PSYCHOMETRIKA—VOL. 46, NO. 4.
DECEMBER, 1981

# QUANTITATIVE ANALYSIS OF QUALITATIVE DATA

FORREST W. YOUNG

L. L. THURSTONE PSYCHOMETRIC LABORATORY
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

This paper presents an overview of an approach to the quantitative analysis of qualitative data with theoretical and methodological explanations of the two cornerstones of the approach, Alternating Least Squares and Optimal Scaling. Using these two principles, my colleagues and I have extended a variety of analysis procedures originally proposed for quantitative (interval or ratio) data to qualitative (nominal or ordinal) data, including additivity analysis and analysis of variance; multiple and canonical regression; principal components; common factor and three mode factor analysis; and multidimensional scaling. The approach has two advantages: (a) If a least squares procedure is known for analyzing quantitative data, it can be extended to qualitative data; and (b) the resulting algorithm will be convergent. Three completely worked through examples of the additivity analysis procedure and the steps involved in the regression procedures are presented.

Key words: exploratory data analysis, descriptive data analysis, multivariate data analysis, nonmetric data analysis, alternating least squares, scaling, data theory.

Perhaps one of the main impediments to rapid progress in the development of the social, behavioral and biological sciences is the omnipresence of qualitative data. All too often it is simply impossible to obtain numerical data; the researcher has the choice of qualitative data or no data at all. Many times it is only possible to determine the category in which a particular datum falls. The sociologist, for example, obtains categorical information about the religious affiliation of her respondents; the botanist obtains categorical information about the family to which his plants belong; and the psychologist obtains categorical information about the psychosis of her patient. Even in the best of circumstances it is often impossible to obtain anything beyond the order in which the data categories fall. When our sociologist observes the amount of education of the respondents in her sample she knows that the observational categories are ordered, but she is unable to assign precise numerical values to the categories. When the psychologist obtains rating scale judgments, the judgments may reasonably be viewed as ordinal, but not always as numerical.

Given the ubiquity of qualitative data, one can understand the long and persistent interest in its quantification. If one could somehow develop a method for assigning "good" numerical values to the data categories, then the data would be quantified and would be susceptible to more meaningful analysis. Curiosity about the topic is nascent in the classical work by Yule [1910], and methods for quantification first began to appear around 1940. Probably the first widely disseminated procedure was Fisher's "appropriate scoring" technique [Fisher, 1938, pp. 285–298] which was introduced at about the same time as a method proposed by Guttman [1941]. Several authors worked on the problem in the early 50's [Burt, 1950, 1953; Hayashi, 1950; Guttman, 1953] with this work being summarized

Presented as the Presidential Address to the Psychometric Society's Annual meeting, May, 1981. I wish to express my deep appreciation to Jan de Leeuw and Yoshio Takane. Our "team effort" was essential for the developments reported in this paper. Without this effort the present paper would not exist. Portions of this paper appear in Lantermann, E. D. & Feger, H. (Eds.) *Similarity and Choice*, Hans Huber, Vienna, 1980. The present paper benefits greatly from a set of detailed comments made by Joseph Kruskal on the earlier paper.

Requests for reprints should be sent to Forrest Young, Psychometric Lab, UNC, Davie Hall, 013-A, Chapel Hill, N.C. 27514.

0033-3123/81/1200-4011$00.75/0

by Torgerson [1958, pp. 338–345]. Much work has occurred recently [Benzicri, 1973, 1977; de Leeuw, 1973; Mardia, Kent & Bibby, 1979; Nishisato, 1980; Saito, 1973; Saporta, 1975; Tenenhaus, Note 1 & Note 2].

In this paper we refer to the process of quantifying qualitative data as "optimal scaling," a term first introduced by Bock [1960]. This is our definition:

> Optimal scaling is a data analysis technique which assigns numerical values to observation categories in a way which maximizes the relation between the observations and the data analysis model while respecting the measurement character of the data.

Note that this is a very general definition: There is no precise specification of the nature of the model, nor is there precise specification of the measurement character of the data. Working with this definition of optimal scaling, we have developed a group of programs for quantifying qualitative data (see Table 1). The programs permit the data to have a variety of measurement characteristics, and permit data analysis with a variety of models. We refer to these programs as ALSOS programs since they use the Alternating Least Squares (ALS) approach to Optimal Scaling (OS).

The ALSOS programs describe qualitative data by quantitative models falling into three general classes: (a) The General Linear Model; (b) The Component (Factor) model; and (c) The General Euclidean Model. As you can see in Table 1, the GLM programs are specifically oriented towards analysis of variance (MANOVALS, ADDALS and WADDALS), regression analysis (MORALS, CORALS, CANALS, OVERALS), discriminant analysis (CRIMINALS) and path analysis (PATHALS). The component programs perform principal components analysis (PRINCIPALS and HOMALS); three-mode component (factor) analysis (ALSCOMP and TUCKALS); and common-factor analysis (FACTALS). The General Euclidean model is fit by ALSCAL and GEMSCAL.

For most of the ALSOS programs the data may be defined at the binary, nominal, ordinal or interval levels of measurement (and the ratio level with the General Euclidean model programs), and may be thought of as having been generated by either a discrete or continuous underlying process. All of these programs also permit arbitrary patterns of missing data. Some permit boundary or range restrictions on the values assigned to the observation categories, and some permit the use of partial orders with ordinal data. Information on obtaining these programs may be obtained as indicated on Table 1.

As we will show in this paper, the ALSOS approach to algorithm construction has one very important implication for data analysis:

> If a procedure is known for obtaining a least squares description of numerical (interval or ratio measurement level) data then an ALSOS algorithm can be constructed to obtain a least squares description of qualitative data (having a variety of measurement characteristics).

### *1. Overview*

Each of the ALSOS programs optimizes an objective loss function by using an algorithm based on the alternating least squares (ALS) and optimal scaling (OS) principles.

The OS principle involves viewing observations as categorical, and then representing each observation category by a parameter. This parameter is subject to constraints implied by the measurement characteristics of the variable (e.g., order constraints for ordinal variables).

The ALS principle involves dividing all of the parameters into two mutually exclusive and exhaustive subsets: (a) the parameters of the model; and (b) the parameters of the data (called optimal scaling parameters). We then proceed to optimize a loss function by alternately optimizing with respect to one subset, then the other (see Figure 1). Note that each subset may itself consist of several subsets which are mutually exclusive and exhaustive. For

Table 1

ALSOS Programs

| Program | Analysis | Data | Source | Primary Reference |
|---|---|---|---|---|
| ADDALS | Additivity analysis (analysis of variance) | Two or three way tables. Nonorthogonal and incomplete designs permitted. | UNC | de Leeuw, Young & Takane (1976) |
| WADDALS | Weighted additivity analysis | Same as ADDALS | UNC | Takane, Young & de Leeuw (1980) |
| MANOVALS | Multivariate analysis of variance | Multi-way tables | RUL | Gifi (1981) |
| MORALS CORALS & CANALS | Multiple and canonical analysis | Mixed measurement level multivariate data | UNC or RUL | Young, de Leeuw & Takane (1976) |
| OVERALS | Canonical analysis | Multiple set mixed measurement level multivariate data | RUL | Gifi (1981) |
| CRIMINALS | Multiple group discriminant analysis | Mixed measurement level predictors | RUL | Gifi (1981) |
| PATHALS | Path analysis | Mixed measurement level multivariate data | RUL | Gifi (1981) |
| PRINCALS & PRINCIPALS | Principal components analysis | Mixed measurement level multivariate data | UNC or RUL | Young, Takane & de Leeuw (1978) |
| HOMALS | Principal components analysis | Multivariate nominal data | RUL | de Leeuw & van Rijkevorsel (1976) |
| ALSCOMP & TUCKALS | Three-mode factor analysis | Mixed measurement level multivariate data | UNC or RUL | Sands & Young 1978 de Leeuw & van Rijkevorsel (1976) |
| FACTALS | Common-factor analysis | Mixed measurement level multivariate data | UNC | Takane, Young & de Leeuw (1978) |
| ALSCAL | Two or three-way multidimensional scaling | Similarity data | UNC | Takane, Young & de Leeuw (1977) |
| GEMSCAL | Two or three-way multidimensional scaling | Similarity data | UNC | Young, Null & De Soete (Note 5) |

Note: The column headed "Source" in Table 1 indicates the address from which the program is available, as follows: UNC: Forrest W. Young, Psychometric Laboratory, Davie Hall 013-A, University of North Carolina, Chapel Hill, NC 27514, U.S.A.; and RUL: Jan de Leeuw, Data Theory, Rijksuniversiteit te Leiden, Breestraat 70, 2311 CS Leiden, The Netherlands.

example, in ALSCAL the model has several parameter subsets, and in the multivariate programs there is a subset of data parameters for each variable.

The optimization proceeds by obtaining the least squares estimates of the parameters in one subset while assuming that the parameters in all other subsets are constants. We call this a conditional least squares estimate, since the least squares nature is conditional on the values of the parameters in the other subsets. Once we have obtained conditional least squares estimates we immediately replace the old estimates of these parameters by the new

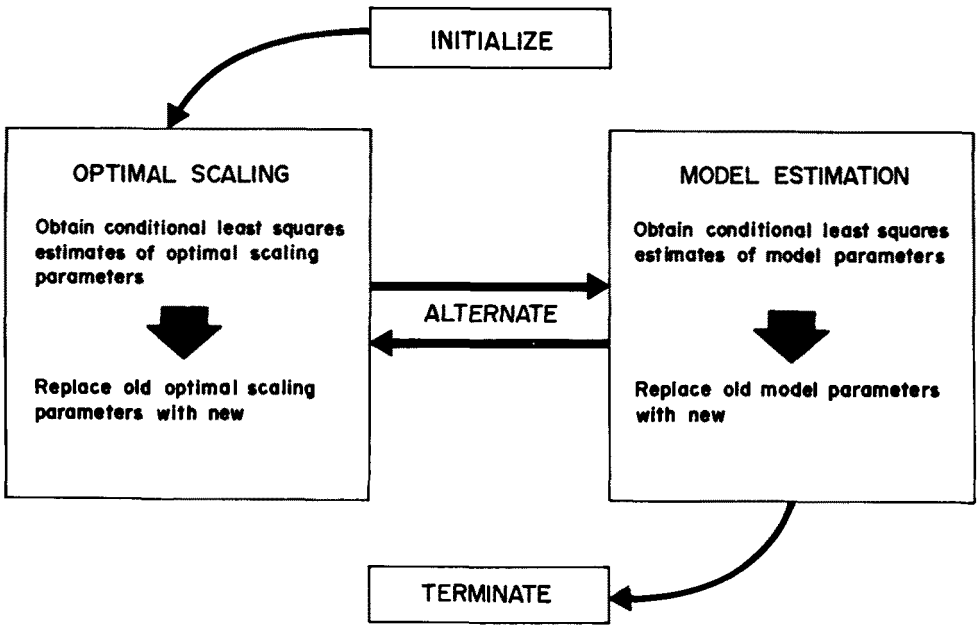


FIGURE 1
Flow of the ALSOS algorithms

estimates. We then switch to another subset of parameters and obtain their conditional least squares estimates. We alternately obtain conditional least squares estimates of the parameters in the model subsets, then in the data subsets, until convergence (which is assured under certain conditions discussed in later portions of this paper) is closely approached. The flow of an ALSOS procedure is diagrammed in Figure 1.

Certain strong correspondences exist between an ALSOS procedure and the NILES approach to algorithm construction developed by Wold and Lyttkens [1969], and the class of numerical analysis algorithms known as successive block algorithms [Hageman & Porsching, 1975]. The main difference between these metric algorithms and the nonmetric ALSOS algorithms is the optimal scaling features of the ALSOS algorithm. The scaling feature permits the analysis of qualitative data, whereas the previous procedures can only analyze quantitative data.

There are also strong connections between the nonmetric algorithms developed by Kruskal [1964, 1965], Roskam [1968], Young [1972], and others. The main difference between these gradient (non-ALS) procedures and ALSOS algorithms is the least squares feature of the model estimation phase.

One of the main advantages of combining the ALS and OS principles is that the OS phase of an ALSOS algorithm does not need to know the type of model involved in the analysis. A parallel and equally important advantage is that the model estimation phase does not need to know anything about the measurement characteristics of the data.

The practical effect of these aspects of ALSOS procedures is enormous: If a least squares procedure exists for fitting a particular model to numerical (i.e, interval or ratio) data, then we can use that procedure in combination with the OS procedures to be discussed to develop an ALSOS algorithm for fitting the model to qualitative data. That's all there is to it! If we can obtain a least squares description of numerical data we can obtain a least squares description of qualitative data. All we have to do is alternate the numerical least squares procedure with the OS procedure which is suited to the measurement characteristics of the data being analyzed.

There is one hooker: The ALSOS procedure does not guarantee convergence on the

globally least squares solution, rather it guarantees convergence on a particular type of local least squares solution. The particular local optimum upon which an ALSOS procedure converges is determined by only one thing, the initialization process. It is possible that two different types of initialization procedures will lead an ALSOS procedure into two different local optima, perhaps giving radically different results. For this reason, and since each phase in an ALSOS procedure is a conditional least squares solution (conditional on the current values of the parameters in the other subset), we refer to the convergence point of an ALSOS procedure as the "conditional global optimum," a somewhat grandiose way of emphasizing that the convergence point is more than simply a local optimum, but may not be the overall global optimum. [The convergence properties of an ALSOS algorithm have been discussed by de Leeuw, Young and Takane (1976) and de Leeuw (Note 3) who prove that such a procedure is indeed convergent if (a) the function being optimized is continuous; and (b) if each phase or subphase of the algorithm optimizes the function.]

