



Additive Structure in Qualitative Data:

An Alternating Least Squares Method
with Optimal Scaling Features

J. de Leeuw, F. W. Young, and Y. Takane

Report Number 140

July, 1975

THE L. L. THURSTONE
PSYCHOMETRIC LABORATORY
UNIVERSITY OF NORTH CAROLINA

CHAPEL HILL, N. C.
27514

Abstract

A method is developed to investigate the additive structure of data which may a) be measured at the nominal, ordinal or cardinal levels, b) be obtained from either a discrete or continuous source, c) have known degrees of imprecision, and d) be obtained in unbalanced designs. The method also permits experimental variables to be at the ordinal level of measurement. It is shown that the method is convergent, and includes several previously proposed methods as special cases. Both Monte Carlo and empirical evaluations indicate that the method is robust.

This research was supported in part by grant MH-10006 from the National Science Foundation to the Psychometric Laboratory of the University of North Carolina. We wish to thank Thomas S. Wallsten for comments on an earlier draft of this paper. Copies of the paper may be obtained from the second author.

0: SUMMARY

In this paper we consider ways to obtain additive representations of data structures. This problem is not new of course, having a long history under the misnomer "analysis of variance", and we are not so presumptuous as to consider all aspects of the problem. Rather, we focus our efforts on a particularly robust way to obtain additive representations for qualitative data structures.

Even this problem is not new. As early as 1938 (Fisher, 1938, pp. 285-298) proposed an eigenvector method for applying the simple additive model to categorical data, a method which has been rediscovered periodically over the years (Hayashi, 1952; Carroll, 1967; Nishisato, 1972, 1973). More recently Kruskal (1965) proposed a gradient procedure for investigating the additive structure of ordinal data (see also Roskam, 1968; deLeeuw, 1969; Lingoes, 1973). Our work is strongly related to deLeeuw's (1973) discussion of methods for analyzing categorical data and by Young's (1972) alternating least squares method for finding additive structure in ordinal data.

Our work is placed in a theoretical framework from which flows an elegant and simple method for investigating additive structure in qualitative data, including as special cases all the methods mentioned in the preceding paragraph. The data may be either nominal, ordinal, a mixture of both, or quantitative. Furthermore, with quantitative data the categories may represent an underlying process which is either discrete or continuous, an important theoretical and practical distinction which is seldomly discussed in this context. It is also very simple, within our

framework, to introduce constraints on the parameters of the additive model. Thus, for example, it is quite simple to specify ordinal constraints for some factor in a design, if there is a priori reason to do so. Finally, our framework allows us to investigate observations arising in certain unbalanced, incomplete factorial designs. If, for example, we have a replicated factorial design, but have been unable to obtain an equal number of observations in all cells of the design, our developments can still be applied.

1. INTRODUCTION

The analysis of additivity has usually been introduced in the context of a statistical model for factorially classified observations, requiring assumptions that are often very strong and unrealistic. In many situations much less specific models are called for, based on much weaker assumptions. We discuss the classical assumptions briefly.

In stochastic versions of the analysis of additivity, one analyzes a model whose assumptions are

$$S_1: \tilde{y}_{ij} = \gamma + \alpha_i + \beta_j + \tilde{\varepsilon}_{ij},$$

S_2 : the $\tilde{\varepsilon}_{ij}$ are independent random variables,

S_3 : the $\tilde{\varepsilon}_{ij}$ have a centered normal distribution with finite variance σ^2 .

(A tilde under a symbol is used to distinguish random variables from fixed constants). Model S generalizes in a straightforward way to incomplete and/or replicated multi-factor situations, in which the number of indices and of corresponding sets of parameters is larger. (In order to avoid cumbersome notation we shall only treat the two factor case in this paper. The generalizations to more complicated factorial designs are obvious.)

Observe that S does not say that there are parameters γ , α_i , β_j such that each additive combination $\gamma + \alpha_i + \beta_j$ is close to the corresponding \tilde{y}_{ij} ; it merely makes a statement about the two-way structure of the expectations $E(\tilde{y}_{ij})$. The variance σ^2 can be arbitrarily large, and if it is unknown (which is the usual case) we can only test hypotheses about the parameters within S (i.e., while assuming S to be true). In many cases S itself is not very reasonable, the trouble being that the parametric assumption S_3

is too strong in many applications, and even the independence assumption S_2 is often not obviously true.

Within the framework of established statistical theory, the logical step out of these difficulties would seem to make weaker, nonparametric assumptions. A straightforward extension of S involves the following nonparametric assumptions:

$$N_1: y_{ij} = \gamma + \alpha_i + \beta_j + \varepsilon_{ij},$$

N_2 " the ε_{ij} are independent random variables,

N_3 : the ε_{ij} have a centered, centrally symmetric, continuous distribution with finite variance.

Unfortunately, the statistical theory based on the assumptions of this model is fragmentary and from the point of view of data analysis inferior to that based on model S .

In the case of model S the natural estimation method and the optimal way of testing hypotheses follow directly from elementary properties of the model. The method of least squares should be used, and the orthogonality properties of the complete factorial design lead to additive partitionings of the sums of squares. These properties lead to optimal tests of hypotheses and are very valuable for summarizing some of the important structures in the data. Model N , on the other hand, leads to robust significance testing and estimation, but the properties of the tests and estimates are usually only approximately known, and the beautiful structure of a complete least-squares analysis is lost. For a summary of some of the results that can be obtained we refer to Puri and Sen (1971).

