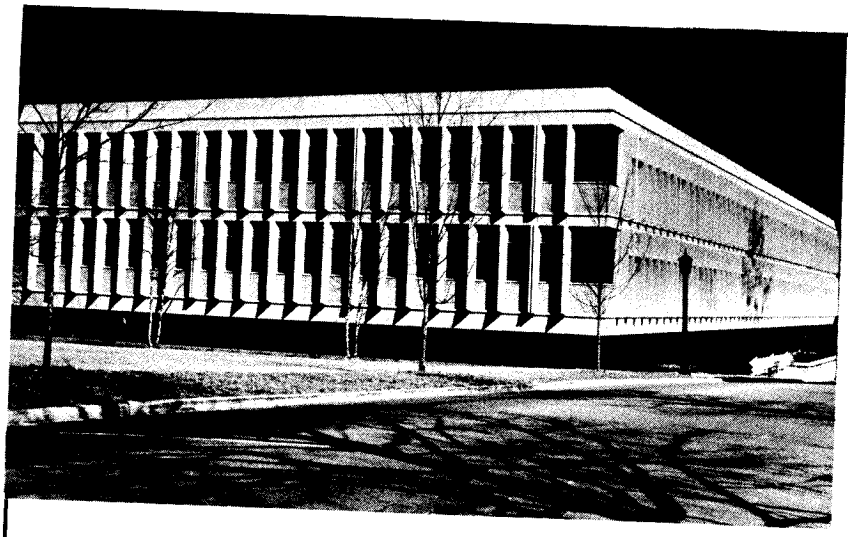
JAN
57



Quantifying Qualitative Data

Forrest W. Young

Jan de Leeuw

and

Yoshio Takane

July, 1976

Report Number 149

THE L. L. THURSTONE
PSYCHOMETRIC LABORATORY
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, N. C.
27514

Abstract

An overview of the theoretical and methodological foundations of an approach to quantifying qualitative data is presented. The two cornerstones of the approach, known as alternating least squares and optimal scaling, are explained. It is emphasized that the approach has two major advantages: a) If a least squares method is known for analyzing quantitative data then a least squares method can be constructed for analyzing qualitative data; and b) under certain fairly general circumstances the approach yields algorithms which are convergent and which have relatively few difficulties with local minima. A system of programs for quantifying qualitative data with either the additive, multiple regression, canonical regression, principal components or multidimensional scaling model is briefly discussed.

Portions of the work reported herein were supported by grants MH10006 and MH26504 from the National Institute of Mental Health to the Psychometric Laboratory of the University of North Carolina.

The current addresses of the authors are:

Forrest W. Young
Psychometric Laboratory
University of North Carolina
Chapel Hill, N.C. 27514

Jan de Leeuw
Datatheorie Instituut
Wassenaarseweg 80
Leiden, Nederland

Yoshio Takane
Department of Psychology
University of Tokyo
Tokyo, Japan

Quantifying Qualitative Data:
An overview of an approach having alternating least squares and optimal scaling features.

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 must either settle for 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 his respondents; the botonist obtains categorical information about the family to which his plants belong; and the psychologist obtains categorical information about the psychosis of his patient. Even in the best of circumstances it is often impossible to obtain anything beyond the order in which the data categories fall. When the sociologist observes the amount of education of the respondents in his sample he knows that the observation categories are ordered, but he 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 by Torgerson (1958, pp. 338-345). Much work has occurred recently, with the most important probably being performed by de Leeuw (1973), Benzicri (1973), and Nishisato (1973).

In this paper we refer to the process of quantifying qualitative data as "optimal scaling," a term first introduced by Boch (1960). By 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 characteristics of the data. Working with this definition of optimal scaling, the author of this paper and his coworkers have developed a system of programs for quantifying qualitative data (de Leeuw, Young & Takane, 1976; Takane, Young & de Leeuw, 1976; Young, de Leeuw & Takane, 1975). The programs permit the data to have a variety of measurement characteristics, and permit data analysis with a variety of models. We refer to this system of programs as the ALSOS system since it uses the Alternating Least Squares (ALS) approach to Optimal Scaling (OS).

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).

The ALSOS system currently includes six programs which quantify qualitative data by applying (a) the simple additive model, (b) the multiple regression model, (c) the canonical regression model, (d) the principal components model, or (e) the multidimensional scaling model. For five of the programs the data may be defined at the binary, nominal, ordinal or interval levels of measurement (and the ratio level with the multidimensional scaling program), and may be thought of as having been generated by either a discrete or continuous underlying process. The ALSOS programs also permit any arbitrary pattern of missing data, permit boundary or range restrictions on the values assigned to the observation categories, and permit the use of partial orders with ordinal data. The most salient characteristics of the programs are summarized in Table 1.

1. Alternating Least Squares

Each of the ALSOS programs optimizes an objective loss function by using an algorithm based on the alternating least squares and optimal scaling 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 (i.e., 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 optimal scaling parameters. We then proceed to optimize a loss function by alternately optimizing with respect to one subset, then the other. We do this

by obtaining the least squares estimates of the parameters in one subset while assuming that the parameters in the other 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 subset. Once we have obtained conditional least squares estimates we immediately replace the old estimates of these parameters by the new estimates. We then switch to the other subset of parameters and obtain their conditional least squares estimates. We alternately obtain conditional least squares estimates of the parameters in one subset, then the other subset, 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 investigated by Wold & Lyttkens (1969), the CANDECOMP algorithm of Carroll & Chang (1970), and the class of numerical analysis algorithms known as successive block algorithms (Hageman & Porsching, 1975). The main difference between these algorithms and an ALSOS algorithm is the optimal scaling feature of the ALSOS algorithm.