Since the initialization procedure is of such importance in the overall process, it is important to employ the best initialization that is available. In all the ALSOS programs we define best initialization to mean that we should optimize the fit of the model to the raw data. Thus, each ALSOS program is initiated by applying a least squares procedure to the raw data under the assumption that the raw data are quantitative, *as the user has coded them.* (Note that a different coding of the data, while still consistent with the data's measurement characteristics, may provide a better start. The start is not "best" in this sense. There is evidence that for ALSCAL, this procedure reduces the frequency of local minimum solutions [Young & Null, 1978].

Once the process is initiated, the procedure for obtaining the conditional least squares estimates of the model parameters is the procedure used to obtain ordinary least squares estimates when the data are numerical. The *only* difference is that the procedure is applied to the optimally scaled data (which is numerical, after all) instead of to the raw data. Since we are applying the model estimation procedure to the optimally scaled data, we are not violating the measurement assumptions of the raw data, whatever they might be. We are not even using the raw data in the model estimation phase, thus we do not need to know its measurement characteristics. Equally important, we do not have to think up a new way of trying to fit the model to qualitative data, we simply use existing procedures for fitting it to quantitative data.

## 2. *Optimal Scaling*

Since the ALS aspects of our work are by now fairly traditional [Wold & Lyttkens, 1969], we do not spend any effort to explicate them. Rather, we fully discuss the OS aspects in the remainder of this paper.

These unique OS aspects of our work permit our ALSOS algorithms to be very flexible in the assumptions the user can make concerning the measurement characteristics of his/her data, as reviewed above.

### *2.1 Measurement Theory*

To appreciate fully the OS aspect of our work we must first discuss the theoretical foundations of our project, our view of measurement theory.

We begin by emphasizing a concept which is crucial to our work: It is our view that all observations are categorical. That is, we view an observation variable as consisting of observations which fall into a variety of categories, such that all observations in a particular category are empirically equivalent. Furthermore, we take this "categorical" view regardless of the variable's measurement characteristics.

Put more simply, it is our view that the observational process delivers observations

which are categorical because of the finite precision of the measurement and observation process, if for no other reason. For example, if one is measuring temperature with an ordinary thermometer (which is likely to generate interval level observations reasonably assumed to reflect a continuous process) it is doubtful whether the degrees are reported with any more precision than whole degrees. Thus, the observation is categorical; there are a very large (indeed infinite) number of different temperatures which would all be reported as say, 40°. Therefore, we say that the observation of 40° is categorical.

At this point we need to define a column vector of $n$ raw observations. We denote this observation vector as $\mathbf{o}$, with general element $o_i$. (Boldface lower case letters refer to column vectors, and italicized lower case letters to scalars.) We also define the model estimates $\hat{\mathbf{z}}$, with general element $\hat{z}_i$, and the optimally scaled observations $\mathbf{z}^*$, with general element $z_i^*$. The elements of $\mathbf{o}$ are organized so that all observations in a particular category are contiguous. The elements of $\hat{\mathbf{z}}$ and $\mathbf{z}^*$ are organized in a fashion having a one to one correspondence with the elements of $\mathbf{o}$. The element $z_i^*$ is the parameter representing the observation $o_i$. The vector $\hat{z}$ is called the "model estimates" because it is the model's estimates, in a least squares sense, of the optimally scaled data $\mathbf{z}^*$.

With these definitions we can formally represent the OS problem as a transformation problem, as follows. We wish to obtain a transformation $\mathcal{t}$ (script letters indicate transformations) of the raw observations which generates the optimally scaled observations,

$$\mathcal{t}[\mathbf{o}] = [\mathbf{z}^*], \tag{1}$$

where the precise definition of $\mathcal{t}$ is a function of the measurement characteristics of the observations, and is such that a least squares relationship will exist between the model's estimates of the scaled data ($\hat{\mathbf{z}}$) and the actual scaled data ($\mathbf{z}^*$), given that the measurement characteristics of $\mathbf{o}$ are strictly maintained. The numerical value assigned to $z_i^*$, then, is the optimal parameter value for the observation $o_i$.

Various types of restrictions are placed on the transformation $\mathcal{t}$, with the type of restriction depending on the measurement characteristics of the data. We distinguish three types of measurement restrictions, termed *measurement level*, *measurement process*, and *measurement conditionality*. As we will see, these three types concern three different aspects of the observation categories. Measurement process concerns the relationships among all of the observations *within* a single category; measurement level concerns the relationships among all of the observations *between* different categories; and measurement conditionality concerns the relationships *within sets* of categories. Each of the several types of processes, levels and conditionalities implies a different set of restraints placed on the transformation $\mathcal{t}$ (1).

In Tables 2, 3, and 4 we summarize the six types of measurement resulting from combining three levels with two processes. A verbal description is given in Table 2, the mathematical restrictions on $\mathcal{t}$ are given in Table 3, and the optimal scaling methods are given in Table 4. Measurement conditionality is discussed at the end of this section.

*Measurement process.* There are two types of measurement process restrictions, one invoked when we assume that the generating process is discrete, and the other when we assume that it is continuous. One or the other assumption must always be made. If we believe that the process is *discrete* (sex is an example of a discrete underlying process) then all observations in a particular category (female or male) should be represented by the same real number after the transformation $\mathcal{t}^d$ (the superscript indicates discreteness) has been made. On the other hand, if we adopt the *continuous* assumption (as we probably should for a weight variable), then each of the observations within a particular category (97.2 Kg., for example) should be represented by a real number selected from a closed interval of real numbers. In the former case the discrete nature of the process is reflected by the fact that we

Table 2

Measurement characteristics
for six types of measurement

| Level | Process | |
|---|---|---|
| | Discrete | Continuous |
| Nominal | Observation categories represented by a single real number | Observation categories represented by a closed interval of real numbers |
| Ordinal | Observation categories are ordered and tied observations remain tied | Observation categories are ordered but tied observations become untied |
| Numerical | Observation categories are functionally related and all observations are precise | Observation categories are functionally related but all observations are imprecise |

choose a single (discrete) number to represent all observations in the category; whereas in the latter case the continuity of the process is reflected by the fact that we choose real numbers from a closed (continuous) interval of real numbers. Formally, we define the two restrictions as follows: The discrete restriction is

$$t^d: (o_i \sim o_m) \rightarrow (z_i^* = z_m^*) \tag{2}$$

Table 3

Measurement restrictions
for six types of measurement

| Level | Process | | | |
|---|---|---|---|---|
| | | Discrete | | Continuous |
| Nominal | $t^d$: | $(o_i \sim o_m) \rightarrow (z_i^* = z_m^*)$ | $t^c$: | $(o_i \sim o_m) \rightarrow (z_i^- = z_m^-) < \left\{ \begin{matrix} z_i^* \\ z_m^* \end{matrix} \right\} < (z_i^+ = z_m^+)$ |
| Ordinal | $t^{do}$: | $(o_i \sim o_m) \rightarrow (z_i^* = z_m^*)$<br>$(o_i \prec o_m) \rightarrow (z_i^* < z_m^*)$ | $t^{co}$: | $(o_i \sim o_m) \rightarrow (z_i^- = z_m^-) < \left\{ \begin{matrix} z_i^* \\ z_m^* \end{matrix} \right\} < (z_i^+ = z_m^+)$<br>$(o_i \prec o_m) \rightarrow (z_i^* < z_m^*)$ |
| Numerical | $t^{dp}$: | $(o_i \sim o_m) \rightarrow (z_i^* = z_m^*)$<br>$z_i^* = \sum_{q=0}^{p} \delta_q o_i^q$ | $t^{cp}$: | $(o_i \sim o_m) \rightarrow (z_i^- = z_m^-) < \left\{ \begin{matrix} z_i^* \\ z_m^* \end{matrix} \right\} < (z_i^+ = z_m^+)$<br>$z_i^* = \sum_{q=0}^{p} \delta_q o_i^q$ |

Table 4

Optimal scaling methods for six types of measurement

| Level | Process | |
|---|---|---|
| | Discrete | Continuous |
| Nominal | Means of model elements | Means of model estimates, followed by primary monotonic transformation |
| Ordinal | Kruskal's secondary monotonic transformations | Kruskal's primary monotonic transformations |
| Numerical | Simple linear (or nonlinear) regression | Simple linear (or nonlinear) regression followed by boundary estimation |

where $\sim$ indicates empirical equivalence (i.e., membership in the same category). The continuous restriction is represented as

$$\ell^c: (o_i \sim o_m) \rightarrow (z_i^- = z_m^-) \leq \begin{Bmatrix} z_i^* \\ z_m^* \end{Bmatrix} \leq (z_i^+ = z_m^+) \tag{3}$$

where $z_i^-$ and $z_i^+$ are the lower and upper bounds of the interval of real numbers. Note that one of the implications of empirical (categorical) equivalence is that the upper and lower boundaries of all observations in a particular category are the same for all the observations. Thus, the boundaries are more correctly thought of as applying to the categories rather than the observations, but to denote this would involve a somewhat more complicated notational system. Note also that for all observations in a particular category the corresponding optimally scaled observations are required to fall in the interval but need not be equal.

*Measurement level.* We now turn to the second set of restraints on the several measurement transformations, the level restraints. With these restraints we determine the nature of the allowable transformations $\ell$ so that they correspond to the assumed level of measurement of the observation variables. There are, of course, a variety of different restraints which might be of interest, but we only mention three here. With these three we can satisfy the characteristics of Stevens' four measurement levels, as well as the measurement level characteristics of missing data, and of binary data.

For the *nominal* level of measurement, and for data that is either missing or are binary, we introduce no measurement-level restraints. The characteristics of these three types of measurement are completely specified by the measurement process restraints. The reason that we need no additional restraints is that for these levels we only know the category of an observation. We know nothing about the relationships of observations in different categories. Thus, these levels are completely specified by restrictions imposed within observation categories, there being no restrictions on the relationships which may exist among observations in different categories.

The difference between nominal, binary and missing data is in the number of observation categories. For nominal data there must be at least three observation categories. When there are only two observation categories the data are binary. (Binary data are somewhat anomolous since they may be thought of as being at any level of measurement.

Since the higher levels of measurement all involve additional between-category restrictions, it is most parsimonious to describe binary data as being at the nominal level.)

*Missing* data, on the other hand, can be viewed as a particular type of data about which we know only one thing: It is missing. Thus, we may view missing data as nominal but having only one category of (non)observation, the "missing" category. It would appear that we could simply call "missing" data an additional category of our nominal data. However, this does not suffice when we have data missing from a set of observations defined at some level of measurement other than nominal. When we introduce the notion of measurement conditionality at the end of this section, we will be able to completely clarify the manner in which we view missing data. Until then, we must be satisfied by simply viewing missing data as being defined at the nominal level, and as having only one "observation" category.

It should be mentioned that data consisting entirely of only one observation category are, logically, equivalent to missing data in their measurement level characteristics. That is, they have no measurement level at all. This is true regardless of the supposed measurement level of the data. Thus, to define the measurement level of a set of observations, the absolute minimum number of observation categories is two, and there must be at least three for any level of measurement above the nominal level.

For the nominal measurement level, and for binary and missing data, there are no level restraints: The characteristics of these data are completely specified by the process restraints. Since there are two types of processes, there are two types of nominal, binary and missing observations. This discrete-nominal level is quite common, with the sex of a person being such a variable. It is clear that this is a nominal (binary) variable, and it is reasonable to assume that the two observation categories (male and female) are generated by a discrete underlying process. An example of a continuous-nominal measurement variable is that of color words. The various observation categories may be blue, red, yellow, green, etc., which, while nominal, actually represent a continuous underlying process (wave length). Even missing data comes in two varieties: Discrete-missing data implies that the observer believes all the observations would have been identical had they been observed; whereas continuous-missing data implies that they wouldn't have been identical. More will be said on this at the end of this section.

For *ordinal* variables, we require, in addition to the process restraints, that the real numbers assigned to observations in different categories represent the order of the empirical observations:

$$\ell^o\colon (o_i \prec o_m) \rightarrow (z_i^* \leq z_m^*) \tag{4}$$

where the superscript on $\ell^o$ indicates the order restriction, and where $\prec$ indicates empirical order. The problem of what to do about ties has already been handled by the process notion. If the variable is discrete-ordinal ($\ell^{do}$), then tied observations remain tied after transformation, whereas, for continuous-ordinal ($\ell^{co}$) variables, tied observations may be untied after transformation. The discrete-ordinal case is well exemplified by data obtained from subjects who order $n - 1$ kinship terms according to their similarity to the $n$'th term. A continuous-ordinal variable might be the income level of one's father, as it is usually obtained in survey data. The observation categories might be "less than \$5,000," "\$5,000–10,000," "\$10,000–20,000," and "more than \$20,000," and one can imagine the continuous process by which such ordered categories are produced.

For *numerical* (interval or ratio) variables we require that the real numbers assigned to the observations be functionally related to the observations. For example (other examples are easily constructed) we might require that the optimally scaled and raw observations be

related by some polynomial rule:

$$\mathcal{I}^p\colon z_i^* = \sum_{q=0}^{p} \delta_q o_i^q. \tag{5}$$

If $p = 2$, for example, we have a quadratic relationship between the optimally scaled and raw observations. When $p = 1$ we obtain the familiar linear relationships used with interval level variables (and with ratio level variables when $\delta_0 = 0$).