Another basic complication is that in many applications even the assumption S_1 or N_1 cannot be applied because the observed data

are qualitative. That is, they consist of a small number of categories for which no precise numerical values are known. This not only violates the assumption of a continuous distribution, but it also makes S_1 and N_1 meaningless because y_{ij} is not defined. In this paper we reformulate the basic structural assumption S_1 or N_1 in such a way that it also applies to categorical data. For this purpose we use the notion of optimal scaling (Fisher, 1938; Buttman, 1941; Burt, 1950; Bock, 1960; Nishisato, 1972; de Leeuw, 1973). We shall assume that the data are in K mutually exclusive and exhaustive categories. We define the K -ary random variables z_{ij}^k which are equal to one if the observation in cell (i,j) of the design is in category k , and equal to zero otherwise. The model we employ is, in this simple case,

$$D_1: \sum_{k=1}^K z_{ij}^k \theta_k = \gamma + \alpha_i + \beta_j + \varepsilon_{ij}.$$

Observe that we have introduced the optimal scaling parameters θ_k , which we use to quantify each of the k categories. It is through restrictions on the optimal scaling parameters θ_k that we can treat qualitative (as well as quantitative) data. If we do not know precise numerical values for the observations we can represent each unique observation by a parameter θ_k and try to parametrize the data (as well as the model) to optimize the fit between the two. (Naturally, there must be fewer categories than observations, or we will have a perfect, but trivial fit). Since we wish to work in the familiar least-squares framework we measure the fit of a particular arbitrary choice of parameters by a suitably normalized version of the loss function

$$\lambda = \sum_{i=1}^n \sum_{j=1}^m \left(\sum_{k=1}^K z_{ij}^k \theta_k - \gamma - \alpha_i - \beta_j \right)^2,$$

and the computational problem is to choose the parameters θ_k , γ , α_i , and β_j in such a way that λ is minimized.

In the several cases we will discuss not all vectors of real numbers are admissible as parameter vectors: That is, the admissible values for θ_k , α_i , β_j , and γ may be subject to certain restrictions. Through these restrictions we cope with a variety of measurement levels. For example, if the data are measured at the ordinal level, then we restrict the value of $\theta_\ell < \theta_k$ if we know that the corresponding data categories stand in this relation. As another example, if we know a priori that the levels of some factor (say factor I) have ordinal properties, then we can restrict the estimate of $\alpha_1 < \alpha_2$, if that is the desired order. Other types of useful parameter restrictions will be discussed in the body of the paper, but we should always keep in mind that our goal is to optimize, within the least-squares framework, the relationship between a possibly restricted set of model parameters α_i , β_j , and γ and a possibly restricted set of optimal scaling parameters θ_k .

An important difference between this approach and the one based on either models S or N is that we have no guarantee that our estimates will be "good" estimates according to any of the accepted statistical criteria. We merely compute estimates, and afterwards we can try to find out how they behave under various more-or-less specific assumptions about the distribution of the z_{ij}^k . Rather than estimate the parameters of a model in the usual sense, we study the properties of a particular transformation or reduction of the data (cf. also de Leeuw, 1973, Chapter I, for

more extensive discussion of the difference between the two approaches).

We use a computational method for optimizing λ which we call additivity analysis by alternating least squares (ADDALS). This is an iterative method which alternates between a) minimizing λ over all admissible optimal scaling parameters θ_k for fixed values of the model parameters α_i , β_j , and γ ; and b) minimizing λ over all admissible model parameters for fixed values of the optimal scaling parameters. In each of the two phases of an iteration the optimization is complete; that is, the values obtained for one of the sets of parameters absolutely minimize the function λ conditional on a fixed set of parameters. Thus, the name alternating least squares: We alternate between two phases, one of which determines the (conditional) least squares estimates for the optimal scaling parameters and the other of which determines the (conditional) least squares estimates for the model parameters. This type of procedure is philosophically much like the NILES/NIPALS procedure developed by Wold and his associates (Wold & Lyttens, 1969) with the distinction that Wold is usually concerned with optimizing only model parameters. The class of procedures used by Wold and by us is known in the mathematical programming literature as block relaxation or nonlinear Gauss-Seidel methods. Although our procedure always converges to a stationary point it may not be the most robust one for each of the special situations outlined above. Thus, we compare our method with others which have been suggested for some of the special cases, with generally satisfactory results. As will be seen, the iterates are very simple (yielding an algorithm which

may be used on small machines) and very quick (enabling the analysis of large problems on large machines).

2: DATA THEORY

In this section we outline the data theory in which the developments of this paper are embedded. This section is divided into three subsections, concerned with the empirical model and measurement aspects of the data theory.

2:1 Empirical aspects

For the sake of simplicity and clarity, we restrict our formal developments to the case where there are only two conditions (called by others, factors, independent variables, components, dimensions, facets, classifications, etc.). The first condition has n levels (values, elements, structs), the second condition has m levels. We shall assume that each combination of levels (cell, structuple) is replicated R times, an assumption which will be relaxed shortly. Finally, we view the experimental design as being the cartesian product of all the conditions and the replication factor.

An assumption fundamental to our work is that an observation is a discrete entity which belongs to a particular observation category. Specifically, an observation is said to be in the same category as another observation if they are indistinguishable from each other in terms of their observational characteristics (other than the time and place of observation). Note that the several categories are mutually exclusive and exhaustive subsets of the entire set of observations. There are K observation categories in total.