2. Quantification with Unknown Models: Theory

2.1 Introduction

One advantage of combining the ALS principle with the OS principle is that the OS phase of the algorithm does not need to know the type of model involved in the analysis. Thus, we can quantify qualitative data without knowing the specific nature of the model.

For the optimal scaling we need a model space and a data space, in the terminology of Young (1975), to obtain the optimal scaling space (see Figure 2).

We assume that there is a model space represented by a vector whose elements are measured at the cardinal (interval or higher) level. The model space is not

Table 1

Programs in the ALSOS system

PROGRAM	ANALYSIS	DATA
ADDALS	Additivity analysis. (Also known as analysis of variance without interaction terms, and as additive conjoint measurement.)	Two or three way tables. Nonorthogonal and incomplete designs permitted. Measurement characteristics apply to both independent and dependent variables.
ALSCAL	Multidimensional scaling, including several individual differences models.	Two or three way tables. Symmetric and asymmetric, conditional and unconditional, and ratio data permitted.
CORALS	Canonical regression.	Multivariate data, with any mixture of measurement characteristics permitted for the several variables.
MORALS	Multiple regression.	Same as CORALS.
PRINCIPALS	Principal components.	Same as CORALS.
HOMALS	Principal components	Multivariate data, all variables being nominal.

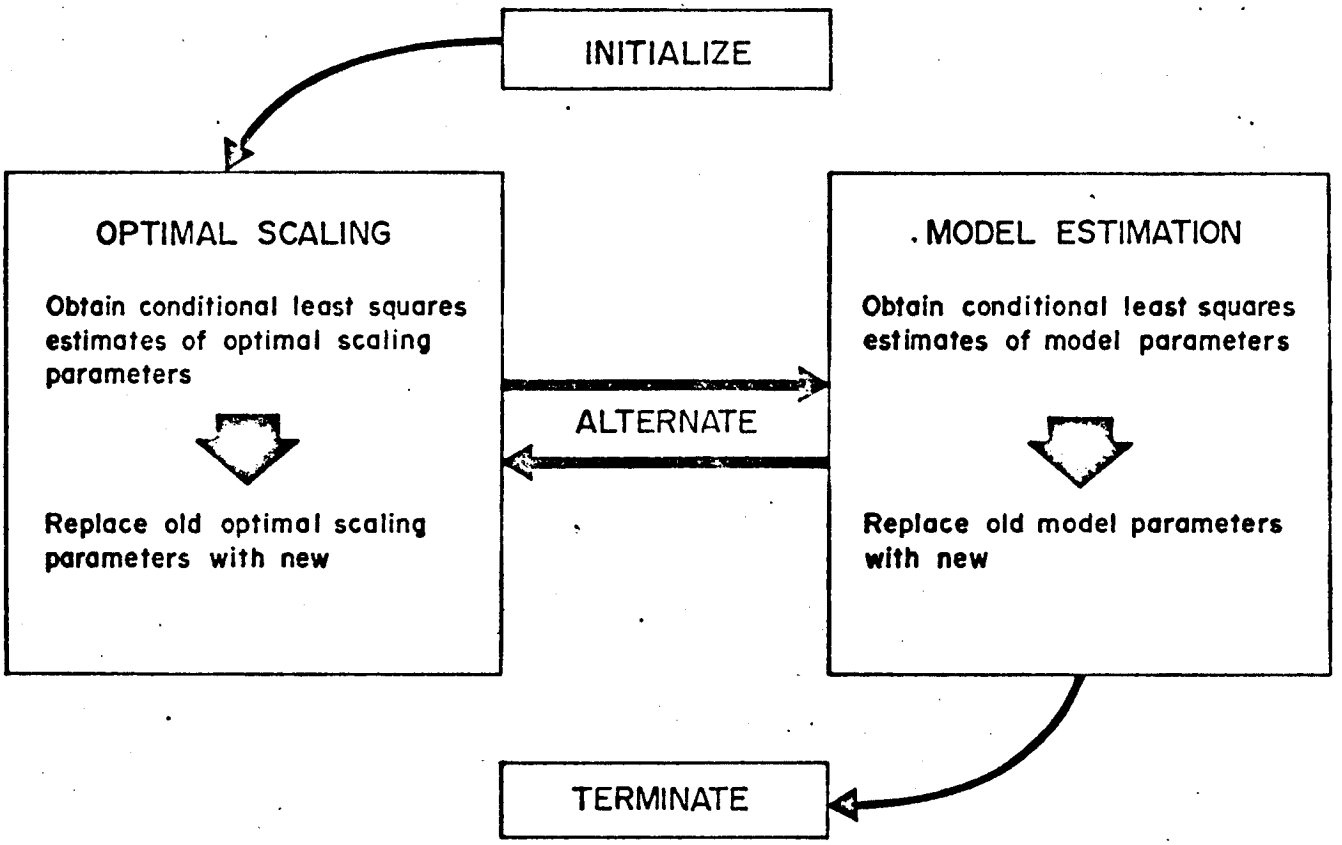


Figure 1: Flow of an ALSOS algorithm.