It is important to note that with numerical variables the role played by the discrete-continuous distinction is that of measurement precision. If we think that our observations are perfectly precise, then we wish that all observations should be related to the optimally scaled observations by exactly the function specified by (5). However, if we think that there is some lack of precision in the measurement situation, then we may wish to let the optimally scaled observations "wobble" around the function specified by (5) just a bit. The former case corresponds to the discrete-interval or discrete-ratio case in which we allow no within observation observation category variation, and the latter case corresponds to the continuous-interval or continuous-ratio case in which we do permit some within category variation. Note that this notion is sensible even when there is only one observation in a particular observation category, as is usually the case.

Let us re-emphasize that even though the data are viewed as categorical, it is just as possible to obtain a categorical datum which is measured at the interval level of measurement but which was generated by a discrete process, as it is possible to obtain a categorical datum which is measured at the nominal level of measurement but which was generated by a continuous process. There is no necessary relationship between the presumed underlying generating process and the level of measurement, and in any case the datum is categorical.

*Measurement conditionality.* The final type of restraints placed on the measurement transformations $\mathcal{I}$ are referred to as conditionality restraints. These restraints operate on the relationships which may exist among observations *within sets* of observation categories. As has been emphasized by Coombs [1964], it may be, for a particular set of data, that the measurement characteristics of the observations are conditional on some aspect of the empirical situation. When this is the case it follows that some of the observations cannot be meaningfully compared with other observations. Thus, we should subdivide all of the observations into groups such that those observations within a group are those which can be meaningfully compared to each other. Then we must restrict the measurement level and process transformations so that they only apply to observations within a group, not between groups. We call such groups *partitions*, since they partition the data into subsets, and we redefine the restrictions given by (2) through (5) so that they only enforce the desired relationships to exist among the observations within a partition. There are no restrictions enforced on the relationships which may exist among observations which are in different partitions.

There are several types of conditionality which can be distinguished, some of which are relevant to certain kinds of data analysis, others to other kinds. We do not go into them in detail here, but choose to mention only two.

One type, *matrix conditional*, is found in the following example. If we have asked several subjects to judge the similarity of all pairs of a set of stimuli, then we usually are unwilling to compare one subject's responses with another. That is, one subject's response of 7 (on, say, a similarity scale of 1 through 9) cannot be said to represent more similarity than another subject's response of 6. Furthermore, we can't say that one subject's response category of 6 means the same as another subject's category of 6. We just are not sure that the several subjects are using the response scale in identical ways. In fact, we are pretty sure

that they do not use the scale identically. Thus, we say that the "meaning" of the measurements are conditional on which subject (matrix) is responding. Thus, we call this type of data matrix-conditional. Furthermore, for matrix-conditional data each matrix is a partition of the data.

Another common type of conditionality, *column conditional*, is exemplified by multivariate data. In multivariate data each column of data represents a measurement variable (such as the sex of a person, the person's socio-economic level, the person's height and weight, her IQ score, etc.). The important notion is that multivariate data are (essentially always) column conditional.

This example is a nice example of the fact that the measurement characteristics of one partition do not have to correspond to those of another partition. One of the variables is sex, which is binary; another is socio-economic level, which is probably continuous-ordinal; a third is height and a fourth weight, which are both ratio and may be reasonably thought of as discrete; and a fifth was IQ score, which may be either ordinal or interval.

Formally, we state that the domain of the measurement transformation $\ell$ is dependent on the type of conditionality. For matrix-conditional data the domain is a single matrix of data and the transformation is denoted $\ell_k$ to indicate that there is a separate transformation for each matrix $k$. For column-conditional data the domain is a single column of a single matrix, and the transformation is denoted $\ell_{jk}$. The previous discussion of measurement level and process was implicitly in terms of unconditional data. While all of the definitions of level and process must be modified appropriately, we do not explicate these modifications as they are both lengthy and obvious.

*Missing data.* We can now fully explicate our missing data notion. We have already stated that missing data can be viewed as being defined at the nominal level, and as having only one "observation" category (that of nonobservation). We noted, however, that we needed one more concept, that of conditionality, to fully explain our missing data notion. Thus, the full idea of just what missing data is can now be stated. We view missing data as observation cells which form their own separate partitions, called, naturally enough, the missing data partitions. All of the missing "observations" in a particular missing data partition fall in one observation category, the "missing" category.

Since the missing data partition has only one category of observation, and since the notion of measurement level refers to restraints between categories, none of the measurement level restraints [Eqs. (4) and (5)] apply. Thus, there is no measurement level for missing data. However, since the notion of measurement process refers to restraints within observation categories, the notion of the measurement process *does* apply to missing data. While this may sound a bit strange, it in fact corresponds to a common concern that a researcher has when faced with what to do about several missing observations. He wonders what it is that caused the missing observations. One of the possibilities is to assume that every missing observation was caused by a common underlying process, whereas another possibility is to assume that the missing observations were caused by a variety of processes. The former view, which says that a single thing caused the data to be missing, implies that all of the missing observations should be assigned a single number by the optimal scaling. Thus, this view corresponds with what we call discrete missing data. The latter view, that a variety of things contributed to the missing observations, implies that a continuum of numbers ought to be assigned. Thus, this view is what we think of as continuous missing data.

Finally, as implied above when we said that there can be several missing data partitions, we wish to point out that missing data can be conditional. For example, if we have multivariate data and there are data missing on two different variables, we would probably assign the missing observations to two separate partitions, one for each variable.

*2.2 Indicator Matrices*

In the previous section we discussed our measurement theory from the perspective of restraints imposed on data transformations. In this section we discuss our theory from a different perspective, that provided by conceiving of the data as being represented by parameters whose values we wish to estimate

Now it may sound a bit unusual to discuss data parameters. After all, we always associate parameters with models. However, with qualitative data it is useful to think of each observation category as being represented by a parameter whose value we wish to estimate in some optimal way. (Gifi, 1981, explores the possibility of several parameters per category.) The value assigned to each observation category parameter is the "quantification" of that category. After determining the best parameter values we have "optimally scaled" the data.

To restate the goal of optimal scaling in this new light: We wish to estimate values for the data (observation category) parameters so that two characteristics are met: First, the estimation must perfectly satisfy the stated measurement restrictions; and second, it must yield a least squares relationship to the model, given that the measurement restrictions are perfectly satisfied, and given certain normalization considerations.

To discuss optimal scaling from the viewpoint of estimating data parameters we must introduce one new notion, called the *indicator matrix*. This matrix represents the data of a specified partition in a way which indicates the category in which each observation resides. There is an indicator matrix for each partition.

For now we define the indicator matrix $U_p$ as an $(n \times n_c)$ binary matrix with a row for each of the $n$ observations in partition $p$, and a column for each of the $n_c$ categories. (This definition will be generalized below.) The elements of $U_p$ indicate category membership:

$$u_{\text{pic}} = \begin{cases} 1 & \text{iff } o_i \in \text{category } c \\ 0 & \text{otherwise} \end{cases}$$

In the remainder of this section we drop the $p$ from $U_p$, but it is to be understood that we are discussing only the data for a specific partition, and that the discussion applies to any and all partitions with no loss of generality.

*Nominal data.* With the definition of $U$ given above we can look at the $\mathscr{E}^{dn}$ (discrete-nominal) transformation as a very simple parameter estimation process. In fact, it is Fisher's optimal scoring technique [Fisher, 1938, pp. 285–298] which consists of estimating the value of the optimally scaled datum $z_i^*$ as the mean of all the model estimates $\hat{z}_j$ which correspond to those observations $o_j$ that are in the same category as $o_i$. Since the $z_i^*$ are the mean of the appropriate $\hat{z}_j$, we minimize the residuals $\| \hat{\mathbf{z}} - \mathbf{z}^* \|$ under the $t^{dn}$ measurement restrictions. However, the index we wish to minimize is a *normalized* residuals index, Kruskal's [1964] Stress index:

$$S = \left[ \frac{||\hat{\mathbf{z}} - \mathbf{z}^*||}{||\mathbf{z}^*||} \right]^{1/2}. \tag{6}$$

Because $\mathbf{z}^*$ appears in both the denominator and numerator, $S$ is not minimized by the values which minimize its numerator. A normalization must be made of the $\mathbf{z}^*$ to minimize (6), as discussed in section 2.4. To emphasize this aspect, we introduce the *unnormalized* scaled data, denoted $\mathbf{z}^u$, which minimize the *unnormalized* residuals $\|\hat{\mathbf{z}} - \mathbf{z}^u\|$, and we reserve $\mathbf{z}^*$ for the normalized scaled data which minimize (6). Formally, $\mathbf{z}^u$ is defined as

$$\mathscr{E}^{dn}: \mathbf{z}^u = U(U'U)^{-1}U'\hat{\mathbf{z}}, \tag{7}$$

where $U$ is defined above. Note that $U'U$ is a diagonal $(n_c \times n_c)$ matrix with a row and column for each observation category, and with the number of observations in each cat-

egory on the diagonal. Also, $U'\hat{\mathbf{z}}$ is an $n_c$ element column vector with the sum of the $\hat{z}_j$'s as its elements. Finally, $(U'U)^{-1}U'\hat{\mathbf{z}}$ is an $n_c$ element column vector with the mean of the appropriate $\hat{z}_j$'s as its elements. *These* are the unnormalized least squares estimates of each of the $n_c$ data parameters for the partition under consideration.

The continuous-nominal situation is more complex than the discrete-nominal situations. The added complexity is introduced because the continuous-nominal situation, as discussed to this point, involves no measurement restrictions. For $\mathscr{t}^{cn}$ we impose the continuous process restrictions [$t^n$, (3)] that each optimally scaled observation should reside in some interval, and we have placed no restrictions on the formation of the intervals. Thus, we could select arbitrarily large upper and lower boundaries which would permit all optimally scaled observations to be set equal to all raw observations, thus minimizing the squared differences trivially and totally.

Naturally, such a process is meaningless. Therefore, we propose an alternative process which yields nonoverlapping contiguous intervals, thus disallowing the trivial possibilities outlined above.

The estimation procedure for the continuous-nominal transformation $\mathscr{t}^{cn}$ involves the following two-phase process: In the first phase we treat the data as though they are discrete-nominal and perform a complete ALSOS analysis based on this assumption. When this process has terminated we enter the second phase in which we treat the data as though they are continuous-ordinal (see below) and perform a second complete ALSOS analysis. Note that in neither phase do we actually assume that the data are continuous-nominal. However, the assumptions that are used do not violate the continuous-nominal nature of the data. In the first phase we use the categorical information to obtain the least squares quantification of each category. In the second phase the quantification from the first phase is used to define an order for the observation categories, which is then used to define interval boundaries.

Three things should be noted about this two-phase procedure. First, it yields a least squares quantification which is consistent with, but stricter than, the continuous-nominal restrictions discussed above. Specifically, the procedure yields nonoverlapping intervals, whereas the restrictions discussed above would permit overlapping intervals. Second, the procedure outlined here is *not* the same as the pseudo-ordinal procedure discussed by de Leeuw, Young & Takane [1976], but is a newer procedure which avoids the divergence problems mentioned in that paper. Third, the procedure is convergent but not strictly least squares because it may converge on a nonoptimal interval order. The only way to avoid this problem is to try all possible interval orders, a prohibitively expensive process.

*Ordinal data.* The estimation procedures for the ordinal transformations $\mathscr{t}^{do}$ and $\mathscr{t}^{co}$ necessitate extending the indicator matrix definition given above. We still define $U$ as a binary matrix, but it is now an $n \times n_b$ matrix, where $n_b$ is the number of *blocks* required to impose the ordinal restriction. An element of $U$ indicates *block* membership in a fashion parallel to the indication of category membership for nominal level.

For the discrete ordinal situation $n_b$ is never greater than $n_c$ (the number of categories) and $U$ represents a merging of observation categories. Given the proper $U$, Young [1975a] has shown that

$$\mathscr{t}^{do}\colon \mathbf{z}^u = U(U'U)^{-1}U'\hat{\mathbf{z}}. \tag{8}$$

Here $U$ is constructed by Kruskal's [1964] *secondary* least squares montonic transformation, which he proved to be least squares. $U$ indicates which categories must be merged (blocked) to satisfy the ordinal restrictions. Note that $U'U$ is a diagonal $(n_b \times n_b)$ matrix containing the number of observations in each block on its diagonal. Also, $(U'U)^{-1}U'\hat{\mathbf{z}}$ is the $n_b$ element vector of the unnormalized optimal scale values that are the

least squares parameter values which preserve the data's discrete-ordinal measurement characteristics. An example is given in section 3.1.

For the continuous-ordinal situation, $n_b$ may or may not be greater than $n_c$. $U$ indicates which *observations* (not categories) must be merged (blocked) in order to preserve the ordinal restrictions. Given the proper $U$, Young [1975a] has shown that

$$\mathscr{t}^{co}: \mathbf{z}^u = U(U'U)^{-1}U'P\hat{\mathbf{z}}, \tag{9}$$

where $U$ and $P$ are constructed by Kruskal's [1964] *primary* least squares monotonic transformation [see de Leeuw, 1975, for a least squares proof, and de Leeuw, 1977a, for an additional ordinal transformation]. The matrix $P$ is a binary $(n \times n)$ block-diagonal permutation matrix. It has $n_b$ blocks, each of which has an order equal to the corresponding element of $U'U$. Each block is a permutation matrix having a single one in each row and column. $P$ has only zeros outside of the blocks. The matrix $U'U$ is interpreted as before (number of observations in each block), and $(U'U)^{-1}U'P\hat{\mathbf{z}}$ contains the unnormalized least squares observation category parameter estimates.