This view of the basic nature of the data allows us to recode the data in a binary form indicating the category membership of each observation. The resulting binary matrix, called

the indicator matrix, has one column for each observation, and one row for each level of each experimental condition, as well as one row for each observation category. Thus, in our situation there are Rnm columns, and $n+m+K$ rows. The rows of the matrix are partitioned into three subsets, as follows. The first set of n rows indicates the level of the first experimental condition, the second set of m rows indicates the level of the second experimental condition, and the last set of K rows indicates the category membership of the observation. The first set of rows is denoted as the submatrix U , the second set as the submatrix V , and the last as the submatrix Z . The column corresponding to the replication r of cell (i,j) has elements equal to unity in row i of submatrix U , in row j of submatrix V , and, if the observation is in category k , in row k of submatrix Z . An example of observations obtained in a 3×2 factorial with two replications and three response categories is presented in Table 1a, with the corresponding indicator matrix in Table 1b.

It is now possible to indicate how we relax the assumption that there are an equal number of replications of each cell. If a particular observation is missing (i.e., there are fewer replications of a particular cell) then we simply assign it to its own unique category. That is, the missing observation is coded in the indicator matrix as belonging to a separate category of which it is the only member. Thus, every missing observation gets assigned to a unique observation category. This shows, at least as far as the coding of the data is concerned, that the assumption of a perfectly balanced factorial design is not a restriction of generality. We shall see later on how the coding conventions affect the results of our analysis.

Table 1a

	Replication			
	1		2	
Condition 1	Condition 2			
level	1	2	1	2
1	1	2	1	2
2	3	1	1	2
3	1	2	1	1

Table 1b

Cell	(1,1) (1,2) (2,1) (2,2) (3,1) (3,2)						(1,1) (1,2) (2,1) (2,2) (3,1) (3,2)					
cond 1	1	1	0	0	0	0	1	1	0	0	0	0
	0	0	1	1	0	0	0	0	1	1	0	0
	0	0	0	0	1	1	0	0	0	0	1	1
cond 2	1	0	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1	0	1
observ	1	0	0	1	1	0	1	0	1	0	1	1
	0	1	0	0	0	1	0	1	0	1	0	0
	0	0	1	0	0	0	0	0	0	0	0	0

2:2 Model aspects

The model involves concepts which parallel those involved in the empirical situation. Corresponding to the two experimental conditions are two vectors of parameters. Just as each condition has levels, each parameter vector has elements, denoted α_i and β_j (we use Greek characters for parameters). There is no notion in the model which corresponds to the empirical notion of replications, since we assume that any differences which arise between replications are random fluctuations not included in the model. (If we were in fact interested in modeling these fluctuations then we would view the "replications" factor as an additional experimental condition). Finally, there is a direct correspondance between the experimental design and the model. Whereas the former involves the cartesian product of all the experimental conditions and the replication factor, the latter involves the factorial combination of all the parameter vectors. For both the cartesian product and the factorial combination we define two real valued functions which generate the data and model spaces, respectively. Thus, the model space is defined by

$$\delta: C_{ij} = \alpha_i + \beta_j.$$

and the data space by

$$t: y_{ij} = \sum_{k=1}^K z_{ij}^k \theta_k.$$

In matrix notation these definitions are

$$\delta: C = U\alpha + V\beta,$$

$$t: Y = Z\theta$$

Finally, as mentioned above, we wish to parameterize the two spaces so that they are as much alike as possible. This objective is realized in the usual way of minimizing the sum of

squared error terms. Thus, we wish to minimize (subject to normalization)

$$\lambda = \sum_{i=1}^n \sum_{j=1}^m \left(\sum_{k=1}^K z_{ij}^k \theta_k - \alpha_i - \beta_j \right)^2,$$

or in matrix terminology

$$\lambda = (Z\theta - U\alpha - V\beta)'(Z\theta - U\alpha - V\beta),$$

by judicious assignment of values to the parameters of the two spaces. The minimization is subject to constraints which we may place on the parameters. These constraints are discussed in the next section.

2:3 Measurement aspects

In this section we discuss those restrictions which optionally may be placed on the data and model parameters. It is through these restrictions that we treat the variety of measurement conditions under which the observations may have been obtained, including the level and precision of measurement, the nature of the process which may have generated the observation, and the measurement characteristics of the experimental conditions themselves. We distinguish three types of parameter restrictions, identification restrictions, model restrictions, and data restrictions, and discuss them in turn.

2:3.1 Identification restrictions

Note that the model

$$c_{ij} = \alpha_i + \beta_j$$

can be written as

$$c_{ij} = \gamma + \alpha_i + \beta_j,$$

with α_i and β_j restricted in such a way that

$$\sum_{i=1}^n \alpha_i = \sum_{j=1}^m \beta_j = 0.$$

These constraints merely serve to identify the model parameters, since without them we can add a constant to all α_i and subtract the same constant from all β_j without affecting the fit. We shall always impose these constraints, but they must be distinguished from other types of constraints which go beyond the basic specifications of the model and data spaces.

2:3.2 Model restrictions

There are two types of optional restrictions which may be placed on the permissible values of α_i and β_j and may be appropriate in certain situations. One type of restrictions is invoked when we know that the levels of one (or both) of the experimental conditions fall in some a priori order. In such a situation we should restrict the corresponding model parameters (α_i or β_j) to be in the desired order. That is, we desire that the parameters be restricted to fall in a known convex cone.

The other type of restriction applies when we know that the levels of an experimental condition are related to each other in some clearly specified functional manner, for example by a linear or polynomial function. In this situation the parameter vector should be restricted to be a function of some fixed and known vector. Here we require that the parameter vector belong to some p -dimensional subspace, with $p < n-1$ (or $m-1$ as the case may be). Note that a p -dimensional subspace is a particular type of convex cone.

2:3.3 Data restrictions

The restrictions on the optimal scaling parameters θ_k are somewhat more complex than the α_i and β_j restrictions presented

in the previous section. These restrictions fall into two classes which are factorially combined to produce six types of data which differ in terms of their measurement characteristics.