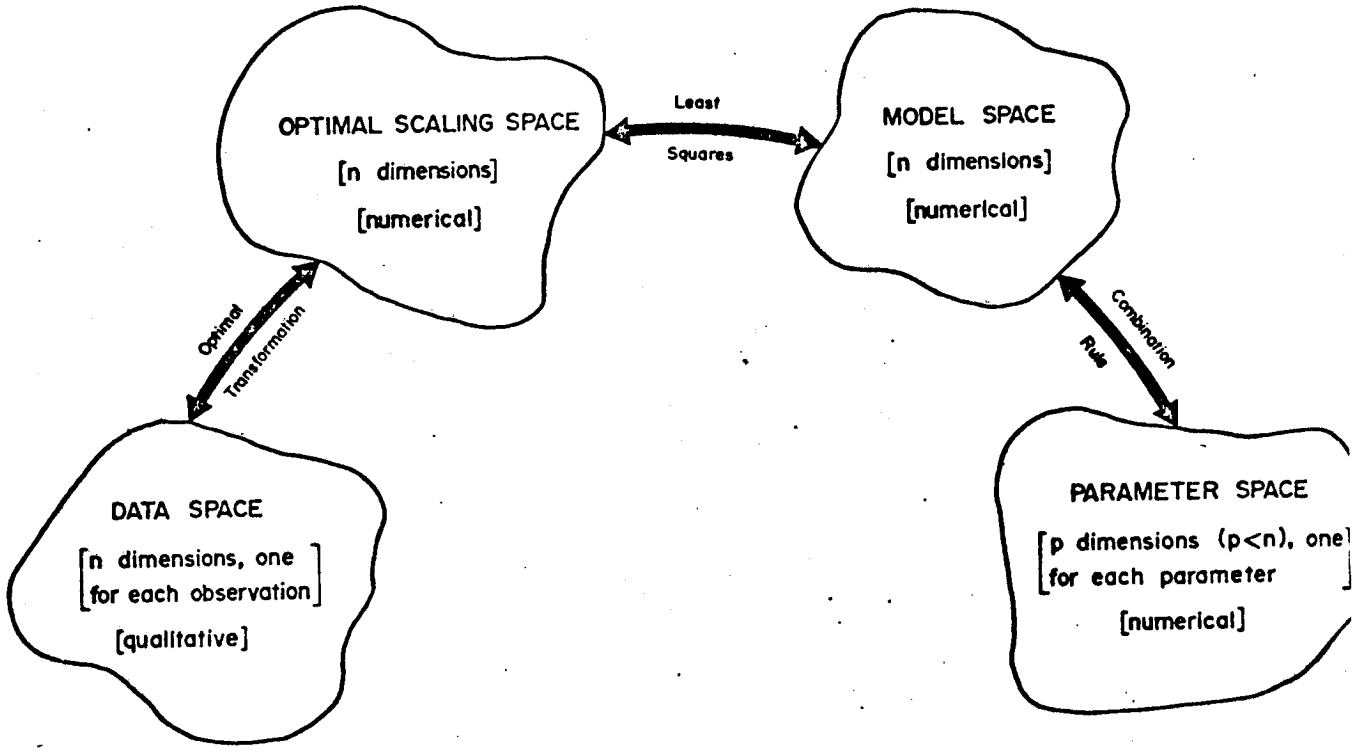


Figure 2: Some aspects of a data theory.

the parameter space. We do not know either the nature of the model (the combination or functional rule by which the model space is computed from the parameters), nor do we know the actual values of the model parameters. All we know is that some parameter values exist somewhere, and that somehow they have been combined together to yield the model space.

We also assume that there is a data space represented by a vector of data. We further assume that the measurement characteristics of the data (whether it is discrete or continuous, what the measurement level is) are known.

The goal of OS is to derive an optimal scaling space which has two characteristics: First, it must perfectly satisfy the measurement characteristics of the data space; and second, it must have a least squares relationship to the model space, given that the measurement characteristics are perfectly satisfied.

2.2 Transformations

To fully understand some of the concepts discussed below we must emphasize 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 most 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 uniquely different temperatures which would all be reported as say, 40°. Therefore, we say that the observation of 40° is categorical.

Let us define a vector of raw observations. We denote this observation vector as $\underline{o}$, with general element o_i. (Underlined lower case letters refer to vectors, and non-underlined lower case letters to scalars.) We also define the model vector $\underline{z}$, with general element z_i, and the optimally scaled observation vector $\underline{z}^*$, with general element z_i^*. The vector $\underline{o}$ is the data space (we assume that the elements in $\underline{o}$ are organized so that all observations in a particular category are contiguous). The vectors $\underline{z}$ and $\underline{z}^*$ are the model and optimally scaled observation spaces, respectively (we assume that their elements are organized in a fashion having a one to one correspondence with $\underline{o}$). The element z_i^* is the parameter representing the observation o_i.

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,

$$(1) \qquad \mathcal{t}[\underline{o}] = [\underline{z}^*] \quad ,$$

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 $\underline{z}$ and $\underline{z}^*$, given that the measurement characteristics 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 two types of measurement restrictions, termed measurement level and measurement process. The process restrictions concern the relationships among all the observations <u>within</u> a single category, whereas the level restrictions concern the relationships among all the observations <u>between</u>

different categories. The measurement implications of the restrictions are summarized in Table 2, and the restrictions are shown in Table 3.