In section 3.1 we present detailed examples of $U$ for $\mathscr{t}^{dn}$ and for $\mathscr{t}^{do}$, as well as $U$ and $P$ for $\mathscr{t}^{co}$. The examples also present the process by which $U$ (and $P$) are constructed for the ordinal situations. It is *very* important to note that only for $\mathscr{t}^{dn}$ do we know $U$ before the analysis takes place: It is simply the category structure of the data. For the ordinal situations we must determine $U$ (and $P$) so that the ordinal properties of the data are maintained. In these cases $U$ (and $P$) are *not* known prior to the analysis, but must be solved for! They are *variables* to be solved for, whereas $U$ is a *constant* when the data are discrete-nominal.

This is a crucial difference with several implications. One implication is that the solution for $\mathbf{z}^u$ is much slower and more complex for ordinal data. Another implication is that the ability to determine degrees-of-freedom is lost with ordinal data. The latter implication implies in turn that inferential procedures are more difficult to determine for ordinal data, as is well known.

*Missing data.* When we have missing data the empty observation cells are removed from whatever partition they were in and placed in one or more separate partitions called missing data partitions. There is one missing data partition for unconditional missing data and more than one for conditional missing data.

For discrete-missing data all of the missing observations in a partition are thought of as residing in one category. Thus $U$ is a column vector of $n$ one's, where $n$ is the number of missing observations in the partition. Equation (7) is applied to calculate $\mathbf{z}^u$ (the optimally scaled missing data) which, due to the nature of $U$, is a vector whose elements are all the mean of $\hat{\mathbf{z}}$, the model estimates of the missing data.

For continuous-missing data the missing data are coded, in $U$, as though they are each in a separate category. Thus, the number of categories equals the number of missing observations, and $U$ is an $(n \times n)$ identity matrix. Therefore, (7) simplifies to $\mathbf{z}^u = \hat{\mathbf{z}}$, and each missing datum is optimally scaled by setting it equal to its model estimate. Note that this way of treating continuous-missing data is not in keeping with the discussion at the end of section 2.1. However, it is mathematically equivalent and simpler to use the present definition of $U$.

*Quantitative data.* While the focus of this paper is on qualitative data, it is worthwhile to spend a moment on the estimation process for quantitative data. With the proper definition of $U$ the $\mathscr{t}^{p}$ transformation can be written, in matrix notation, as

$$\mathscr{t}^{p}: \mathbf{z}^u = U\delta. \tag{10}$$

Here $U$ is a matrix with a row for each observation and with $p + 1$ columns, each column

being an integer power of the vector $\mathbf{o}$ of observations. The first column is the zero'th power (i.e., all ones), the second column is the first power (i.e., is $\mathbf{o}$ itself), the third column is the squares $\mathbf{o}^2$, etc. The unnormalized least squares estimates of $\mathbf{z}^u$ is

$$\mathcal{L}^p\colon \mathbf{z}^u = \underline{U(U'U)^{-1}U'\hat{\mathbf{z}}}. \tag{11}$$

Note that in this case $U$ is, once again, known before the analysis takes place. It is, then, only for the ordinal cases that $U$ is unknown prior to the analysis.

*2.3 Conic Projection*

It is important to note that for all of the types of measurement characteristics discussed here, the corresponding transformation $\mathcal{L}$ may be viewed, for each partition, as though we are regressing the model estimates $\hat{\mathbf{z}}$ onto the raw observation $\mathbf{o}$ in an unnormalized least squares sense and under the appropriate measurement restrictions. In particular, each $\mathcal{L}$ can be represented by a projection operator of the form

$$E = U(U'U)^{-1}U' \tag{12}$$

where the particular definition of $U$ depends on the measurement characteristics, as noted above. This means that we can make the important point that

$$\mathbf{z}^u = E\hat{\mathbf{z}}. \tag{13}$$

When we formally note that the least squares notion is defined as

$$\phi^2 = \| \mathbf{z}^u - \hat{\mathbf{z}} \| = (\mathbf{z}^u - \hat{\mathbf{z}})'(\mathbf{z}^u - \hat{\mathbf{z}}) \tag{14}$$

and when we define $F = I - E$, then we see that

$$\phi^2 = \hat{\mathbf{z}}'F\hat{\mathbf{z}} \tag{15}$$

emphasizing the fact that each of the transformations can be viewed as optimizing the vector product of the model estimates and some linear combination of the very same model estimates, where the linear combination is determined by the measurement restrictions. This point has been emphasized in a more restricted situation by Young [1975a], and was first noted in the present context by Young, de Leeuw, & Takane [1976].

Geometrically, the projection operator $E$ projects the model estimates $\hat{\mathbf{z}}$ onto the nearest surface of a data cone $\mathbf{o}$. The projection is the unnormalized optimally scaled data $\mathbf{z}^u$.

Speaking geometrically, the model's estimates, the optimally scaled data, and the raw data can each be seen as subspaces of a space whose dimensionality is very high. We can also picture the model's parameters as existing in a parameter space.

Figure 2 presents the geometric relations among the model, data and optimal scaling subspaces, as well as the parameter space. Note that the model, data and optimal scaling subspaces are subspaces of a single "*problem*" space of dimensionality $n$, with each observation represented by a dimension of the space. We refer to this space as the "problem" space because it is in this space that we characterize and solve the data analysis problem under consideration. Note that the problem space is a space of real numbers, and that the space has a dimension for all observations in all partitions including missing data partitions (if there are any).

We emphasize that the parameter space is *not* a subspace of the problem space. The parameter space is of dimensionality $p$, one dimension for each of the $p$ parameters. Usually $p$ is much less than $n$, the reduction in dimensionality representing the parsimony of the model's description of the data. We also emphasize that the model, scaling and data subspaces have fewer than $n$ dimensions, but are subspaces of the problem space. Fur-

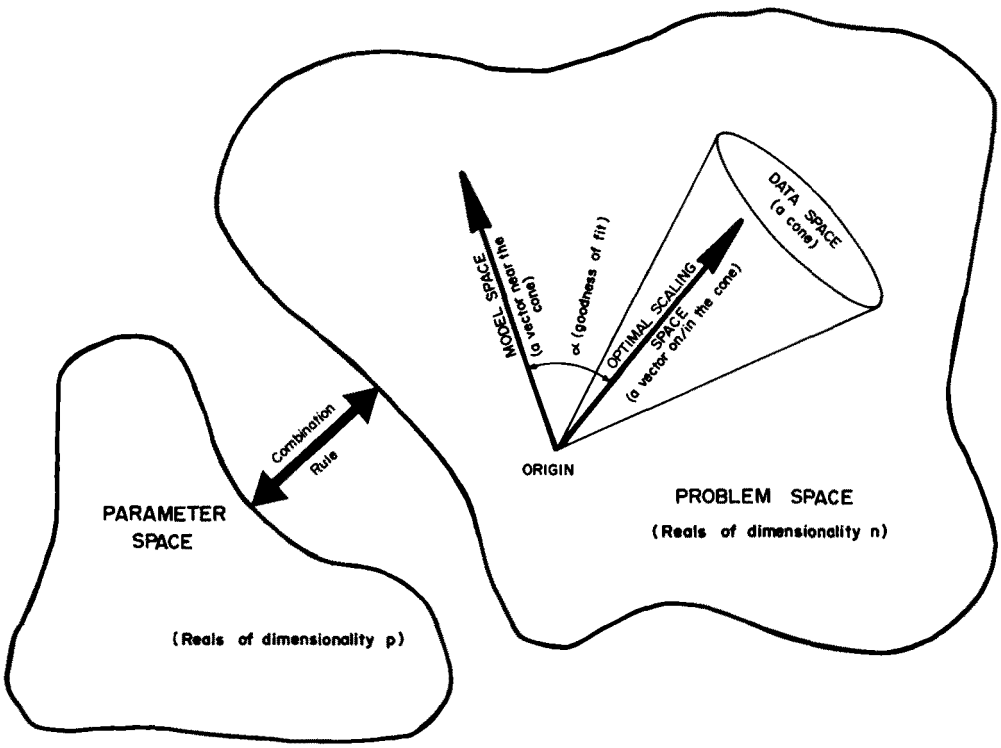


FIGURE 2
Geometrical foundations of the ALSOS algorithms with emphasis on conic projection

thermore, each partition is represented by its own unique model, scaling, and data subspace, so when there are $m$ partitions the problem space contains $m$ model subspaces, $m$ optimal scaling subspaces, and $m$ data subspaces. In Figure 2 we only display the geometry for one partition for simplicity and without loss of generality.

In the problem space we have, geometrically, represented the model estimates and the optimally scaled data subspaces as vectors and the raw data subspace as a cone. Furthermore, the two vectors and cone all intersect at the origin of the problem space. We have chosen the type of representation for each of the three subspaces for specific reasons. We represent the optimal scaling subspace as a geometric vector running through the origin to emphasize the fact that the elements of the algebraic vector $\mathbf{z}^u$ define a point in the problem space, and that, if we form the geometric vector which connects that point to the origin of the problem space, then all of the other points on the geometric vector are equivalent to $\mathbf{z}^u$ at the ratio level of measurement. In terms of the restrictions discussed above, any point in the optimal scaling subspace in Figure 2 is equivalent to any other point. (The normalization restrictions will select a specific point $\mathbf{z}^*$ on the optimal scaling vector, as discussed in section 2.4.) We represent the model subspace as a geometric vector for the same type of reasons.

On the other hand, we represent the data subspace as a geometric cone, not a geometric vector. Although the representation is different, the reasoning underlying the representation is the same: For the data subspace a cone properly represents the measurement characteristics, whereas for the model and optimal scaling subspace a geometric vector is the proper representation. If you reflect on the restrictions given in (2) through (5), you will see they can all be represented geometrically as cones (some restrictions imply certain degenerate cones, for example vectors). This point has been discussed by de Leeuw, Young & Takane [1976] and by de Leeuw [1975; Note 3].

You will note that the optimal scaling vector is represented as being on the surface of the cone. Since the optimal scaling and data subspaces are completely equivalent in terms of the measurement characteristics of the data, the optimal scaling vector must be contained in the data cone. Since the model and optimal scaling subspaces are as nearly alike as possible in a least squares sense, the optimal scaling vector must be "near" the model vector. Thus it is usually the case that the optimal scaling vector is on the surface of the cone, since the surface is the part of the cone which is generally closest to the model subspace. (The only time that the optimal scaling vector is inside the cone is when the model subspace also happens to be in the cone, which only happens when the model perfectly fits the data.)

Finally, note the angle $\alpha$ between the model and optimal scaling vectors. The angle $\alpha$ represents the goodness-of-fit between the two vectors, the fit being measured by (14). The smaller the angle the better the fit. When the angle is zero the fit is perfect (this usually means that the model and optimal scaling vectors are inside the data cone, but it may mean that the two are on the surface of the cone). Note that there is a difficulty associated with a model subspace consisting entirely of zeros. In this case, (14), the fit between the model and optimal scaling vectors, is perfect, and the angle $\alpha$ in Figure 2 is zero. However, the fit is perfect only in a trivial and uninteresting sense. Thus we must ensure that whatever procedures we adopt will not yield a solution at the origin of the problem space. Such solutions are avoided by normalizing the length of the model and optimal scaling vectors to some arbitrary nonzero length.

*2.4 Normalization*

As we just mentioned, a trivial and undesirable way of minimizing (14) is to set the model subspace $\hat{\mathbf{z}}$ equal to zero. Then $\mathbf{z}^u$ is also zero for all transformations, and hence $\phi^2$ is zero for each partition. It is for this reason that the remarks about normalization conditions were made just prior to (6). In this section we discuss the normalization.

Several different normalizations are used in the ALSOS programs. All of the normalizations are introduced to avoid solutions represented by the origin of the problem space (see Figure 2) or other types of trivial solutions. The several normalization conditions have been discussed by Kruskal and Carroll [1969], Sands and Young [1980], Young [1972]; and de Leeuw [Note 3]. Two of these conditions are equivalent to defining either

$$\phi_a^2 = \frac{(\mathbf{z}_a^* - \hat{\mathbf{z}})'(\mathbf{z}_a^* - \hat{\mathbf{z}})}{\hat{\mathbf{z}}'\hat{\mathbf{z}}}, \tag{16}$$

or

$$\phi_b^2 = \frac{(\mathbf{z}_b^* - \hat{\mathbf{z}})'(\mathbf{z}_b^* - \hat{\mathbf{z}})}{\mathbf{z}_b^{*\prime}\mathbf{z}_b^*}, \tag{17}$$

where $\mathbf{z}_a^*$ and $\mathbf{z}_b^*$ are the "normalized" versions of $\mathbf{z}^u$ which optimize $\phi_a^2$ and $\phi_b^2$, respectively.

Now it should be clear that $\mathbf{z}^u$ minimizes (16) since we know from section 2.3 that it minimizes the numerator of (16), and since $\mathbf{z}^u$ is not involved in the denominator of (16). Thus,

$$\mathbf{z}_a^* = \mathbf{z}^u. \tag{18}$$

Also, by the measurement characteristics of $\mathbf{z}^u$, and as pictured in Figure 2,

$$\mathbf{z}_b^* = b\mathbf{z}^u = \mathbf{z}^* \tag{19}$$

where $b$ is a "normalization value" which is to be determined. Notice that we are specifically using $\mathbf{z}^*$ to refer to the normalization of $\mathbf{z}^u$ which minimizes (17).