The first class of restrictions is concerned with the measurement level of the data, and is precisely the same as that discussed in the previous section. That is, there are convex cone restrictions on θ_k when the data are known to be ordinal, and subspace restrictions on θ_k when the data are numerical. Just as with the model parameters, the data parameters may also be unrestricted which, when combined with the process restrictions discussed in the next paragraph, implies that the observations are measured at the nominal level.

The second class of restrictions on the optimal scaling parameters θ_k corresponds to our assumptions about the process which generated the observations. If we believe that the process is discrete then we restrict all the observations in a particular category to be represented by a single, discrete number. Thus in this case the optimal scaling parameter θ_k is a single number for each K , which may be viewed as a point on the real number continuum. On the other hand, if we believe that the process is continuous then we define θ_k to be an interval of numbers so that all the observations in a particular category are represented by an interval of numbers.

By factorially combining the three level of restrictions (no restrictions, cone restrictions, and subspace restrictions) with the two process restrictions (point and interval) we obtain six types of restrictions on the parameterization of θ_k , which correspond to six different types of measurement, as follows.

When we combine the "no" level restrictions with either one of the two types of process restrictions we obtain two different forms of what are commonly called nominal data. The point process restrictions are appropriate to data defined at the nominal level. In this case all observations in a given category are assigned a single number, with there being no restrictions between the various categories. We call this well known case the discrete-nominal case. On the other hand, when interval restrictions are invoked we obtain permissible parameterizations of θ_k which are appropriate to what we call continuous-nominal data. Here we assign a range of numbers of observations in each category, with no restrictions between categories. Obviously, the requirement that all observations in a category must be quantified by an interval is much too weak, as any arbitrary quantification always satisfies the restrictions if the category intervals are wide enough. Thus, we need to specify additional restraints. One possibility for achieving meaningful and non-trivial boundaries is to view the supposedly continuous-nominal data as actually being continuous-ordinal (to be discussed in a moment), but with the order of the categories unknown. The restrictions are of the cone type, but they are not convex cones. In fact the permissible region is the union of a finite number of convex cones.

When we combine the cone level restrictions with either of the process restrictions we obtain the two commonly discussed forms of ordinal data, corresponding to how tied observations are handled. The cone-point combination yields what we call discrete-ordinal data for which all observations in the same

category are to be parameterized by the same number. This is what Kruskal (1964) has called the secondary approach to ties. The cone-interval combination yields what we call continuous-ordinal data, for which all observations in a particular category are represented by an interval of numbers. Kruskal calls this the primary approach to ties.

When we combine the subspace level restrictions with either of the process restrictions we obtain a measurement level which corresponds to two forms of numerical (quantitative, cardinal) data. What is most commonly thought of as numerical data is obtained when the point process restriction is combined with the subspace level restriction, since in this case all observations which are equal (i.e., in the same category) remain equal (are parameterized by a single θ_k) and all observations which are not equal (in different categories) are functionally related. On the other hand, when we combine the subspace and interval restrictions we obtain a form of numerical data whose measurement characteristics take into consideration the precision of measurement, since in this case each observation is functionally related to every other observation within a certain degree of tolerance, the degree being specified by the width of the interval around each observation. Note that there is a subtle difference between the present usage of interval restrictions and the previous usage. Whereas previously we assumed that the boundaries of the intervals were determined internally (i.e., according to the nature of the data and model), we now assume that the boundaries are specified externally before the data are analyzed. Thus we assume that the researcher can specify an upper boundary θ_k^+ and

a lower boundary θ_k^- on each observation category. Generally for data defined at the interval level of measurement there is but one observation in each category so we usually are specifying an interval for every single observation. If we use these intervals to specify the precision of measurement then in many situations we will wish to specify an interval of constant width for all observations, with the midpoint of the interval being equal to the observation. That is, we only need to specify θ^Δ from which we can determine $\theta_k^+ = \theta_k + \theta^\Delta$ and $\theta_k^- = \theta_k - \theta^\Delta$. There are other interesting uses of the continuous-numerical parameter restrictions. For example, external boundary constraints can be used to impose nonnegativity (by setting $\theta_k^- = 0$ and $\theta_k^+ = \infty$) or other types of range restraints. External boundary constraints can also be used to impose constancy on certain portions of the data by setting $\theta_k^- = \theta_k^+ = p_k$, where p_k is a known constant.

3: METHOD

In this section we present the alternating least squares (ALS) method which obtains estimates of the optimal scaling parameters θ_k and the additive model parameters α_i and β_j which optimize λ . In the first subsection we discuss the decompositions of the function λ from which flow the ALS procedure as applied to the additive model (the ADDALS algorithm). In the next subsection we discuss parameter restrictions and their least squares implementation in ADDALS. In the third section we outline the ADDALS algorithm for finding the jointly optimal (restricted) parameterization of the model and data spaces and prove the convergence of the algorithm under all restrictions except the pseudo-ordinal restrictions. In the fourth section we show that a) the ADDALS algorithm is equivalent to the analytic method proposed independently by Fisher (1938), Hayashi (1952), Carroll (1969) and Nishisato (1972) for discrete-nominal data; b) the ADDALS algorithm is essentially equivalent to the MONANOVA algorithm proposed by Kruskal (1965) for ordinal data (discrete or continuous); c) the ADDALS algorithm is equivalent to the widely used ANOVA methods for analyzing discrete-numerical data; and d) the ADDALS algorithm is equivalent to the widely used procedure proposed by Yates (1933) to solve for the optimal values of missing discrete-numerical data. Finally, it is observed that ADDALS obtains least squares parameter estimates in a wide range of other situations for which, to the authors' knowledge, least squares methods have not been previously proposed.

