optimal KYST solution, namely the solution having the lowest stress value. The results reported in the following figures and tables are those obtained from these optimal solutions and from solutions having a normal termination, meaning that the iterations terminated because the stress value had stabilized.

Example 1. Hexagons. In this example, 12 points were arranged in two hexagons, one inside of the other. The arrangement is shown in the upper left-hand panel, panel (a), in Figure 6. Two different values of the uncertainty parameter were used: 2.5 and 1.0. The larger value was assigned to each one of the six points forming the inner hexagon, and the smaller value to each one of the six points forming the outer hexagon. The outer hexagon was about two units wide.

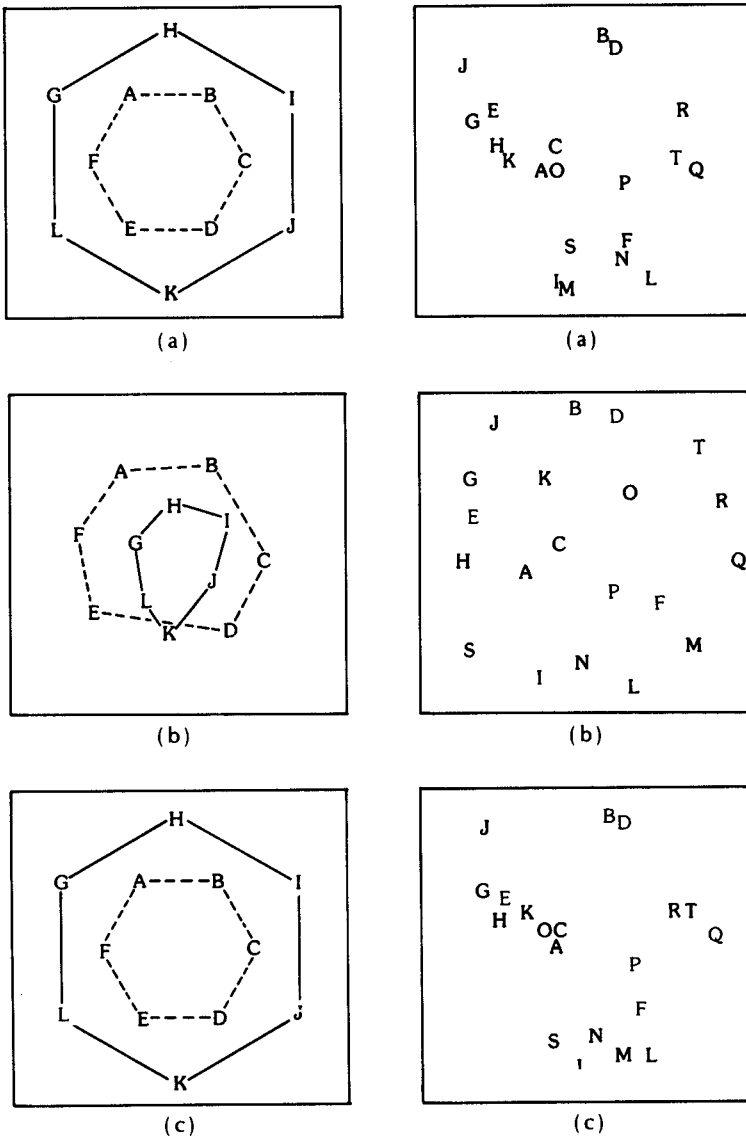


FIGURE 6.

The original and recovered configurations in Examples 1 and 2. Example 1 is on the left-hand side and Example 2 on the right-hand side of the figure. Panels (a) show the original configurations, panels (b) the configurations recovered from KYST, and panels (c) the configurations recovered from the ML approach.

The simulated data consisted of 30 replications of each one of the 45 interpoint distances.

The recovered configurations obtained from KYST and the ML approach are shown in the left-hand portion of Figure 6, panels (b) and (c), respectively. It can be seen that the KYST solution interchanged the position of the two hexagons. The hexagon that initially was on the outside is now contained almost entirely inside the hexagon that initially was on the inside.

This result is as expected, because, as shown in Section 1, the expected value of the interpoint distances will tend to reflect, almost entirely, the uncertainty values of the stimuli when these values are large relative to the true distance. Therefore, the inner hexagon, having large variances, will tend to be "perceived" by a nonmetric program as a large hexagon, having large interpoint distances.

The solution obtained from the ML approach, as shown in Figure 6, appears to be extremely accurate. There is no detectable difference between the true and recovered configurations.