There are two types of 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 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 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

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

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

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

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

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

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}:	$(o_i \sim o_m) \rightarrow (z_i^- = z_m^-) \underline{<} \begin{Bmatrix} z_i^* \\ z_m^* \end{Bmatrix} \underline{<} (z_i^+ = z_m^+)$
Ordinal	t^{do}:	$(o_i \sim o_m) \rightarrow (z_i^* = z_m^*)$ $(o_i \prec o_m) \rightarrow (z_i^* < z_m^*)$	t^{co}:	$(o_i \sim o_m) \rightarrow (z_i^- = z_m^-) \underline{<} \begin{Bmatrix} z_i^* \\ z_m^* \end{Bmatrix} \underline{<} (z_i^+ = z_m^+)$ $(o_i \prec o_m) \rightarrow (z_i^* < z_m^*)$
Numerical	t^{dp}:	$(o_i \sim o_m) \rightarrow (z_i^* = z_m^*)$ $z_i^* = \sum_{q=0}^{p} \delta_q o_i^q$	t^{cp}:	$(o_i \sim o_m) \rightarrow (z_i^- = z_m^-) \underline{<} \begin{Bmatrix} z_i^* \\ z_m^* \end{Bmatrix} \underline{<} (z_i^+ = z_m^+)$ $z_i^* = \sum_{q=0}^{p} \delta_q o_i^q$

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.

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 t 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.

For nominal variables, there are no level restraints: The characteristics of nominal variables are completely specified by the process restraints. Since there are two types of processes, there are two types of nominal variables; discrete-nominal and continuous-nominal. The discrete nominal variable is quite common, with the sex of a person being such a variable. It is clear that this is a nominal 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).

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:

(4) $t^{o}: \ (o_i < o_m) \rightarrow (z^*_i < z^*_m)$

where the superscript on t^{o} indicates the order restriction, and where $<$ 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 (t^{do}) then tied observations remain tied after transformation, whereas for continuous-ordinal (t^{co}) variables tied observations may be untied after transformation. The discrete-ordinal case is well examplified 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:

(5) $t^{p}: \ z^*_i = \sum_{q=0}^{p} \delta_q o_i^q \ .$

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 relationship 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 equation (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 equation (5) just a bit. The former case corresponds to the discrete-interval or discrete-ratio case in which we allow no within 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.

2.3 Geometrical interpretation

Figure 3 presents the geometric relations among the model, data and optimal scaling spaces, as well as the parameter space. Note that the model, data and optimal scaling spaces are pictured as all being components 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 each of the missing observations (if there are any).

We emphasize that the parameter space is not part 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