3.1: Decompositions

We now introduce the index r for replications explicitly into our equations, by defining the quantified observations as

$$r^{y_{ij}} = \sum_{k=1}^K r^{z_{ij}^k} r^{\theta_k}$$

in the unpartitioned case, or

$$r^{y_{ij}} = \sum_{k=1}^{K(r)} r^{z_{ij}^k} r^{\theta_k}$$

in the partitioned case (the number of categories need not be the same for each replication). From the familiar theory of the analysis of variance we copy the decompositions of $r^{y_{ij}}$ into orthogonal components. We use dots to indicate indices over which we have averaged. The decomposition we use is

$$r^{y_{ij}} = .y_{..} + (.y_{i.} - .y_{..}) + (.y_{.j} - .y_{..}) + (.y_{ij} - .y_{i.} - .y_{.j} + .y_{..}) + (r^{y_{ij}} - .y_{ij}).$$

We define

$$\begin{aligned} \hat{\mu} &= .y_{..}, \\ \hat{\alpha}_i &= .y_{i.} - .y_{..}, \\ \hat{\beta}_j &= .y_{.j} - .y_{..}, \\ \hat{\gamma}_{ij} &= .y_{ij} - .y_{i.} - .y_{.j} + .y_{..}, \\ r^{\hat{\epsilon}}_{ij} &= r^{y_{ij}} - .y_{ij}, \\ r^{\hat{\delta}}_{ij} &= r^{\hat{\epsilon}}_{ij} + \gamma_{ij}. \end{aligned}$$

Observe that all these quantities depend on the θ_k , but we suppress this dependence to keep the notation simple.

It is well known that $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$ are least squares estimates of the corresponding parameters in the model

$$r^{y_{ij}} = \mu + \alpha_i + \beta_j + r^{\delta}_{ij},$$

i.e., they minimize the sum of squares of the residuals $r\delta_{ij}$. The corresponding minimum residuals are, of course, precisely $r\delta_{ij}$. In the same way $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$ are the least squares estimates in the model

$$r y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + r \epsilon_{ij},$$

and $r\hat{\epsilon}_{ij}$ is the corresponding minimum residual. Although we are really only interested in the first model (any departure from simple additivity is assumed to be error) it is sometimes informative to decompose the residual into a systematic interaction and error term.

In ordinary analysis of variance the decomposition of $r y_{ij}$ into orthogonal components defines an additive decomposition of the sum of squares of the $r y_{ij}$ into components, each of which is the sum of squares of one component of the $r y_{ij}$. In this paper we use the same orthogonality properties to partition our loss functions

$$\lambda = \sum_{r=1}^R \sum_{i=1}^n \sum_{j=1}^m (r y_{ij} - \mu - \alpha_i - \beta_j)^2$$

into loss function components corresponding to each subset of the parameters. The relevant partition is given in Table 2.

In the case in which the parameters are not restricted in any sense minimization can obviously be accomplished by minimizing each of the components over the relevant subset of the parameters. This makes each of the three deviation components equal to zero because we set $\mu = \hat{\mu}$, $\alpha_i = \hat{\alpha}_i$, and $\beta_j = \hat{\beta}_j$. In the constrained case a similar result is true if the constraints on the parameters are separated (there are constraints on α , constraints on β , and no constraints that involve both α and β). Thus, the overall minimization problem separates into a number of simpler minimization

Table 2

deviation from optimal mean	$Rn (\hat{\mu} - \mu)^2$
deviation from optimal row scores	$Rm \sum_{i=1}^n (\hat{\alpha}_i - \alpha_i)^2$
deviation from optimal column scores	$Rn \sum_{j=1}^m (\hat{\beta}_j - \beta_j)^2$
SUBTOTAL: deviation from optimal parameterization	$\sum_{r=1}^R \sum_{i=1}^n \sum_{j=1}^m \{(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) - (\mu + \alpha_i + \beta_j)\}^2$
optimal minimum loss	$\sum_{r=1}^R \sum_{i=1}^n \sum_{j=1}^m ({}_r \hat{\delta}_{ij})^2$
total loss for given parameterization	$\sum_{r=1}^R \sum_{i=1}^n \sum_{j=1}^m ({}_r y_{ij} - \mu - \alpha_i - \beta_j)^2$

subproblems. As mentioned previously we are only interested in the additive model in this paper, and the decomposition of the $r\delta_{ij}$ into an interaction term γ_{ij} and an error term $r\epsilon_{ij}$ is not really relevant. It is obvious, however, that Table 2 can be modified very easily to include the interaction parameters. In de Leeuw, Young & Takane (1975) we have done this, and have discussed restrictions on the interactions of form which has recently been studied extensively in the statistical literature (for example, Corsten and van Eynsberger, 1972).

To derive the second decomposition of our loss function we define

$$\hat{y}_{ij} = \mu + \alpha_i + \beta_j,$$

and (in the unpartitioned case)

$$\hat{\theta}_k = (M_k)^{-1} \sum_{i=1}^n \sum_{j=1}^m \hat{y}_{ij} \sum_{r=1}^R r^z_{ij}^k,$$

with

$$M_k = \sum_{r=1}^R \sum_{i=1}^n \sum_{j=1}^m r^z_{ij}^k.$$

Clearly, M_k is the total number of observations in category k , and $\hat{\theta}_k$ is the average \hat{y}_{ij} value of the observations in this category. Consequently, $\hat{\theta}_k$ is the unrestricted least squares estimator of θ_k for given μ , α , β . Note that $\hat{\theta}_k$ is a function of μ , α , and β but we suppress this dependence to simplify the notation. The additive partition of λ , corresponding to the problem of minimizing the loss over θ for fixed α , β , μ is given in Table 3.