Example 2. Random Points. The configuration used in this example consisted of 20 points randomly located in a 2 dimensional unit square. The actual points are shown plotted in the upper right panel, panel (a), of Figure 6. The points are labelled A through T, which corresponds exactly to the order of the magnitudes of the uncertainty values assigned to the points. Point A has the lowest uncertainty value and point T has the highest. The actual values ranged from 0 to .5 and were randomly selected from a uniform distribution.

The simulated data in this example consisted of 10 replications of each one of the 190 interpoint distances.

The configurations recovered by KYST and the ML approach are shown in the two lower right-hand panels of Figure 6, panels (b) and (c). Since these points are arranged randomly, it is somewhat difficult to determine the characteristic features of the KYST solution. These characteristics can be highlighted by determining the amount that each point in the KYST configuration has drifted, either toward or away from the centroid of the configuration, relative to its true position. To be more explicit, let us call the drift of any particular point positive if it has moved farther away from the centroid, and negative if it has moved closer to the centroid.

Although it is not obvious from Figure 6, it turns out that the drifts of the points in the KYST configuration are not completely random. The correlation between the drift of each point and its uncertainty value equals .4 ($p < .04$). This indicates that the KYST solution tends to move points having large uncertainty values away from the center of the configuration, and, conversely, it tends to move points having low uncertainty values closer to the center of the configuration. This result, on a smaller scale, is exactly what was observed in Example 1. The "stimulus drift" effect here is smaller, because the uncertainty values used ranged over a much smaller interval.

The ML solution, shown in the lower right-hand panel of Figure 6, is actually moderately close to the original configuration. The correlation of the interpoint distances between the ML and the true configuration equals .958, compared to .763 for the KYST solution.

Example 3. Expected Values. In this example, unlike the previous two, the simulated data consisted of the expected values of all of the interpoint distances, rather than just a finite number of independently sampled replications. This was done to simplify issues somewhat. We wish to show in this example what happens to the nonmetric solution when the uncertainty values become quite large and when the nonmetric results cannot possibly be attributed to perturbations resulting from samples having a finite size.

The configurations used in this example were generated by randomly selecting points along the bell-shaped curve of the normal distribution. Three different configurations were obtained, containing 10, 15, and 20 points each. However, only the results for the 10 point configuration are plotted in Figure 7, although some of the results for all three configurations are shown in Table 3. There were, in fact, no fundamental differences in the solutions obtained from these three configurations.