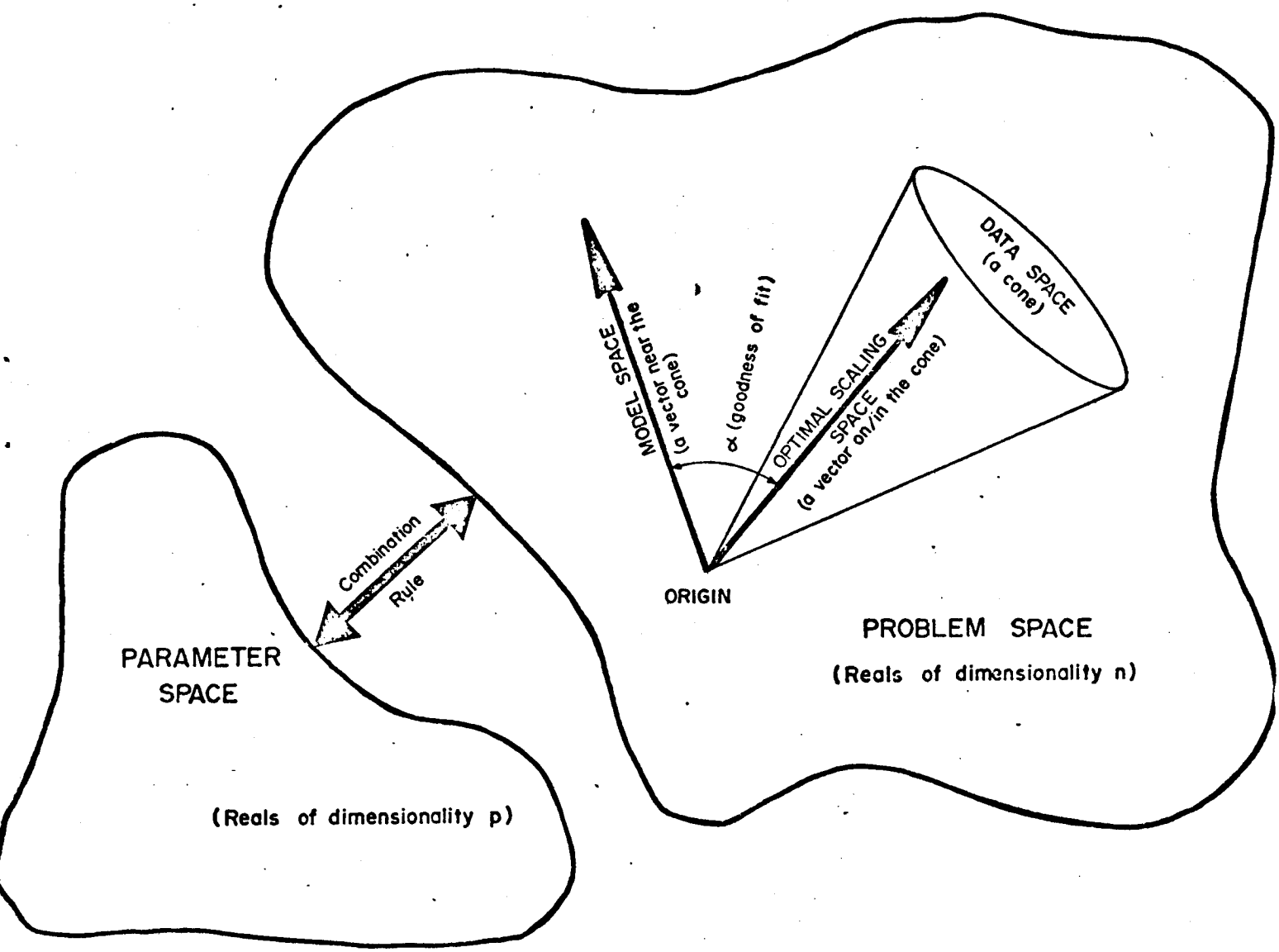


Figure 3: Geometry of a data theory.

representing the parsimony of description inherent in the model. As shown in the figure, the parameter and model spaces are related by a rule for mapping from one space to the other, a rule which we call the combination rule (Young, 1975). More will be said about this later.

In the problem space we have represented the model and optimal scaling spaces as vectors and the data space as a cone. Furthermore, the two vectors and cone all intersect at the origin of the problem space. We choose the type of representation for each of the three spaces for specific reasons. We represent the optimal scaling space as a vector running through the origin to emphasize the fact that the elements of $\underline{z}^*$ define a point in the problem space, and that if we form the vector which connects that point to the origin of the problem space then all of the other points on the vector are equivalent to $\underline{z}^*$ at the ratio level of measurement. In terms of the restrictions discussed above, any point in the optimal scaling space in Figure 3 is equivalent to any other point. We represent the model space as a vector for the same type of reasons.

On the other hand, we represent the data space as a cone, not a vector. Although the representation is different, the reasoning underlying the representation is the same: For the data space a cone properly represents the measurement characteristics, whereas for the model and optimal scaling spaces a vector is the proper representation. If you reflect on the restrictions given in equations 2 through 5, you will see that 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). You will note that the optimal scaling vector is represented as being on the surface of the cone. Since the optimal scaling and data spaces 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 spaces 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 space. (The only time that the optimal scaling vector is inside the cone is when the model space also happens to be in the cone, which only happens when the model perfectly fits the data.)

Finally, note the angle α between the model and optimal scaling spaces. The angle α represents the goodness-of-fit between the two spaces, 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 space consisting entirely of zeros. In this case the fit between the model and optimal scaling spaces is perfect ($\alpha=0$), but 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. Generally, such solutions are avoided by normalizing the length of the model and optimal scaling vectors to some arbitrary non-zero length.

3. Quantification with Unknown Models: Methods

3.1 Introduction

As stated above, the goal of an optimal scaling algorithm is to derive a space of optimally scaled data which has two characteristics: First, it must perfectly satisfy the measurement characteristics of the data space; and second, it must have a least squares relationship with the model space, given that the measurement characteristics are perfectly satisfied. In this section we discuss the methods used to obtain the optimal transformations t^{d}, t^{o}, t^{do}, t^{co}, and t^{p}.

Each of these transformation methods satisfies the stated measurement characteristics and is least squares. Thus each method is an example of our expanded definition of optimal scaling. The methods are summarized in Table 4.

3.2 Methods

For the two nominal level transformations t^d (discrete-nominal) and t^c (continuous-nominal) the estimation process is very simple and, at least for t^d, quite well known (Fisher, 1938, pp. 285-298). The t^d procedure consists, simply enough, of defining an element z_i^* as the mean of all the z_i which correspond to observations o_i in a particular category. Since the z_i^* are the mean of their corresponding z_i, we obtain a least squares fit given the restrictions placed by the measurement characteristics on t^d (Eq. 2). Formally, z_i^* is stated, under the discrete-nominal restriction, as

$$(6) \qquad t^d: \ \underline{z}^* = \underline{U}(\underline{U}'\underline{U})^{-1}\underline{U}'\underline{z}$$

where $\underline{U}$ is a binary matrix with a row for every observation and a column for every observation category. The elements of $\underline{U}$ indicate category membership:

$$(7) \qquad u_{ic} = \begin{cases} 1 \text{ iff } o_i \in \text{category } c \\ 0 \text{ otherwise} \end{cases}$$

The continuous-nominal situation is a bit more complex. The added complexity is introduced because the continuous-nominal situation, as discussed to this point, involves no measurement restrictions. For t^c (Eq. 3) we just have the requirement 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.

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

Naturally, the process proposed in the previous paragraph is meaningless. Therefore, we propose an alternative process which involves additional restrictions on the relationships between the intervals. Specifically, we propose a procedure which yields non-overlapping contiguous intervals, thus disallowing the trivial circumstances outlined in the previous paragraph.