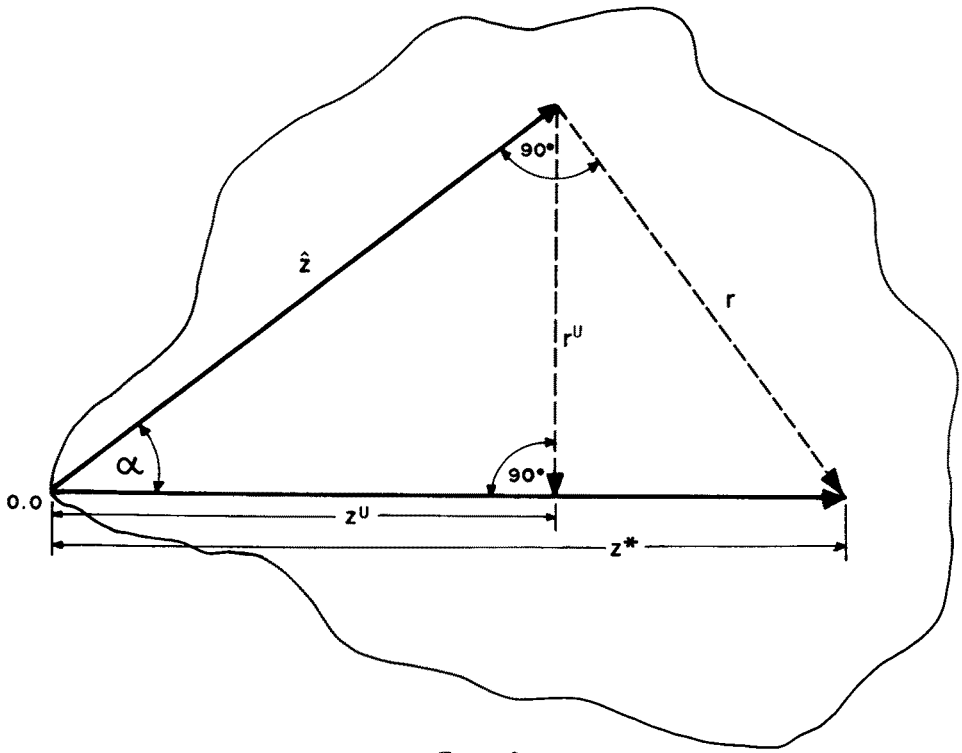


FIGURE 3
Geometrical representation of the normalization aspect of ALSOS algorithms

By looking at Figure 3 we may understand the relationships between $\phi^2$ and $\phi_b^2$, and the relationships between $\mathbf{z}^u$ and $\mathbf{z}^*$. This figure presents, in more detail, a portion of the problem space shown in Figure 2. Specifically, we are looking down at a portion of the *surface* of the data cone, with the surface represented by the irregularly shaped area. Above the cone's surface is shown the model vector $\hat{\mathbf{z}}$. Note that it emanates from the origin of the problem space and data cone, the origin denoted o.o. The orthogonal projection of the model vector onto the surface of the cone gives $\mathbf{z}^u$, the *unnormalized* optimally scaled data. As we saw in section 2.3, this projection is represented by the operator $E$ (12) which minimizes $\phi^2$ (14), the *unnormalized* index of fit. Geometrically, the projection minimizes the angle $\alpha$ between $\hat{\mathbf{Z}}$ and $\mathbf{z}^u$, and thus the length of the vector of residuals $\mathbf{r}^u$, and thus (14) which is simply the square of the length of the residuals vector.

However, $\mathbf{z}^u$ does *not* minimize $\phi_b^2$, even though it minimizes $\phi^2$, as we shall now demonstrate. Recall that $\alpha$, the angle between $\hat{\mathbf{z}}$ and $\mathbf{z}^u$, has been minimized by orthogonally projecting $\hat{\mathbf{z}}$ onto the cone's surface. It is simple to see that the projection defines a right triangle such that

$$\sin^2 \alpha = \frac{\mathbf{r}^{u2}}{\hat{\mathbf{z}}^2} = \frac{\mathbf{r}^{u\prime}\mathbf{r}^u}{\hat{\mathbf{z}}'\hat{\mathbf{z}}} = \phi_a^2, \tag{20}$$

since the orthogonal projection of $\hat{\mathbf{z}}$ *onto* the cone's surface requires a right angle at the surface of the cone (indicated by the lower 90° angle). However, if we project orthogonally *from* $\hat{\mathbf{z}}$ (indicated by the upper 90° angle) onto the surface of the cone in the plane defined by $\hat{\mathbf{z}}$ and $\mathbf{z}^u$, we obtain a projection $\mathbf{z}^*$ and a *new* right triangle such that

$$\sin^2 \alpha = \frac{\mathbf{r}^2}{\mathbf{z}^{*2}} = \frac{\mathbf{r}'\mathbf{r}}{\mathbf{z}^{*\prime}\mathbf{z}^*} = \phi_b^2, \tag{21}$$

and, since

$$\mathbf{r} = \mathbf{z}^* - \hat{\mathbf{z}}, \tag{22}$$

we see that

$$\sin^2 \alpha = \frac{(\mathbf{z}^* - \hat{\mathbf{z}})'(\mathbf{z}^* - \hat{\mathbf{z}})}{\mathbf{z}^{*\prime}\mathbf{z}^*}. \tag{23}$$

Thus the vector $\mathbf{z}^*$ minimizes $\phi_b^2$ (17). Furthermore, when $\mathbf{z}^u$ is used in $\phi_a^2$ and $\mathbf{z}^*$ in $\phi_b^2$, it is the case [from (20) and (23)] that

$$\phi_a^2 = \phi_b^2. \tag{24}$$

Thus, these two apparently different formulas are in fact equal, and it makes no difference which normalization is chosen.

We have not, however, discovered how to obtain $\mathbf{z}^*$ from $\mathbf{z}^u$; that is, we still need to determine the value of $b$ in (19). The value of $b$ is obtained by noting that

$$\cos^2 \alpha = \frac{\mathbf{z}^{u\prime}\mathbf{z}^u}{\hat{\mathbf{z}}'\hat{\mathbf{z}}} \tag{25}$$

and that

$$\cos^2 \alpha = \frac{\hat{\mathbf{z}}'\hat{\mathbf{z}}}{\mathbf{z}^{*\prime}\mathbf{z}^*}, \tag{26}$$

Thus

$$\frac{\hat{\mathbf{z}}'\hat{\mathbf{z}}}{\mathbf{z}^{*\prime}\mathbf{z}^*} = \frac{\mathbf{z}^{u\prime}\mathbf{z}^u}{\hat{\mathbf{z}}'\hat{\mathbf{z}}}, \tag{27}$$

and

$$\mathbf{z}^{*\prime}\mathbf{z}^* = \frac{(\hat{\mathbf{z}}'\hat{\mathbf{z}})(\hat{\mathbf{z}}'\hat{\mathbf{z}})}{(\mathbf{z}^{u\prime}\mathbf{z}^u)}. \tag{28}$$

Noting that the values within parentheses are scalars, we see that

$$\begin{aligned}\mathbf{z}^{*\prime}\mathbf{z}^* &= \frac{(\hat{\mathbf{z}}'\hat{\mathbf{z}})(\mathbf{z}^{u\prime}\mathbf{z}^u)(\hat{\mathbf{z}}'\hat{\mathbf{z}})}{(\mathbf{z}^{u\prime}\mathbf{z}^u)(\mathbf{z}^{u\prime}\mathbf{z}^u)} \\ &= \left[\frac{(\hat{\mathbf{z}}'\hat{\mathbf{z}})}{(\mathbf{z}^{u\prime}\mathbf{z}^u)}\mathbf{z}^{u\prime}\right]\left[\mathbf{z}^u\frac{(\hat{\mathbf{z}}'\hat{\mathbf{z}})}{(\mathbf{z}^{u\prime}\mathbf{z}^u)}\right].\end{aligned} \tag{29}$$

Thus, it follows that

$$\mathbf{z}^* = \left[\mathbf{z}^u\frac{(\hat{\mathbf{z}}'\hat{\mathbf{z}})}{(\mathbf{z}^{u\prime}\mathbf{z}^u)}\right] \tag{30}$$

Therefore, in (19) we see that

$$b = \frac{(\hat{\mathbf{z}}'\hat{\mathbf{z}})}{(\mathbf{z}^{u\prime}\mathbf{z}^u)}. \tag{31}$$

A little study of Figure 3 or (25) will reveal that

$$\begin{aligned}b &= \frac{1}{\cos^2 \alpha} \\ &= \frac{1}{1 - \phi_b^2}\end{aligned} \tag{32}$$

Thus we also note that

$$\phi_a^2 = \phi_b^2 = 1 - \frac{1}{b}. \tag{33}$$

Finally, the orthogonality of $\mathbf{z}^u$ and $\mathbf{r}^u$ allows us to also show that

$$b = \frac{(\hat{\mathbf{z}}'\hat{\mathbf{z}})}{(\hat{\mathbf{z}}'\mathbf{z}^u)}. \tag{34}$$

These relationships among the various expressions for $b$ were first noted by Sands and Young [1980]. The fact that optimizing the unnormalized loss function by a projection operator is simply related to the more difficult problem of optimizing a normalized loss function was first discussed by de Leeuw [1975] and de Leeuw, Young and Takane [1976] and proved by de Leeuw [Note 3].

Both Sands and Young [1980] and de Leeuw [Note 3] also discuss relations between two versions of Kruskal's [1965] second stress formula

$$\phi_c^2 = \frac{(\mathbf{z}_c^* - \hat{\mathbf{z}})'(\mathbf{z}_c^* - \hat{\mathbf{z}})}{(\hat{\mathbf{z}} - \bar{\hat{\mathbf{z}}})'(\hat{\mathbf{z}} - \bar{\hat{\mathbf{z}}})}, \tag{35}$$

and

$$\phi_d^2 = \frac{(\mathbf{z}_d^* - \hat{\mathbf{z}})'(\mathbf{z}_d^* - \hat{\mathbf{z}})}{(\mathbf{z}_d^* - \bar{\mathbf{z}}_d^*)'(\mathbf{z}_d^* - \bar{\mathbf{z}}_d^*)}, \tag{36}$$

where the bar over a symbol indicates a constant vector of means of the indicated vector. Reasoning like that presented above leads to the conclusion that

$$\mathbf{z}_c^* = \mathbf{z}^u \tag{37}$$

and that

$$\mathbf{z}_d^* = (\mathbf{z}^u - \hat{\mathbf{z}})\left[\frac{(\hat{\mathbf{z}} - \hat{\mathbf{z}})'(\hat{\mathbf{z}} - \hat{\mathbf{z}})}{(\mathbf{z}^u - \bar{\mathbf{z}}^u)'(\mathbf{z}^u - \bar{\mathbf{z}}^u)}\right] + \hat{\mathbf{z}}. \tag{38}$$

*2.5 Partitions*

The final point to be made in this section concerns what we term "measurement partitions." In some sets of data all of the observations are thought of as having arisen from a single measurement source. Furthermore, with some of these sets of data the measurement source generates data in such a way that all of the observations are reasonably assumed to have the same measurement characteristics. For example, when a subject makes similarity judgments concerning pairs of stimuli, then all of the judgments can reasonably be thought of as having been generated by a single source (the subject) and as all having the same measurement characteristics (discrete ordering of the similarity judgments). However, for other types of data it is clearly the case that the data arise from several measurement sources, or on several scales.

For example, when we obtain multivariate survey data with variables such as sex, age, hair color, income, educational background and political preference from a set of people, we would probably think of each variable as a unique measurement source having its own separate measurement characteristics: Sex is binary; age is ratio; hair color is nominal; income may be interval, educational background may be ordinal; and preference is ordinal. In this case we would wish to partition the data space into a set of mutually exclusive and exhaustive subspaces (one for each variable).

While the notion of partitions most clearly relates to multivariate data, the notion is also useful for other types of data. For example, Coombs' [1964] notion of conditional similarities data (for which a subject rank orders the similarity of $n-1$ "comparison" stimuli with respect to the $n$'th "standard" stimulus, and then does this $n$ times, each time with a different stimulus as the "standard") is a situation in which a single measurement source (the subject) generates $n$ different measurement scales (the rank orders). For this type of data measurement partitions are also of great use.

When the data are partitioned, the OS phase of an ALSOS procedure is slightly more complicated than when they are not partitioned, but only slightly. The difference is that we must perform the OS and normalization for each partition separately, one partition at a time. Since the partitions are mutually exclusive, and since the OS is performed for each partition separately, the measurement characteristics of one partition need bear no special relationship to those of another partition. This means, for example, that many of the procedures oriented towards multivariate data (see Table 1) can analyze data with mixtures of measurement characteristics. These data, which we call mixed measurement level data, can have one set of measurement characteristics for one variable, and a completely different set for another variable. A multivariate procedure (MORALS) is discussed in Section 3.2.

Note that for partitioned data the overall loss function is defined as the root-mean-square of the loss functions for each partition. Thus, if $\phi_i^2$ denotes the normalized loss function for the $i$'th of $p$ partitions, we define the overall loss as

$$\phi = \left[\frac{1}{p}\sum_{i}^{p}\phi_i^2\right]^{1/2} \tag{39}$$

There is a very important consideration here, however, which must not be overlooked. It is usually imperative that right after performing the optimal scaling for a particular partition we immediately replace the old optimal scaling with the new optimal scaling. As will become clear from the next portion of this paper, the immediate replacement is imperative when the partitions are dependent. (Partitions are "dependent" if the optimal scale values for at least one partition, assuming the others are fixed, are a function of the optimal scale values for at least one other partition.) Since dependence is generally a characteristic of multivariate data, the programs which analyze such data involve immediate replacement. This point has been emphasized in Young, de Leeuw, & Takane [1976].