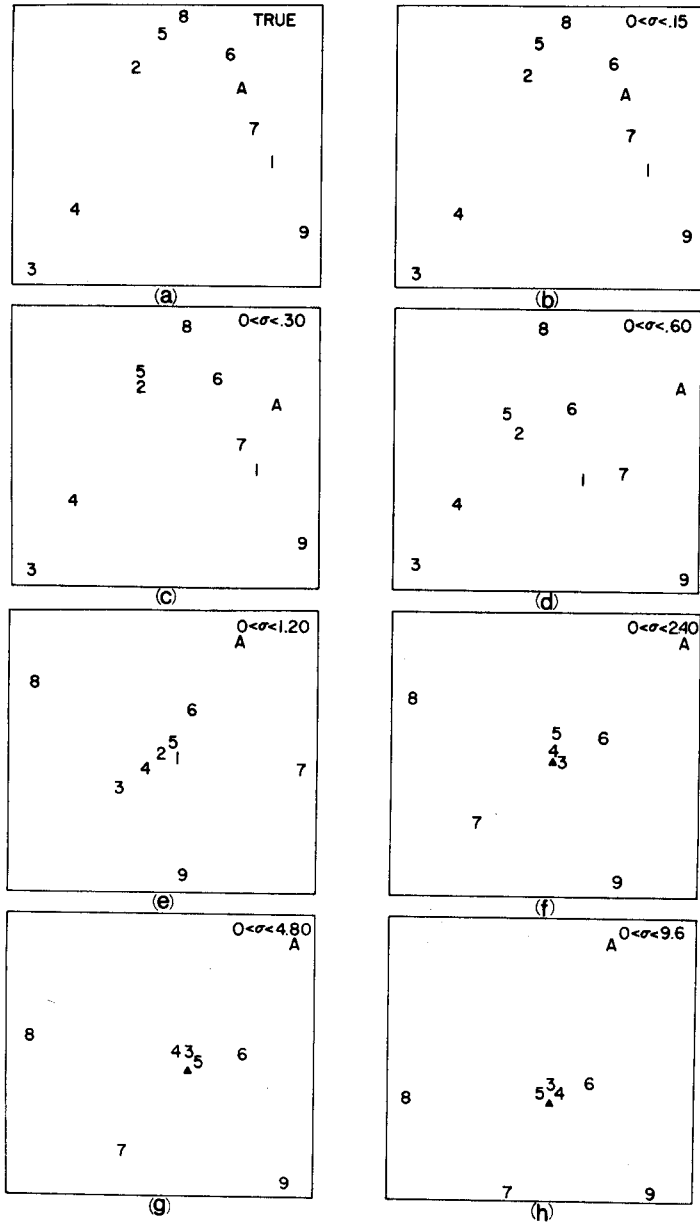


FIGURE 7.

The original and recovered configurations in Example 3. The higher numbered points have higher variances assigned to them. Point A has the highest variance. The delta symbol indicates a multiple point, consisting of points 1 and 2.

Table 3

Values of Stress Obtained Under Seven Different Levels of σ_B^*

N	Stress Formula	Values of σ_B^*						
		.15	.30	.60	1.2	2.4	4.8	9.6
10	1	.00	.04	.10	.14	.12	.12	.10
10	2	.01	.09	.24	.31	.20	.18	.15
15	1	.00	.00	.02	.07	.14	.16	.16
15	2	.00	.01	.03	.14	.27	.31	.30
20	1	.00	.01	.03	.07	.14	.17	.19
20	2	.00	.02	.06	.15	.28	.32	.35

*The standard deviation σ_B is the upper bound of uncertainty assigned to each point.

The values of the uncertainty parameter assigned to each point varied between 0 and an upper bound σ_B . Seven different values of σ_B , ranging from .15 to 9.6, were used. The exact values are shown in Table 3. The relative size of these uncertainty values can be appreciated by noting that the coordinates of the configurations were standardized to have a standard deviation of one on each of the two axes.

Figure 7 shows the configurations recovered by KYST for each one of the seven simulations. The upper left-hand panel in this figure shows what the true configuration looks like. The points in each panel are labelled 1 through 9 followed by A. This labelling reflects the order of magnitude of the uncertainty parameters associated with each point. Point 1 has the lowest value, while point A has the highest. The minimum stress values obtained by KYST in each simulation is shown in Table 3.

From the seven recovered configurations, shown in Figure 7, it is quite evident that the KYST solution degenerates considerably as the level of σ_B increases. When the level of σ_B is quite high, the recovered configuration bears no resemblance to the true configuration. What actually happens is that the higher numbered points, those assigned higher uncertainty values, gradually move to the outside of the recovered configuration, while the remaining points gradually move toward the center. (The delta symbol in Figure 7 designates a multiple point, in this case points 1 and 2.)

These results show in a more extreme form what was discernible in the previous two examples. When the uncertainty value of a point is large, its location in the configuration recovered by a nonmetric algorithm is determined almost entirely by the magnitude of its uncertainty value, not by its actual position in the true configuration. The recovered configuration, therefore, need not bear any resemblance to the actual configuration. And, as shown in this example, the recovered configuration can degenerate considerably if the range of uncertainty values is large enough.

It may be useful to comment briefly on the effects of using (3) to calculate the expected values for this example. As indicated previously, this equation can only be considered accurate when D'_{ij} is large. It might be thought, therefore, that the inaccuracies of (3) would contribute to the degeneracies observed at the higher uncertainty values, when

D'_{ij} is small. Actually, just the opposite is the case. The use of this equation slows down the rate of degeneracy, because it underestimates the expected value when the uncertainty values are large. For example, when $D_{ij} = 1.414$, $\sigma_i = 5$ and $\sigma_j = 10$, (3) gives 13.75 for the expected value, while the value obtained from direct numerical integration equals 14.02. Thus, more accurate estimates of the expected value of the interpoint distances would actually speed up the rate of degeneracy, because points having large uncertainty values would move to the periphery of the space sooner.

The stress values shown in Table 3 illustrate another aspect of the nonmetric solutions. The degree of recovery is not always monotonically related to the stress values obtained. When the configuration has 10 points, the value of stress, as shown in rows one and two of Table 3, initially increases as the level of σ_B increases, but eventually decreases. This same general tendency can be seen in the third and fourth rows of Table 3, for the 15 point configuration, except that the reduction of the stress does not occur until higher levels of σ_B are reached. It appears likely that the stress values for the 20 point configuration, reported in the last two rows of Table 3, would also have exhibited this same nonmonotonicity property had higher values of σ_B been explored. Thus, the stress values of the nonmetric algorithms do not always provide a reliable clue as to the accuracy of the recovered configuration. Degenerate solutions apparently can have lower stress values than solutions which are substantially closer to the true solution.