The continuous-nominal transformation t^c 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 using Kruskal's primary least-square monotonic transformation. Note that in neither phase do we actually assume that the data are continuous-nominal. However, the assumptions which 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. This order is then used to help define interval boundaries. Two things should be noted: First, the procedure outlined here yields a least squares quantification which is consistent with, but slightly stricter than, the continuous-nominal restrictions specified in Eq. 3. Specifically, the procedure yields non-overlapping intervals, whereas the restrictions specified by Eq. 3 would permit overlapping intervals. Second, the procedure outlined here is <u>not</u> the same as the pseudo-ordinal procedure discussed by de Leeuw, Young & Takane (1976), but is a newer procedure which

avoids the problems mentioned in that paper. Specifically, the new procedure does not suffer from the oscillations and discontinuities present in the former procedure.

The two ordinal transformations t^{do} (discrete-ordinal) and t^{co} (continuous-ordinal) are defined by Kruskal's least squares monotonic transformation. Our discrete process corresponds to his secondary procedure, and our continuous process to his primary procedure. Young (1975) has shown that both thansformations may be formally stated as

(9) $\quad t^{o}: \ \underline{z}^* = \underline{U}(\underline{U}'\underline{U})^{-1}\underline{U}'\underline{z}$.

In the continuous-ordinal case $\underline{U}$ is a binary matrix indicating the $\underline{z}$ which must be tied to satisfy the ordinal restrictions, and in the discrete-ordinal case $\underline{U}$ is a binary matrix indicating the $\underline{z}$ which must be tied to satisfy both the ordinal and categorical restrictions. Kruskal (1964) has shown that the discrete-ordinal case is least squares, and de Leeuw (1975) has shown that the continuous-ordinal case is least squares.

The least squares solution for $\underline{z}^*$ under the restrictions of the t^{p} transformation is well known. The t^{p} transformation can be written in matrix notation as

(10) $\quad t^{p}: \ \underline{z}^* = \underline{U}\delta$

where $\underline{U}$ is a matrix with a row for each observation and with p+1 columns, each column being an integer power of the vector $\underline{o}$ of observations. The first column is the zero'th power (i.e., all ones), the second column is the first power (i.e., is $\underline{o}$ itself), the third column is the squares $\underline{o}^2$, etc. The least squares estimate of $\underline{z}^*$ is

(11) $\quad t^{p}: \ \underline{z}^* = \underline{U}(\underline{U}'\underline{U})^{-1}\underline{U}'\underline{z}$.

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

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

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

(13) $\underline{z}^* = \underline{E}\underline{z}$.

When we formally note that the least squares notion is defined (under suitable normalization conditions) as

(14) $\phi^2 = (\underline{z}^*-\underline{z})'(\underline{z}^*-z)$

and when we define $\underline{F} = \underline{I}-\underline{E}$, then we see that

(15) $\phi^2 = \underline{z}'\underline{F}\underline{z}$

emphasizing the fact that each of the transformations can be viewed as optimizing a relationship between the model space and some linear combination of the very same model space, where the linear combination is determined by the measurement restrictions. This point has been emphasized in a more restricted situation by Young (1975), and was first noted in the present context by Takane, Young & de Leeuw (1976). Geometrically, the projection operator projects the model space $\underline{z}$ onto the nearest surface of the data space cone (see Figure 3).

3.3 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 been generated by a single measurement device. Furthermore, with some of these sets of data the measurement device generates data in such a way that all of the observations are reasonably assumed to be on the same measurement scale. 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 "device" (the subject) and as having been generated on a single scale (the rank order of the similarity judgments). However, for other types of data it is clearly the case that the data are generated by several measurement devices, or on several scales. For example, when we obtain measurements about sex, age, hair color, income, educational background and political preference from a set of people, we would probably think of each measurement variable as being derived from a unique measurement device. In this case we would wish to partition the data space into a set of mutually exclusive and exhaustive subspaces (one for each variable) whereas in the first case we would simply view the entire data space as a single space. 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 in our view a situation in which a single measurement device (the subject) generates n different measurement scales (the rank orders). For this type of data the notion of measurement partitions is 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 for each partition separately, one partition at a time (see Figure 4). 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 with the procedures oriented towards multivariate data (MORALS, CORALS, & PRINCIPALS) we can analyze data