We can use this last partition of the total sums of squares to illustrate our technique for handling missing data and unbalanced designs. Remember that each missing observation has its

own category, and the corresponding category score θ_k is unrestricted. This means that the optimal score for the category equals the corresponding \hat{y}_{ij} value, and the missing cell does not contribute to the loss at all. Minimizing λ over our artificially balanced design is equivalent to minimizing a loss function that is the sum of squares of the deviations of data and model values in non-missing cells only. This is true for obtaining either θ_k or α_i and β_j .

3:2 Use of restrictions

In this section we discuss the implementation of the most important types of restrictions on the parameters in the two computational subproblems (minimizing λ for fixed θ over α, β, γ and minimizing λ for fixed α, β, γ over θ).

For the first problem it may be that we know, a priori, an appropriate order for the levels of I or J, and therefore may desire to restrict the parameters, for example, so that $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_n$, and/or $\beta_1 \leq \beta_2 \leq \dots \leq \beta_m$. Our first decomposition (Table 1) shows that the optimal α under these restrictions can be found by applying the familiar isotonic regression methods (Barlow, et. al., 1972; Barlow & Brunk, 1972). Actually, general partial orders on the α_i or the β_j could be incorporated in this way, but the following developments only cover the linearly ordered case, with Kruskal's (1964) two methods for incorporating ties. Although our developments are limited to ordinal restrictions on the model parameters, we could restrict the α_i and β_j in other ways. For example, the α_i (or β_j) could be required to be related by the linear function

$$\alpha_i = a + b\alpha_{i+1}$$

or by some other polynomial function. In such a case the decomposition shows that ordinary linear regression can be used to compute the least squares estimates of the linearly related α_i and β_j . (See de Leeuw, Young, & Takane, 1975, for developments of this notion.)

From the second decomposition (Table 3) it follows that explicit interval restrictions of the form $\theta_k^- \leq \theta_k \leq \theta_k^+$ with known θ_k^+ and θ_k^- (e.g., continuous-numerical data) can be handled very easily. If $\hat{\theta}_k$ is in the interval then the optimal θ_k is equal to $\hat{\theta}_k$. If $\hat{\theta}_k$ is outside the interval then the optimal θ_k is equal to the nearest endpoint of the interval (e.g., equal to θ_k^+ if $\hat{\theta}_k > \theta_k^+$ or equal to θ_k^- if $\hat{\theta}_k < \theta_k^-$). Cone restrictions on θ_k can be handled by monotone or linear regression again. Using the primary or secondary approach to ties takes care of continuous or discrete ordinal data, and of discrete categorical data (in the last case we set the optimal θ_k equal to $\hat{\theta}_k$).

Only continuous-nominal data present a problem. In the pseudo-ordinal case we want the optimal θ_k to fall into disjoint intervals, but the order of the intervals on the real line is unknown. The obvious best procedure is to try out all possible orders of intervals, compute the optimal θ_k by monotone regression with the primary approach for each interval order, and keep the best order to define the optimal θ_k for this iteration. This can lead to rather unpleasant computations if the number of categories is at all large, and it introduces severe discontinuities in our transformation, which would affect the convergence behavior of our algorithm. A second alternative (which is used in ADDALS) is to derive the optimal order of the intervals from the order of the $\hat{\theta}_k$. This will be a satisfactory approximation in most

Table 3

deviation from optimal
unrestricted quantification

$$\sum_{k=1}^K M_k (\hat{\theta}_k - \theta_k)^2$$

optimal minimum loss

$$\sum_{r=1}^R \sum_{i=1}^n \sum_{j=1}^m \left(\sum_{k=1}^K r_{ij}^k \hat{\theta}_k - \hat{y}_{ij} \right)^2$$

total loss for given
parameterization

$$\sum_{r=1}^R \sum_{i=1}^n \sum_{j=1}^m \left(\sum_{k=1}^K r_{ij}^k \theta_k - \hat{y}_{ij} \right)^2$$

cases. Again, discontinuities may present a problem and convergence is not assured, but we can fix the order of the intervals at the current optimum in the final iterations, and treat the data as continuous-ordinal in the remaining cycles. This guarantees convergence.

3:3 Convergence

In the previous section we showed that each of the two subproblems can be solved in a very elementary way. Of course, this still does not prove anything about the efficiency or convergence of the complete process of alternating subproblems one and two. Let us formalize this process somewhat. Define the metric projection $P_C(x)$ of a point x in a Euclidian space on a closed convex subset C in a same space by $y_0 = P_C(x)$ if and only if y_0 is the unique minimizer in C of $\|x-y\|$ (where the double bars indicate sums of squares). Another suggestive name for $P_C(x)$ is the nearest point in C to x . Suppose now that there are two closed convex sets C_1 and C_2 with associated projections P_1 and P_2 . We define an iterative process by starting with $k=0$ and some arbitrary y_0 , and proceed by first obtaining $P_1(y_k)$ (the nearest point in C_1 to y_k) and setting $x_k = P_1(y_k)$, and then obtaining $P_2(x_k)$ (the nearest point in C_2 to x_k) and setting $y_{k+1} = P_2(x_k)$. We then increase k and repeat. Convergence of this scheme can be proved by using available theorems in the literature. There are, in the first place, theorems dealing explicitly with cyclic projection on a finite sequence of convex sets. The most general results have been given by Gubin, Polyak, and Raik (1967). In the second place there are some general theorems dealing with convergence of block relaxation of convex functions. A representative

reference is Cea and Glowinski (1973). A useful convergence theorem for nonconvex functions (with statistical applications) is given by Oberhofer and Kmenta (1974). Finally there are a number of general convergence theorems for relaxation processes, of which the most familiar one is given by Zangwill (1969). It follows from these theorems that the sequence x_k converges to a fixed point x_∞ of $P_1 \circ P_2$, and that a fixed point of $P_1 \circ P_2$ is a point of C_1 nearest C_2 . Moreover, y_k converges to a fixed point y_∞ of $P_2 \circ P_1$, which is a point of C_2 nearest C_1 . Consequently, the distance between x_∞ and y_∞ is the minimum of all possible distances between x in C_1 and y in C_2 .