These results have implications for simulation studies that use the Hefner model to generate "error". If nonmetric algorithms are used to estimate the coordinates of a multidimensional configuration, then the recovered configuration should not be expected to represent accurately the true configuration, even if large sample sizes are used. The correct approach, we believe, is to use an estimation procedure that is appropriate to the model under consideration. In the case of the Hefner model, this means using a procedure that takes into consideration the nonmonotonicity properties of that model. The ML estimation procedure is one such procedure.

It is interesting to note what the Young-Householder [1938] metric solution looks like when applied to the simulated data of this example. The metric results are almost precisely the same as the nonmetric ones. The metric solutions degenerate precisely in the same way as do the nonmetric solutions when the level of uncertainty increases. Furthermore, the eigenvalues obtained do not provide a reliable clue as to the dimensionality of the actual configuration. The larger the uncertainty values, the more the eigenvalues tend to approach each other. Thus, the Young-Householder metric approach does not seem any more suitable for estimating the parameters of the Hefner model than does the nonmetric approach.

8. Conclusion

Simple procedures have been described for obtaining maximum likelihood estimates of the parameters in the Hefner model. These procedures involve various simplifying approximations of the likelihood function and simple expressions for obtaining the starting values of the ML iterations. From the results of the simulations reported, it appears that the ML estimates of the coordinates are reasonably accurate, even for the incomplete, unreplicated set of data considered. The estimates of the variances, however, appear to be considerably less accurate. What is needed, evidently, is a systematic study of the small sample properties of the ML estimates.

The question of determining the dimensionality of the stimulus space has not been specifically discussed. Likelihood ratio tests can be used for this. Such tests are quite simple and straight-forward. What is not so simple is the question of power. A systematic study of the power of the test for small sample sizes would be especially desirable.

- Patnaik, P. B. The non-central chi-square and F-distributions and their applications. *Biometrika*, 1949, 36, 202-232.
- Ramsay, J. O. Some statistical considerations in multidimensional scaling. *Psychometrika*, 1969, 34, 167-182.
- Ramsay, J. O. Maximum likelihood estimation in multidimensional scaling. *Psychometrika*, 1977, 42, 241-266.
- Richardson, M. W. Multidimensional psychophysics. *Psychological Bulletin*, 1938, 35, 659-660.
- Sankaran, M. On the non-central chi-square distribution. *Biometrika*, 1959, 46, 235-237.
- Schönemann, P. H. On metric multidimensional unfolding. *Psychometrika*, 1970, 35, 349-367.
- Sherman, C. R. Nonmetric multidimensional scaling: A Monte Carlo study of the basic parameters. *Psychometrika*, 1972, 37, 323-355.
- Spence, I. & Domoney, D. W. Single subject incomplete designs for nonmetric multidimensional scaling. *Psychometrika*, 1974, 39, 469-490.
- Suppes, P. & Zinnes, J. L. Basic measurement theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.) *Handbook of Mathematical Psychology*, Vol. I. New York: Wiley, 1963, Pp. 1-76.
- Thurstone, L. L. A law of comparative judgment. *Psychological Review*, 1927, 34, 273-286.
- Young, F. W. Nonmetric multidimensional scaling: recovery of metric information. *Psychometrika*, 1970, 35, 455-473.
- Young, F. W. & Cliff, N. Interactive scaling with individual subjects. *Psychometrika*, 1972, 37, 385-415.
- Young, G. & Householder, A. S. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 1938, 3, 19-21.
- Zinnes, J. L. & Griggs, R. A. Probabilistic multidimensional unfolding analysis. *Psychometrika*, 1974, 39, 327-350.
- Zinnes, J. L. & MacKay, D. B. Multidimensional scaling models: the other side. In I. Borg (Ed.) *Multidimensional Data Representations: When and Why*. Ann Arbor, Mich.: Mathesis Press, 1981, 517-42.
- Zinnes, J. L. & Wolff, R. P. Single and multidimensional same-different judgments. *Journal of Mathematical Psychology*, 1977, 16, 30-50.

Manuscript received 10/15/79

First revision received 3/20/80

Second revision received 9/9/81

Final version received 6/4/82