If, in fact, the partitions are dependent, then there is one additional consideration. Let's say, for the multivariate data case, that we have completed a cycle of optimal scaling and replacement for each variable. Now let's say that we repeat the optimal scaling of one of the variables. If we do this, then the second optimal scaling of the variable does not yield the same quantification as the first optimal scaling. Why is this? Because the variables are dependent. The quantification obtained by optimally scaling one variable depends on the quantification of each of the other variables.

While this all sounds somewhat bothersome, it can be shown [de Leeuw, Young & Takane, 1976] that were we to perform "inner" iterations ("inner" with respect to the scheme in Figure 1) of the cycle of optimal scaling and replacement, then this process would converge to a point where the quantifications would no longer change upon repeated optimal scaling. In our work we do not perform such inner optimal scaling iterations, however, only performing the process once for each variable (or partition) before switching to the model estimation phase (see Figure 1). Our experience has been that such inner iteration only serves to decrease the overall efficiency of the procedure, and de Leeuw [Note 3] has proven that the number of inner interations has no effect on the eventual convergence point.

## 3. *ALSOS Algorithms*

In this section we discuss two ALSOS algorithms in detail, the ADDALS and MORALS algorithms. In conjunction with the ADDALS discussion we present the discrete-nominal, discrete-ordinal and continuous-ordinal transformation processes (in terms of indicator matrices) in detail.

### *3.1 ADDALS Algorithm*

In this section we discuss the overall ADDALS algorithm [de Leeuw, Young & Takane, 1976] for additivity (conjoint) analysis. The steps of the algorithm are presented in Figure 4. We discuss the ADDALS algorithm first because it is the simplest.

The ADDALS algorithm describes tabular data by using the simple additive model. This is the "main effects" analysis of variance model, the analysis of variance model which has no interaction term:

$$z_{ij}^* \simeq \alpha_i + \beta_j + \mu. \tag{40}$$

Note that we have reorganized the vector $\mathbf{z}^*$ with element $z_i^*$ into a two-way table $Z^*$ with element $z_{ij}^*$.

The initialization of ADDALS is very simple (see step START of Figure 4). We simply call the raw data the initial "optimally scaled" data ($Z^* = \mathbf{O}$). These initial "scaled" data serve as the input to the model estimation step that is next.

The model estimation phase of ADDALS (step MODEL in Figure 4) begins by estimating the parameters $\alpha_i$, $\beta_j$ and $\mu$ of (40). The estimation method is well known: We use the grand mean of $Z^*$ to estimate $\mu$, and the mean of the $i$'th row's (or $j$'th column's) deviation from $\mu$ to estimate $\alpha_i$ (or to estimate $\beta_j$). This is the same as with regular analysis of variance, except we use the optimally scaled data $Z^*$ in place of the raw data $\mathbf{O}$.

ADDALS ALGORITHM

| | | |
|---|---|---|
| START: | READ O AND ITS MEASUREMENT CHARACTERISTICS. NORMALIZE O AND SET $Z^* = O$. | INITIALIZATION |
| MODEL: | CALCULATE $\alpha_I$, $\beta_J$, AND $\mu$ AS THE ROW, COLUMN AND GRAND MEANS OF $Z^*$. | MODEL PARAMETERS |
| | $\hat{Z}_{IJ} = \alpha_I + \beta_J + \mu$ | MODEL ESTIMATES |
| FIT: | $\phi = \left[\|\|\hat{Z}-Z^*\|\|/\|\|Z^*\|\|\right]^{1/2}$ | FIT |
| | IF $\phi$ IS SMALL, OR IF THE CHANGES IN $\phi$, $\alpha_I$, $\beta_J$ AND $\mu$ ARE SMALL, GO TO QUIT, OTHERWISE GO TO SCALE. | TERMINATION |
| SCALE: | $Z^U = U(U'U)^{-1}U'\hat{Z}$ | SCALING |
| | $Z^* = Z^U \left[\frac{\|\|\hat{Z}\|\|}{\|\|Z^U\|\|}\right]$ | NORMALIZATION |
| | GO TO MODEL | |
| QUIT: | OUTPUT RESULTS AND STOP. | |

FIGURE 4
The major steps in the ADDALS algorithm

Following the parameter estimation step we calculate the model estimates $\hat{z}_{ij}$ by applying the model in (40). These $\hat{Z}$ are the values which minimize $\phi$ for the particular $Z^*$ we have on the current iteration.

The goodness of fit $\phi$ is calculated in step FIT. If $\phi$ is small (good fit) or if $\phi, \alpha_i, \beta_j$ and $\mu$ all haven't changed much from the previous iteration (convergence), we quit. Otherwise we proceed to rescale the data.

The model estimates $\hat{\mathbf{z}}$ from step MODEL serve as input to the optimal scaling (step SCALE). (Note that we have just reshaped the $p$ by $q$ matrix $\hat{Z}$ into a vector $\hat{\mathbf{z}}$ which has $p$ times $q$ elements. This simplifies the notation.) The projection operator $U(U'U)^{-1}U'$ is applied to the estimates $\hat{\mathbf{z}}$ to obtain the unnormalized scaled data $\mathbf{z}^u$, which in turn are normalized to obtain the optimally scaled data $\mathbf{z}^*$. These $\mathbf{z}^*$ are the values which minimize $\phi$ for the particular $\hat{\mathbf{z}}$ we have on the current iteration. The indicator matrix $U$, of course, is defined by whatever process corresponds to the measurement characteristics of the data $\mathbf{O}$, as discussed in section 2.2. We now reshape $\mathbf{z}^*$ into $Z^*$ and return to step MODEL to obtain new model estimates based on the newly scaled data $Z^*$.

In the remainder of this section we present three detailed and completely worked out artificial examples of the ADDALS algorithm. The three examples all involve analyzing the same 3 × 3 table of data, but under three different sets of measurement assumptions: discrete-nominal, discrete-ordinal, and continuous-ordinal. The table of data is (arbitrarily):

| | | |
|---|---|---|
| *A* | *C* | *B* |
| *A* | *B* | *B* |
| *C* | *A* | *B* |

These data could have been obtained, for example, in a 3 × 3 experiment in which the two experimental variables are wind-speed (none, slow, fast) and temperature (−10°C, 0°C, and 10°C), and in which we ask people to judge the relative perceived temperatures as being cold, (*C*); colder, (*B*); and coldest, (*A*).

*Discrete-nominal.* In Figure 5 we present the first two iterations of ADDALS when the data are presumed to be discrete-nominal. The first iteration is shown on the left panel, the

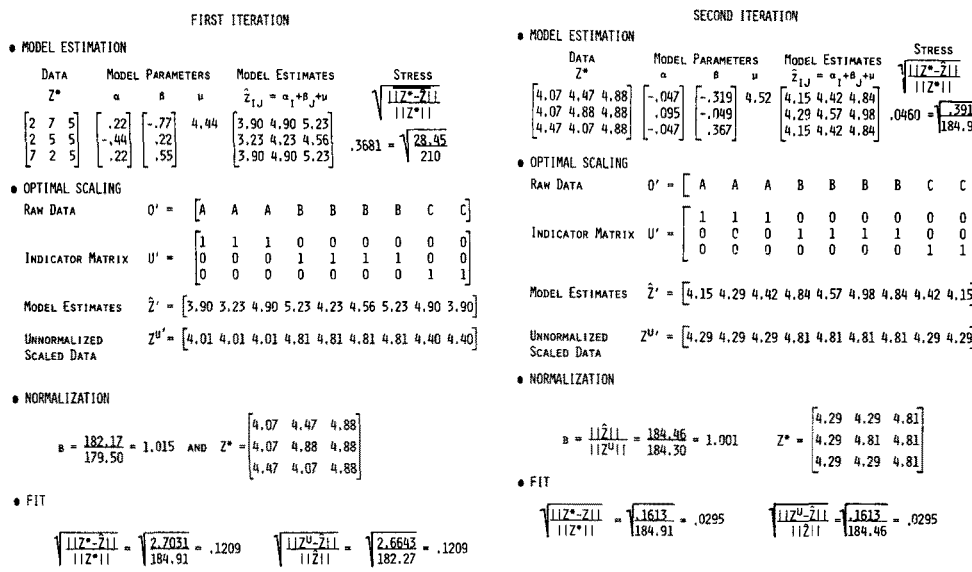


FIGURE 5
Discrete-Nominal ADDALS example

second on the right. Each panel is divided into four sections called Model Estimation, Optimal Scaling, Normalization, and Fit. Note that the data on the first iteration ($Z^*$, the left-most matrix in the Model Estimation section of the left panel) use different symbols to code the categories than the $A$, $B$, and $C$ given above. We must use numbers, not letters, because the initialization requires them. Thus, we have chosen to set $A = 2$, $B = 5$, and $C = 7$. This initial $Z^*$, then, is arbitrary, and any other initial assignment of numbers to the categories $A$, $B$, and $C$ would suffice. The initial assignment may cause the algorithm to obtain a local minimum as a solution, instead of the global minimum [Young & Null, 1978]. Thus, we should be careful at this point. In fact, it is desirable to try several different assignments and to observe their effect on the results, particularly when the categories are truely unordered. For the example given (perceived temperature), the categories are potentially ordered, and we have chosen numbers which are in that potentially "correct" order. However, the analysis is nominal, thus the initial order need not be preserved.

To the right of $Z^*$ are the values of the model parameters $\alpha_i$, $\beta_j$, and $\mu$. The mean of $Z^*$ is $\mu$, and the deviation of the row and column means from $\mu$ are the $\alpha_i$ and $\beta_j$, respectively. To the right of the parameters is the matrix of estimates $\hat{Z}$ and the measure of fit (called stress since this is Kruskal's Stress index). The parameters yield estimates $\hat{Z}$ which are a least stress fit to $Z^*$, conditional on the arbitrary initial coding of the three observation categories. Note that what has been done to this point is a classical ANOVA of the $Z^*$ using the main effects model.

ADDALS now proceeds to the optimal scaling on the first iteration. Figure 5 presents the raw data vector **o**, with the observations coded as $A$, $B$ and $C$, and with all observations in a given category being adjacent to each other. (The order of the observations in this vector is irrelevant. The order shown simplifies the presentation.) Directly below **o** is $U$, the indicator matrix (note that the transpose of $U$ is shown). $U$ has three columns, one for each category, and nine rows, one for each observation. Finally, note that the vector of nine model estimates, $\hat{\mathbf{z}}$, appears directly below $U$. The order of the elements of $\hat{\mathbf{z}}$ is *not* arbitrary, but is in the order dictated by the order of the elements in the vector **o**. The first element in the *vector* $\hat{\mathbf{z}}$ (3.90) corresponds to the first observation in the *vector* **o** (an $A$) because the value $3.9\overline{0}$ comes from a cell in the *matrix* $\hat{Z}$ which corresponds to an $A$ observation in the **O** *matrix*. The correspondence between the elements of the vectors $\hat{\mathbf{z}}$ and **o** is the same throughout: Each value in the vector $\hat{\mathbf{z}}$ is the model's least squares estimate of the current numerical coding (quantification) of the category given in the vector **o** above.

The input to the optimal scaling step is the indicator matrix $U$ and the model estimates $\hat{\mathbf{z}}$. The output is the unnormalized scaled data $\mathbf{z}^u$. You will recall that $\mathbf{z}^u = U(U'U)^{-1}U'\hat{\mathbf{z}}$, according to (7). We note that the diagonal of $U'U$ is [3, 4, 2], which is the number of observations in each category. Of course DIAG$[(U'U)^{-1}]$ = [1/3, 1/4, 1/2]. Furthermore, the projection operator $E = U(U'U)^{-1}U'$ is a block diagonal matrix with three blocks, one for each category. The order of a block equals the number of observations in the corresponding category, and all elements in a block equal the reciprocal of the category frequency. Finally, the transpose of $(U'U)^{-1}U'\hat{\mathbf{z}}$ is [4.01, 4.81, 4.40]. These three values are the unnormalized scale values for the three observation categories, also called the unnormalized observation category parameter values. Note that these three values are *not* in the anticipated order: Category $B$, which had the middle initial value (5), now has the largest value.

ADDALS now proceeds to the normalization step in Figure 5. As discussed in section 2.4 and illustrated in Figure 3, the vector $\mathbf{z}^u$ which was just computed minimizes the unnormalized fit to $(\|Z^u - \hat{Z}\|)$, but not the normalized Stress index $(\|Z^* - \hat{Z}\|/\|Z^*\|)^{1/2}$ at the top of Figure 5. As we showed in section 2.4, to minimize the normalized Stress index we must compute the normalization constant $b$ given by (31). For the first iteration $b = 1.015$, as is shown on the bottom left portion of Figure 5. This yields the matrix of (normalized)

optimally scaled data $Z^*$ (given next in the Figure) and a (normalized) Stress value of .1209. We see that the optimal scaling has improved the fit from .3681 to .1209, a big improvement.

Note that we can also calculate the Stress using the unnormalized $Z^u$ if we make sure to use the model estimates in the denominator [(16), above]. The value for this formula, which is shown at the bottom of the left panel of Figure 5, must also be .1209. The question might be asked, then, why normalize if we can use the unnormalized $Z^u$ to calculate Stress just by using a slightly different Stress formula? The answer is that while it is true that we can use $Z^u$ to calculate Stress, we *must* have $Z^*$ in order to start the next iteration properly.