These results can be applied directly to the case in which there are interval restrictions on the θ_k , and cone restrictions on the α_i and β_j . If both θ_k and α_i, β_j are restricted by cone restrictions, however, the results are without value. Cones intersect at the origin, and often the origin is the only point in the intersection. The theorems quoted above prove that both θ and α, β converge to zero in this case, which is a trivial and undesirable result.

We reformulate our problem by specifying that we are only interested in solutions which are "normalized" in some sense. This normalization (an extra restriction on either θ or α, β, γ or both) is chosen in such a way that the trivial solutions are excluded, with the computations being only slightly more complicated. The remainder of this section analyzes the normalization problem in some detail. As a first natural normalized loss function we consider

$$\mu = \frac{\|x-y\|}{\|y\|}$$

which has to be minimized over x in C_1 and y in C_2 , with C_1 and C_2 two convex cones. For fixed y in C_2 we still find the minimum of μ over x in C_1 by computing $P_1(y)$, but the problem of finding the minimum over y in C_2 for fixed x in C_1 seems more complicated. It has been proven, however, by Kruskal and Carroll (1969), that the solution of this subproblem is still proportional to $P_2(x)$. Moreover, the alternative normalized loss function

$$\zeta = \frac{\|x-y\|}{\|x\|}$$

is connected to μ by the simple relationship

$$\min_{y \in C_2} \mu = \min_{y \in C_2} \zeta$$

for all values of x . Consequently, using ζ instead of μ does not make any difference. If we combine the results of Kruskal and Carroll with the fact that for any convex cone C it is true that $P_C(\alpha x) = \alpha P_C(x)$ for all $\alpha \geq 0$, we find the important result that our previous alternating projection procedures also minimize the same subproblems for normalized loss functions, at least if we normalize the solutions in the proper way. Moreover, the normalizing can be done whenever we want to; it is not necessary to normalize after each iteration, although we do. Finally, it does not matter which of the two natural normalizations of λ we use, the results in each iteration will differ only by a proportionality factor, and the ultimate solutions will always be identical. Observe that an equivalent formulation of the normalized problem is the maximization of $x'y$ under the condition $x \in C_1$, $y \in C_2$, and under the normalization conditions $\|x\|=1$ and $\|y\|=1$.

This shows that we minimize the angle between the vectors x and y in their cones, without paying attention to their length. An alternative elementary proof of the Kruskal and Carroll results, with applications to ALS, is given by de Leeuw (1975). Convergence for normalized iterations follow in the same way as before from the general convergence theorems for relaxation processes.

The structure of the iterative process is now obvious. A program is available from the second author.

3:4 Relation to earlier work

If the convex sets C_1 and C_2 are linear subspaces, then the projectors P_1 and P_2 are independent of x and y . Thus, there are orthogonal projection matrices A and B such that

$$y_{k+1} = Bx_k = BAy_k,$$

and

$$x_{k+1} = Ay_{k+1} = ABx_k.$$

It follows that in this case our ALS method is equivalent to the power method for computing the dominant eigenvalue and corresponding eigenvector of BA and AB . Since the method proposed by Fisher (1938), and rediscovered by Hayashi (1952), Carroll (1969), and Hishisato (1972) finds the eigenvalue/eigenvector pair of the same matrices, it is clear that ALS is equivalent to these methods. In this specific case the previous methods are more efficient. It must be emphasized that some authors compute further subdominant eigenvalues and eigenvectors of these matrices to obtain a multidimensional qualification in an ALS framework.

Our missing data technique has been proposed, in the case in which there are no constraints on the model parameters and the non-missing observations are known real numbers, by Yates (1933)

(see also Wilkinson, 1958). The iterative technique has also been used by some authors as a computationally convenient way to estimate parameters in unbalanced designs. It is easy to show that the technique solves the least squares method by an iterative method based on a regular splitting of the design matrix. The theory of such methods has been studied very recently by Berman and Plemmons (1974).

It is also interesting to study the relationship of ALS and gradient methods, since Kruskal (1965) has proposed a gradient method for continuous or discrete ordinal data, with no constraints on the model parameters. We first consider the general unnormalized problem of minimizing $||x-y||$ over $x \in C_1$ and $y \in C_2$. It is well known that the function

$$v(x) = \min_{y \in C_2} ||x-y|| = ||x-P_2(x)||$$

is continuously differentiable, with gradient vector $x-P_2(x)$. The gradient projection method (Levitin & Polyak 1966) sets

$$x^+ = P_1[x-K(x-P_2(x))],$$

with the step size K chosen in such a way that sufficient decrease of $v(x)$ is guaranteed. Levitin and Polyak show that $K=1$ is an admissible step size, and by setting $K=1$ in the update equation we find the ALS method $x^+ = P_1 \circ P_2(x)$. Thus, our ALS algorithm is a convergent gradient projection algorithm with constant step size. In the normalized case we find $v(x)$, such that