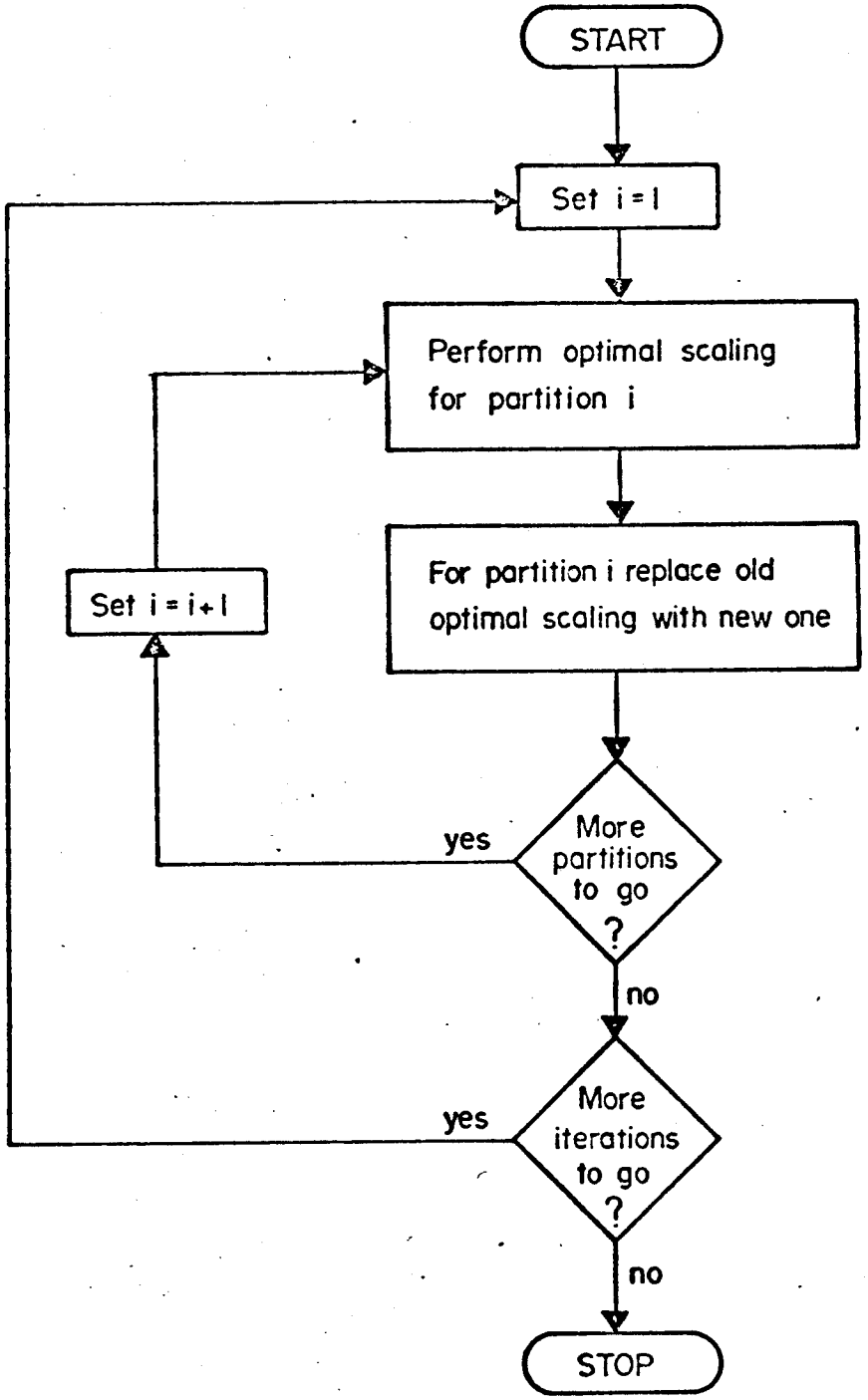


Figure 4: Flow of optimal scaling with partitioned data.

with any mixture of measurement characteristics. There is a very important consideration here, however, which must not be overlooked. It is sometimes imperative that right after performing the optimal scaling for a particular subset 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 subsets are not independent (where independent will be defined later). Such independence is, in fact, not generally a characteristic of multivariate data, thus in the programs which analyze such data the replacement is made immediately. This point has been emphasized in Young, de Leeuw, & Takane (1975).

If, in fact, the partitions are not independent, 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 not independent. The quantification obtained by optimally scaling one variable depends on the quantifications of each of the other variables. While this 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 (as in Figure 4), 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 has no effect

on the eventual convergence point. Since performing only one iteration (instead of iterating to convergence) can be viewed as a type of relaxation procedure, it may be that the improved overall efficiency is related to the same factors which often times cause relaxation procedures to be more efficient than non-relaxation procedures (Hageman & Porsching, 1975).

4. Quantifying Data Having Unknown Measurement Characteristics

It was stated above that one of the chief 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 of combining ALS and OS is that the model estimation phase of an ALSOS algorithm does not need to know anything about the measurement characteristics.

The practical effect of this aspect of an ALSOS procedure 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 discussed in the previous section 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 two things, the initialization process and the inner iteration structure. It is possible that two different types of initialization procedures or inner iteration structures 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, 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 & Takane (1976) 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 of the programs in the ALSOS system, we define "best initialization" to mean that we should clearly optimize something relevant to the problem being solved. We take this something to be the fit of the model to the raw data. Thus, each of the ALSOS programs is initiated by applying the numerically least squares procedure to the raw data under the assumptions that the raw data are quantitative.

The procedures for obtaining the conditional least squares estimates of the model parameters are the familiar procedures used to obtain ordinary least squares estimates when the data are numerical. The only difference is that the procedures are applied to the vector $\underline{z}^*$ of optimally scaled data (which is numerical, after all) instead of to the vector $\underline{o}$ of raw observations. Since we are applying the numerical model estimation procedure to the optimally scaled data and not to the raw 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.