The second iteration is shown in the right panel of Figure 5. We are only going to comment on the values of Stress on this iteration. We see that Stress is .0460 after the second model estimation, and goes down further to .0295 after the second optimal scaling. These values of fit will always improve (decrease) until ADDALS converges on a point of no further change. This is the *convergent* nature of all ALSOS algorithms: The fit never worsens and eventually stabilizes. Note that if we had used different initial values for the observation categories the algorithm would still remain convergent, but would perhaps have converged on a different Stress value. If so, the larger Stress value would be a local minimum [Young & Null, 1978].

ADDALS would take a few more iterations to reach a point where Stress ceases to improve by very much, and would then stop. We do not report the rest of these iterations. Note that we stop after the model estimation because at that point the data are scaled in a fashion which yields the parameters and Stress just calculated.

*Discrete-ordinal.* We now discuss the discrete-ordinal analysis of these same data. The first two iterations are shown in Figure 6. The discrete-ordinal ADDALS algorithm is exactly the same as the discrete-nominal ADDALS algorithm, with but a single exception: An order constraint is imposed on the observation categories during the optimal scaling. The constraint is introduced via the $U$ matrix.

One implication of this relationship between the discrete-nominal and discrete-ordinal algorithms is that the model estimation step on iteration one is precisely the same (compare Figures 5 and 6). In fact, the optimal scaling step starts out in the same way for both levels of measurement. Specifically, for the discrete-ordinal case the indicator matrix $U$ starts out to be the same as the indicator matrix used for the discrete-nominal case: It simply indicates the category structure. Thus, the FIRST TRY (left panel of Figure 6) computes the same $Z^u$ as is computed for the discrete-nominal case (left panel of Figure 5). However, when the $Z^u$ values are inspected we see that they are *not* in the required order: The middle raw observation category (5) has been assigned the largest $Z^u$ (4.81), and the largest category (7) a smaller $Z^u$ (4.40).

To cope with this order violation we modify $U$ and have a SECOND TRY. The new $U$ still has nine rows, one for each observation, but it has only two columns, one for the smallest observation category and one for the two order violating categories. Thus, we have merged the two violating categories into one "block." We now repeat the calculation of $Z^u$ using this new $U$ and check to see if its values are properly ordered. They are, so we proceed to the normalization step. Of course, if the $Z^u$ entries were still disordered we would have tried again with the order violating columns of $U$ merged.

Note that we have just looked in detail at the critical difference between the nominal and ordinal levels. For nominal, $U$ is known before analysis, remains constant, and simply indicates the observation category structure. For ordinal (discrete or continuous), $U$ is *not* known but must be determined. It is *not* constant, but is variable. And for discrete-ordinal $U$ does *not* indicate category structure, but does indicate blocks of categories which must be merged to maintain order.

We will not discuss the remainder of Figure 6 in detail; rather, we let you peruse it at

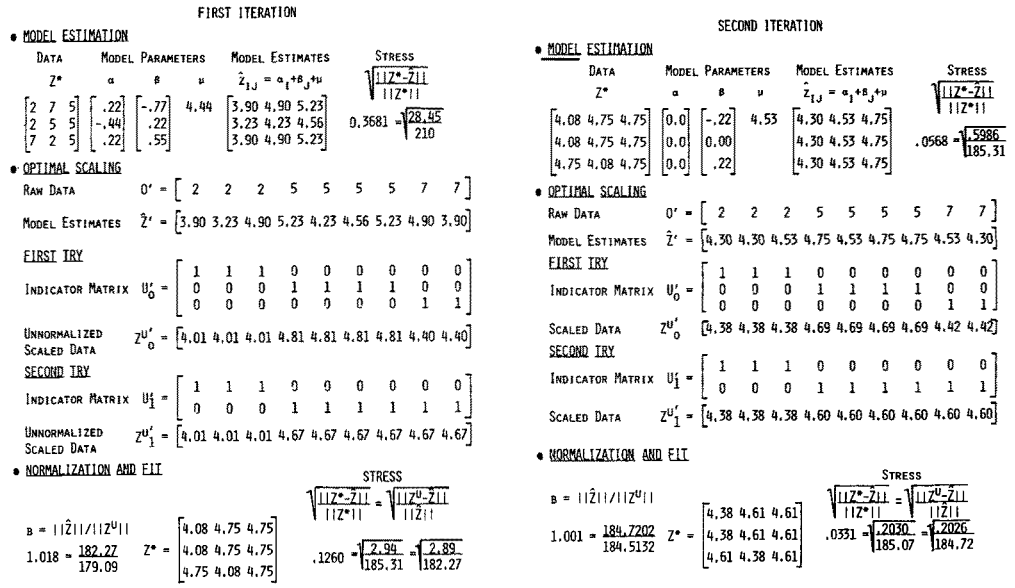


FIGURE 6
Discrete-Ordinal ADDALS example

your leisure. Note that the normalization procedures are the same as with any other measurement characteristics, although the numbers are different than in Figure 5 because of the ordinal constraint in the optimal scaling. Note also that the Stress values are all larger than with the discrete-nominal assumptions, because we have had to impose the order constraints.

*Continuous-ordinal.* The algorithm for continuous-ordinal is precisely the same as the previous two measurement cases except for two differences, both in the optimal scaling step.

The first difference is that a permutation matrix $P$ is used. The order of $P$ equals the number of observations. Furthermore, $P$ is block diagonal, with the number of blocks equaling the number of categories, and the order of each block equaling the number of observations in each category. The permutation matrix permutes (sorts) the model estimates $\hat{Z}$ into order within each observation category.

For the example we've been discussing (Figure 7) the permutation matrix on the first iteration is

$$\begin{bmatrix} 0 & 1 & 0 & & & & & & & \\ 1 & 0 & 0 & & & & & & & \\ 0 & 0 & 1 & & & & & & & \\ & & & 0 & 1 & 0 & 0 & & & \\ & & & 0 & 0 & 1 & 0 & & & \\ & & & 1 & 0 & 0 & 0 & & & \\ & & & 0 & 0 & 0 & 1 & & & \\ & & & & & & & 0 & 1 & \\ & & & & & & & 1 & 0 & \end{bmatrix}$$

This matrix, when applied to the model estimates, yields the permuted $\hat{Z}$ shown in Figure 7. Note that $P$ is a variable, varying between iterations, just as does $U$.

The second difference between continuous-ordinal and the previous two measurement cases is that $U$ starts out being the identity matrix whose order equals the number of observations (not the number of categories). Note that $U$ does not start out reflecting the

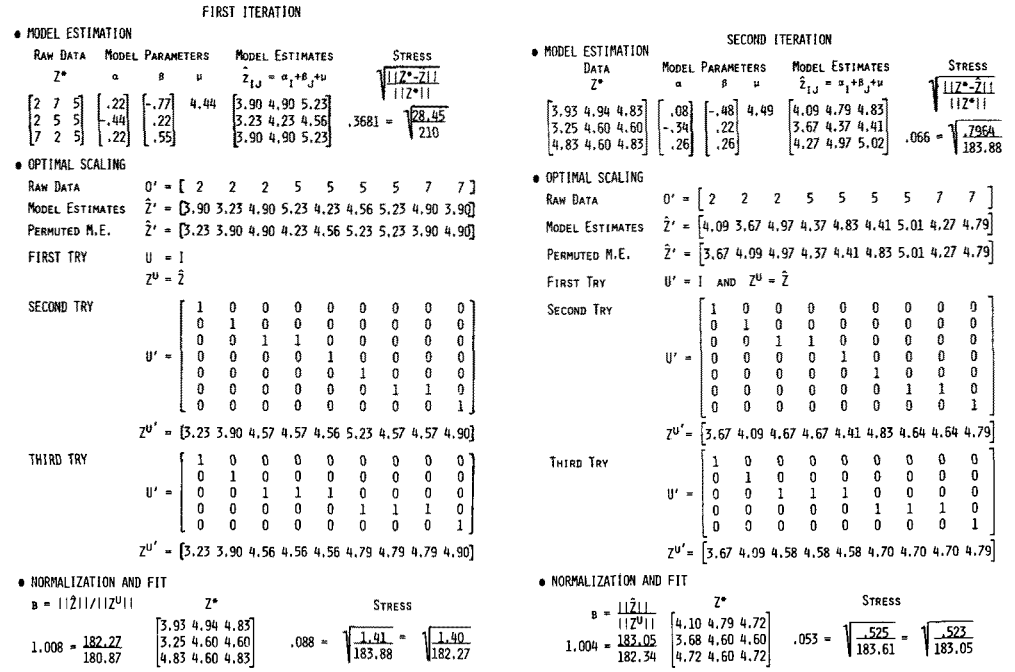


FIGURE 7
Continuous-Ordinal ADDALS example

category structure. Because of this $U$ does not end up reflecting the category structure, *nor* the structure of categories which must be blocked to preserve order. Rather, $U$ ends up indicating which observations (not categories) must be blocked to preserve order. We are not going to review Figure 7 in detail.

### *3.2 The MORALS Algorithm*

In this section we briefly review the overall MORALS algorithm [Young, de Leeuw, & Takane, 1976] for multiple regression with multivariate data whose variables each have their own independent measurement characteristics. We discuss only the algorithm. We present no detailed examples like those in the previous section, as we deem them unnecessary.

The important aspect of MORALS is that it permits the multivariate data to have any mix of measurement types: Some variables can be nominal, others ordinal, and yet others interval. Similarly, any variable can be discrete or continuous. This flexibility applies to the dependent variable as well as the independent variables. In fact, the algorithm has been extended to the case where there are multiple dependent variables with mixed measurement characteristics [CORALS and CANALS by Young, de Leeuw & Takane, 1976] and to the case where there are multiple sets of variables instead of two sets, each set having mixed measurement variables [OVERALS, by Gifi, 1981]. Closely related is the PATHALS algorithm for path analysis with mixed measurement level data [Gifi, 1981], and the CRIMINALS algorithm for discriminant analysis with mixed measurement level predictors [Gifi, 1981].

The reason that we choose to discuss the MORALS algorithm is because it is the simplest algorithm which involves the concept of measurement partitions, a concept not illustrated by the ADDALS algorithm. The partitions notion, discussed in section 2.5, is very appropriate to multivariate data, since it is usually the case that the observations on one variable are not directly comparable to those on another variable.

MORALS ALGORITHM

| | | |
|---|---|---|
| START: | READ Y AND X AND THEIR MEASUREMENT CHARACTERISTICS. | |
| | SET $\underline{Y}^*=\underline{Y}$ AND $X^*=X$ . | INITIALIZATION |
| MODEL: | $\beta = (X^{*\prime}X^*)^{-1}X^*\underline{Y}^*$ | MODEL PARAMETERS |
| FIT: | CALCULATE MULTIPLE $R^2$. IF IT HASN'T IMPROVED "ENOUGH" FROM LAST ITERATION, QUIT. | TERMINATION |
| SCALE: | $\hat{\underline{Y}} = X^*\beta$ | MODEL ESTIMATES |
| | $\underline{Y}^U = U_Y(U_Y'U_Y)^{-1}U_Y'\hat{\underline{Y}}$ | OPTIMAL SCALING |
| | $\underline{Y}^* = \underline{Y}^U\left(\frac{\|\|\hat{\underline{Y}}\|\|}{\|\|\underline{Y}^U\|\|}\right)$ | NORMALIZATION |
| LOOP: | FOR J=1,M VARIABLES | |
| | $\hat{\underline{X}}_J = \frac{1}{\beta_J}(\underline{Y}^* - \sum_{I\neq J}\beta_I\underline{X}_I^*)$ | MODEL ESTIMATES |
| | $\underline{X}_J^U = U_J(U_J'U_J)^{-1}U_J'\hat{\underline{X}}$ | OPTIMAL SCALING |
| LOOPEND: | $\underline{X}_J^* = \underline{X}_J^U\left(\frac{\|\|\hat{\underline{X}}_J\|\|}{\|\|\underline{X}_J^U\|\|}\right)$ | NORMALIZATION |
| | GO TO ITERATE | |

FIGURE 8
The major steps in the MORALS algorithm

The structure of the MORALS algorithm is presented in Figure 8. In this figure **y** is a vector of $K$ raw observations on one dependent variable, and $X$ is a matrix of $K$ raw observations on $M$ independent variables. Each of the $M + 1$ variables has its own measurement characteristics and has its own partition. Thus, there are $M + 1$ partitions, $M + 1$ indicator matrices, and $M + 1$ optimal scaling steps.

The START step is similar to that used in the ADDALS algorithm: The initial "optimally scaled" data are simply the raw data. The MODEL step is simply a multiple regression analysis of the optimally scaled variables $y^*$ and $X^*$. The FIT step is similar to the ADDALS FIT step.

The new aspect is the SCALE step. Notice that it is divided into two major sections, the first for the single dependent variable (at step SCALE), and the second for the M independent variables (at step LOOP). For each variable we calculate the model's estimate of that variable ($\hat{\mathbf{y}}$ or $\hat{\mathbf{x}}_j$), then use the estimate with the appropriate indicator matrix ($U_y$ or $U_j$) to calculate the unnormalized optimally scaled data ($\mathbf{y}^u$ or $\mathbf{x}_j^u$), and then perform the normalization to obtain $\mathbf{y}^*$ or $\mathbf{x}_j^*$. Notice that the newly computed $\mathbf{y}^*$ or $\mathbf{x}_j^*$ replace their previous values. The model estimate equation for the dependent variable is straightforward, and the one for the independent variable has been explained by Young, de Leeuw and Takane [1976]. The scaling and normalization steps are the same as with the other algorithms.

The difference, then, is that we have several partitions and that we do the model

estimation, scaling, and normalization for each one. It is important to point out that the partitions are *not independent*, that is, the values being calculated for one partition are dependent on the values calculated for all other partitions. To assure convergence and to maintain the ALS aspect of an algorithm with nonindependent partitions we must immediately replace the previous scaled data with the newly computed (normalized) scaled data.

The nonindependence of the partitions also brings up another important point. If, after the LOOPEND in Figure 8 we return to step SCALE instead of step MODEL, and if we repeated the scaling of each variable, we would obtain a new and different scaling which would fit better than before. Thus, to make the scaling of all variables least squares (in an overall sense) we would have to perform "inner" iterations on the scaling of the variables until convergence is reached on their scaling. However, we have found this to be inefficient, and instead return to the MODEL step to obtain improved values for the model parameters.

### *3.3 The ALSCAL Algorithm*

Most of the procedures that have been developed on the ALSOS principle are simple in the model estimation phase. For example, the PRINCIPALS and PRINCIP procedures apply the principal components model to mixed measurement level multivariate data [Young, Takane, & de Leeuw, 1978; de Leeuw & van Rijkevorsel, Note 4]. For these algorithms the model estimation phase is nothing more than a standard eigenvalue decomposition of the optimally scaled data.

The only procedure which involves a fairly complicated model estimation phase is the ALSCAL algorithm [Young, Takane, & Lewyckyj, 1978, 1980; Young & Lewyckyj, 1979, 1980], for performing individual differences multidimensional scaling [Takane, Young, & de Leeuw, 1977]. However, the complexity of the model estimation phase lies in the very nature of the model: There are several sets of parameters which are not mutually independent (as, for example, are the several sets of parameters of the additive model), and which are not all linearly related to the loss function (as is also the case in the additive model). These characteristics of the model can be seen from the equation defining the model:

$$\hat{z}_{ijk} = \sum_{a=1}^{t} v_{ia} w_{ka} (x_{ia} - y_{ja})^2 \tag{41}$$

where $\hat{z}_{ijk}$ is a tabular reorganization of the model estimates $\hat{\mathbf{z}}$, with subscripts $i$ and $j$ referring to objects or events about which we have some sort of similarity information, and subscript $k$ referring to situations (subjects, experimental conditions, etc.) under which the similarity information is observed. The parameters $v_{ia}$ are "stimulus weights" of the asymmetric Euclidean model [Young, 1975b], $w_{ka}$ are subject weights of the individual differences model discussed by Carroll and Chang [1970] and Horan [1969], $x_{ia}$ are stimulus-object points in a Euclidean space, and $y_{ja}$ are ideal points for Coombs' unfolding model [1964] or attribute points for preference data.

When we say that the several sets of parameters are not mutually independent we mean that estimating the values of at least one set of parameters involves estimates already obtained for at least one of the other sets of parameters. When parameters are not independent, the values of the parameters in one set affect the values estimated for the parameters in the other set. This way of looking at the difficulty immediately suggests a solution to the problem, however. All we have to do is to define an ALS "inner" interation which estimates parameters, one set at a time. Thus, for ALSCAL, which is based on the model in (41), the inner iteration has four phases each using the values of the parameters in three of the sets (and the optimally scaled data) to obtain conditional least squares estimates for the parameters in the fourth set. Once the parameters in a set are estimated, they are immediately

used to replace their old values, and the procedure moves on to another one of the four model parameter sets. This four phase ALS procedure could be iterated until convergence is obtained (there would be inner iterations).

Actually, ALSCAL does not use the inner iteration procedure outlined in the previous paragraph. It would be very slow to require the inner iterations of the model estimation phase to converge before going on to the optimal scaling phase. Experience shows that we should only cycle through the four phases of the inner iteration once, defining that to be a complete model estimation phase. Note that the considerations about nonindependent data partitions apply in precisely the same fashion to nonindependent model parameter sets.

The second source of complexity in the ALSCAL algorithm is the nonlinear relationship between the stimulus-object points $x_{ia}$ and $y_{ja}$ and the model estimates $\hat{z}_{ijk}$. We do not go into this problem here except to say that the solution we use is to apply the ALS principle yet a third time (defining what might be called "innermost" iterations) to estimate the conditional least squares value for a single point's coordinates, one coordinate at a time, under the assumption that all of the other coordinates are constant. This innermost iteration involves $n*t$ phases, one for each of the $n$ points on each of the $t$ dimensions.

The ALSCAL algorithm involves a concept which does not arise in the other algorithms: The parameters of the model are not mutually independent. The algorithm, then, serves to illustrate one method for coping with parameter dependence, namely the use of inner iterations to reapply the ALS principle. The algorithm also serves to illustrate that we do not have to iterate the inner iterations until convergence is reached (one "iteration" can suffice).

As mentioned above, the notion of inner iteration is involved in the ALSOS system in one other critical place: the method for optimally scaling data which are partitioned into dependent partitions. When we view the observation categories as parameters and the optimal scale values assigned to each category as parameter values, then we see that we need knowledge of some parameter values in estimating other parameter values. This is precisely the definition of dependence given above, except that the problem occurs in the optimal scaling phase of the algorithm instead of in the model estimation phase. Note that data partitions are not always dependent [for example, the data partitions discussed by de Leeuw, Young, & Takane, 1976, for ADDALS, and by Takane, Young, & de Leeuw, 1977, for ALSCAL are independent] just as parameters are not always dependent. However, when dependence exists the ALS inner iteration approach is a viable approach to deal with the problem.

## *4. Conclusions*

The combination of alternating least squares and optimal scaling, which forms the foundation of the ALSOS approach to algorithm construction, has two primary advantages: (a) If a least squares procedure is known for analyzing numerical data, then it can be used to analyze qualitative data simply by alternating the procedure with the optimal scaling procedure appropriate to the qualitative data; and (b) under certain fairly general circumstances the resulting ALSOS algorithm is convergent and has no difficulties associated with estimating step size. It is the opinion of the author that the second advantage implies that ALSOS algorithms have fewer local minimum problems than gradient procedures which require step-size estimation.

We do not mean to imply that an ALSOS algorithm is the be-all and end-all of algorithms. It is not. It is simply a relatively straightforward approach to algorithm construction which has certain nice convergence properties. The resulting algorithm may not be very simple. With ALSCAL, for example, even though each step is not very complicated, the overall structure is rather complex due to the necessity of inner and innermost iter-

ations. Furthermore, in some circumstances there are some indeterminacies of construction which may have great effect on the overall speed of the algorithm (such as the number of inner iterations performed on each outer iteration). Finally, perhaps the biggest drawback is that the ALSOS approach does not guarantee convergence on the global optimum, but on a potentially local optimum. Since the convergence point is conditional on the initialization point, it is sometimes the case that the initialization procedure can become very complicated, and may be very crucial. We conclude, however, that the ALSOS approach to algorithm construction provides flexible and well-behaved methods for quantitative analysis of qualitative data.

## REFERENCE NOTES


1. Tenenhaus, M. Principal components analysis of qualitative variables. Report No. 175/1981. Jouy-en-Josas, France, Centre d'Enseignement Superieur des Affaires, 1981.
2. Tenenhaus, M. Principal components analysis of qualitative variables. Les Cahiers de Recherche No. 175/1981. Jouy-en-Josas, France, CESA, 1981.
3. de Leeuw, J. A normalized cone regression approach to alternating least squares algorithms. Unpublished note, University of Leiden, 1977b.
4. de Leeuw, J., & van Rijkevorsel, J. How to use HOMALS 3. A program for principal components analysis of mixed data which uses the alternating least squares method. Unpublished mimeo, Leiden University, 1976.
5. Young, F. W., Null, C. H., & De Soete, G. The general Euclidean Model. 1981 (in preparation).


## REFERENCES


Benzecri, J. P. *L' analyse des donnees—Tome II :* Correspondances Dunod, Paris, 1973.

Benzecri, J. P., Histoire et Prehistoire de l'analyse des donnees; l'analyse des correspondence. *Les Cahiers de l'Analyse des Donnees* (Volume II), Paris, 1977.

Bock, R. D. Methods and applications of optimal scaling. Psychometric Laboratory Report #25, University of North Carolina, 1960.

Burt, C. The factorial analysis of qualitative data. *British Journal of Psychology, Statistical Section*, 1950, *3*, 166–185.

Burt, C. Scale analysis and factor analysis. *British Journal of Statistical Psychology*, 1953, *6*, 5–24.

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

Coombs, C. H. *A Theory of Data.* New York: Wiley, 1964.

de Leeuw, J. *Canonical analysis of categorical data.* University of Leiden, The Netherlands, 1973.

de Leeuw, J. Normalized cone regression. Leiden, The Netherlands: University of Leiden, Data Theory, mimeographed paper, 1975.

de Leeuw, J. Correctness of Kruskal's algorithms for monotone regression with ties. *Psychometrika*, 1977a, *42*, 141–144.

de Leeuw, J., Young, F. W., & Takane, Y. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 1976, *41*, 471–503.

Fisher, R. *Statistical methods for research workers.* (10th ed.) Edinburgh: Oliver and Boyd, 1938.

Gifi, A. *Nonlinear multivariate analysis* (preliminary version). University of Leiden, Data Theory Department. 1981.

Guttman, L. The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.) *The prediction of personal adjustment.* New York: Social Science Research Council, 1941.

Guttman, L. A note on Sir Cyril Burt's "Factorial Analysis of Qualitative Data," *The British Journal of Statistical Psychology*, 1953, *7*, 1–4.

Hageman, L. A., & Porsching, T. A. Aspects of nonlinear block successive over-relaxation. *SIAM Journal of Numerical Analysis*, 1975, *12*, 316–335.

Hayashi, C. On the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, 1950, *2*, 35–47.

Horan, C. B. Multidimensional scaling: Combining observations when individuals have different perceptual structures. *Psychometrika*, 1969, *34*, 139–165.

Kruskal, J. B. Nonmetric multidimensional scaling. *Psychometrika*, 1964, *29*, 1–27, 115–129.

Kruskal, J. B. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society*, Series B, 1965, *27*, 251–263.

Kruskal, J. B., & Carroll, J. D. Geometric models and badness-of-fit functions. In P. R. Krishnaiah (Ed.), *Multivariate analysis (Vol. 2)*. New York: Academic Press, 1969.

Mardia, K. V., Kent, J. T., & Bibby, J. M. *Multivariate analysis*. London: Academic Press, 1979.

Nishisato, S. *Analysis of categorical data : Dual scaling and its applications*. University of Toronto Press, 1980.

Roskam, E. E. *Metric analysis of ordinal data in psychology*. Voorschoten, Holland: VAM, 1968.

Saito, T. *Quantification of categorical data by using the generalized variance*. Soken Kiyo, Nippon UNIVAC Sogo Kenkyn-Sho, 61–80, 1973.

Sands, R., & Young, F. W. Component models for three-way data: An alternating least squares algorithm with optimal scaling features. *Psychometrika*, 1980, *45*, 39–67.

Saporta, G. Liaisons entre plusieurs ensembles de variables et codages de donnes qualitatives. These de Doctorat de 3eme cycle, Paris, 1975.

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

Takane, Y., Young, F. W., & de Leeuw, J. An individual differences additive model: An alternating least squares method with optimal scaling features. *Psychometrika*, 1980, *45*, 183–209.

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

Wold, H., & Lyttkens, E. Nonlinear iterative partial least squares (NIPALS) estimation procedures. *Bulletin ISI*, 1969, *43*, 29–47.

Young, F. W. A model for polynomial conjoint analysis algorithms. In R. N. Shepard, A. K. Romney, & S. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences*. New York: Academic Press, 1972.

Young, F. W. Methods for describing ordinal data with cardinal models. *Journal of Mathematical Psychology*, 1975a, *12*, 416–436.

Young, F. W. An asymmetric Euclidian model for multi-process asymmetric data. U.S.-Japan Seminar on Multidimensional Scaling, 1975b.

Young, F. W., de Leeuw, J., & Takane, Y. Multiple (and canonical) regression with a mix of qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 1976, *41*, 505–529.

Young, F. W., & Lewyckyj, R. *ALSCAL Users Guide*. Carrboro, NC; Data Analysis and Theory, 1979.

Young, F. W., & Lewyckyj, R. The ALSCAL procedure. In SAS Supplemental Library User's Guide, Reinhardt, P. (Ed.). SAS Institute, Raleigh, NC, 1980.

Young, F. W., & Null, C. H. Multidimensional scaling of nominal data: The recovery of metric information with ALSCAL. *Psychometrika*, 1978, *43*, 367–379.

Young, F. W., Takane, Y., & de Leeuw, J. The principal components of mixed measurement level data; An alternating least squares method with optimal scaling features. *Psychometrika*, 1978, *43*, 279–282.

Young, F. W., Takane, Y., & Lewyckyj, R. ALSCAL: A nonmetric multidimensional scaling program with several individual differences options. *Behavioral Research Methods and Instrumentation*, 1978, *10*, 451–453.

Young, F. W., Takane, Y., & Lewyckyj, R. ALSCAL: A multidimensional scaling package with several individual differences options. *American Statistician*, 1980, *34*, 117–118.

Yule, G. U. *An introduction to the theory of statistics*. London: Griffin, 1910.

*Manuscript received 7/22/81*
*Final version received 7/22/81*