The procedure for ADDALS, the algorithm for applying the simple additive model (i.e., no interaction terms) to qualitative data is an excellent example of the simplicity of the model estimation process in an ALSOS algorithm (de Leeuw, Young & Takane, 1976). The procedure for obtaining the best estimates for the parameters of the additive model

$$z_{ijk} = \alpha_i + \beta_j + \gamma_k + \mu \tag{16}$$

(where we have reorganized the previous model vector $\underline{z}$ with element z_i into a three-way table with element z_{ijk}) is very well known: We obtain row means to estimate α_i, column means to estimate β_j, plane means to estimate γ_k and the grand mean to estimate μ. The only difference is that we base the means on the optimally scaled datum z^*_{ijk} instead of the raw datum o_{ijk} (both now represented in tabular format). It is important to note that with the additive model we do not need to define any type of inner iterations, even when the data are incomplete. Thus, for the additive model the conditional global optimum is conditional only on the values used to initialize the algorithm.

Most of the rest of the procedures that we have developed on the ALSOS principle are equally simple in the model estimation phase. The MORALS and CORALS procedures apply the multiple or canonical regression model to mixed measurement level multivariate data (Young, Takane, & de Leeuw, 1976). The model estimation phase is identical to a standard multiple or canonical regression algorithm, except that it is applied to the optimally scaled data, not the raw data. The PRINCIPALS procedure applies the principal components model to mixed measurement level multivariate data (Takane, Young & de Leeuw, 1976), and involves nothing more than a standard eigenvalue decomposition of the optimally scaled data in the model estimation phase. We should emphasize

that each of these procedures are applied to multivariate data, and that the ramifications of such partitioned data (discussed in section 3.3) apply. In particular, for each of the procedures mentioned in this paragraph the conditional global optimum is conditional on both the initialization procedure and on the inner iteration structure. If we were to permit each inner iteration to procede to convergence, as suggested in section 3.3, then the conditionality would only be upon the initialization procedure. Our experience is, however, that such an approach is very time consuming and has no essential effect on the convergence point. Thus, we have opted to only use a single inner iteration.

The only procedure which involves a fairly complicated model estimation phase is the ALSCAL algorithm for performing individual differences multidimensional scaling (Takane, Young & de Leeuw, 1976). 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:

$$(17) \qquad z_{ijk} = \sum_{a=1}^{t} v_{ia} w_{ka} (x_{ia} - y_{ja})^2$$

where, again, z_{ijk} is a tabular reorganization of the model space vector $\underline{z}$, 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 Baker, Young & Takane (1976) (Young, 1975b) asymmetric model, w_{ka} are subject weights of the individual differences model discussed by Carroll & Chang (1970) and Horan (1969),

x_{ia} are stimulus-object points in a Euclidian 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 we need to know the values of one of the sets of parameters in order to derive the best estimate of another set of parameters. When parameters are not independent the values of the parameters in one set effect 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" iteration which estimates parameters, one set at a time. Thus, for ALSCAL, which is based on the model in Eq. 17, 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 is iterated until convergence is obtained.

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 again 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 non-independent data partitions apply in precisely the same fashion to non-independent 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 space 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 point at a time, under the assumption that each of the coordinates of all remaining points are constant. This innermost iteration involves n phases, one for each of the n points.

We have gone into fairly great detail concerning the ALSCAL algorithm because it involves an important source of complexity 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" may 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 (1976) for ALSCAL are independent) just as parameters are not always dependent. However, when dependence exists the LS inner iteration approach is a viable approach to deal with the problem.

5. 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 will have relatively few difficulties with local minima.

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 straight forward 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 iterations. Furthermore, in some circumstances there are some indeterminacies of construction which may have great effect on the overall speed of the algorithm(such as in 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, only on the conditional global 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 would conclude, however, that the ALSOS approach to algorithm construction is both more flexible and equally or more robust than previous approaches to quantifying qualitative data.

References

Baker, R.F., Young, F.W. & Takane, Y. An asymmetric multidimensional scaling model (in preparation) 1976.

Benzecri, J.P. L'analyse des donnees - Tome II: Correspondances DUNOD, Paris, 1973.

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

Bouroche, J.M., Saporta, G., Tenenhaus, M. Generalized canonical analysis of qualitative data. Paper presented at the U.S. Japan Seminar on Theory, Methods and Applications of multidimensional scaling and related techniques, 1975.

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 multidimensional 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. Leider, The Netherlands: University of Leiden, Data Theory, mimeographed paper, 1975.

de Leeuw, J., Young, F.W. & Takane, Y. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. Psychometric Laboratory Report #140, University of North Carolina, 1975 (and Psychometrika, in press, 1976).

Fisher, R. Statistical Methods for Research Workers. (10th ed.) Edinburgh: Oliver and Boyd, 1938.

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.

Nishisato, S. Optimal scaling and its generalizations. Department of Measurement and Evaluation. Ontario Institute for Studies in Education. I 1972 - II 1973 - III 1975.

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

Takane, Y., Young, F.W. & de Leeuw, J. How to use PRINCIPALS. Unpublished users manual, 1975.

Takane, Y., Young, F.W. & de Leeuw, J. Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. Psychometric Laboratory Report #147, University of North Carolina, 1975 (and Psychometrika, in press, 1976).

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. Methods for describing ordinal data with cardinal models. Journal of Mathematical Psychology, 1975, 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. Psychometric Laboratory Report #146, University of North Carolina, 1975.

Yule, G.U. An Introduction to the Theory of Statistics. London: Griffin, 1910.