$$v(x) = \min_{y \in C_2} \frac{||x-y||}{||y||} = \frac{||x-P_2(x)||}{||x||},$$

which is continuously differentiable if $||x|| \neq 0$, with gradient

$$g(x) = ||x||^{-1} (x - P_2(x)) - v(x) x.$$

Again, we can choose the stepsize in a gradient projection algorithm in such a way that it becomes equivalent to ALS, except possibly for a different normalization of intermediate solutions. If one of the cones in the normalized problem is a linear subspace we can collect a basis for the subspace in T, and minimize

$$v(x) = \min_{y \in C_2} \frac{||Tx - y||}{||Tx||}$$

unconditionally over x. Kruskal's MONANOVA (1965) is the special case in which C_2 is the polyhedral convex cone of monotone transformations. In the same way as before we show that the iterations of ALS can be interpreted (up to proportionality factors) as gradient iterations, with a particular choice of the step size. In MONANOVA the step size is determined by a completely different procedure, which may or may not be more efficient.

In a paper dealing with another special case of our situation, Bradley, Katti & Coons (1962) define

$$u(y) = \min_x \frac{||Tx - y||}{||Tx||}$$

and minimize $u(y)$ over C_2 by a coordinate descent method. The relationship of this method and ALS is complicated, although the basic idea of decomposing the optimization problem in a cyclic sequence of simpler problems is the same for both methods. It follows from the convergence theory of the methods we have shown to be equivalent to our method that convergence of ALS in these cases is at most linear (and can degenerate to convergence of order zero in some cases). In the computational literature a large number of methods are available that can be used to speed

up convergence. In particular, our analysis shows that choosing a different step size in gradient projection methods corresponds to over or underrelaxing the ALS iterations. Our examples show that in some instances convergence of ALS is quite slow, and experimenting with a relaxation parameter may be quite useful.

4: RESULTS AND DISCUSSION

In this section we present the results of applying ADDALS to several sets of data whose structures have been investigated by methods which are special cases of ADDALS. For these data we expect our results to be very much like the previous results. We also present the results of ADDALS analysis of artificial data to evaluate other special ADDALS cases. We will first discuss nominal data, then ordinal, then numerical.

4:1 Nominal data

Due to the equivalence of the iterative ADDALS method and the analytic eigenvector method when the data are discrete-nominal it is unnecessary to determine whether ADDALS will behave robustly with artificial error-free data. It will. However, we should point out certain types of discrete-nominal data (with or without error) which do not yield results which are unique up to a linear transformation. An obvious example is data which consist of unique categories, i.e., for which there is only one observation in each category. For such data, any parameterization of α_i and β_j yields a perfect, but meaningless, solution. A necessary condition for a unique solution, then, is that one category contains at least two observations. This condition is by no means sufficient, however. Consider, for example, the 3 x 3 table with 3 observations in each of 3 categories:

A	A	A
B	B	B
C	C	C.

In this case the row effects are completely determinant and the column effects are only determined to be equal at all levels. As

another, more subtle example, consider the 3 x 4 table with eight observations categories:

A	D	B	E
B	E	C	F
C	F	G	H.

If these categories are assumed to be discrete (not continuous) then the rows are connected (since each shares categories with another row) but the columns are only partially connected (since column one shares categories only with column three, and column two only with column four). Thus, the rows are determined up to a linear transformation, but the columns are determined up to two separable transformations, one for columns one and three, and another for two and four, due to the fact that two and four share no categories with one and three. Thus, an important condition to obtain results defined at the interval level from discrete-nominal data is that all rows (columns) be connected by common categories. It does not seem to be necessary that a row (column) share at least one category with all other rows (columns), but rather that a row (column) share at least one category with a second row which shares a category with a third, etc. Of course, these are but examples, and we do not mean to imply that they represent a complete argument for a necessary, let alone a sufficient condition which must be met to obtain a quantitative analysis. In the case of replicated data, for example, the condition given above can undoubtedly be weakened.

We have found that ADDALS yields results which are within a linear transformation of those obtained by the analytic eigenvector procedure for discrete-nominal data which meet the necessary

condition given above. Fisher (1938, pps. 285-298) demonstrated his eigenvector method by analyzing data concerning twelve samples of human blood tested with twelve sera, where the observations were one of five chemical reactions (this is a balanced, unreplicated 12×12 factorial design with 5-category data assumed by Fisher and ourselves to be discrete). ADDALS obtained a solution with $\lambda = .5397$ in 8 iterations with a random start (the criterion to terminate the iterative process in this and all other analyses, unless otherwise stated, is that the improvement in λ^2 must be less than .0005). The ADDALS parameter estimates are related to Fisher's estimates by a perfectly linear transformation. Carroll (1969) demonstrated his CCM method (which is identical to Fisher's proposal) with data obtained in an experimental situation described by three variables: the wave form, modulation percentage and modulation frequency of a tone. The experimental design was a factorial $2 \times 3 \times 4$, balanced and unreplicated. The data analyzed by Carroll were the clusters into which each of the 24 tones were placed by a clustering program, there being 5 such clusters. Our analysis (assuming discrete process) yielded results indistinguishable from Carroll's analysis, except for a linear transformation ($\lambda = .4477$, 34 iterations, random start).

We now investigate the behavior of ADDALS via the use of an artificial example in which the true population values underlying the discrete-nominal observations are known. In Table 4a we present the population values for the example, and in Table 4b we present the observation categories (this is a 6×6 balanced design with 2 replications, and with 5 observation categories in the first replication and 3 different observation categories in

Table 4a
Population Values

		β_j					
		1	2	3	4	5	6
α_i	2	3	4	5	6	7	8
	4	5	6	7	8	9	10
	7	8	9	10	11	12	13
	9	10	11	12	13	14	15
	12	13	14	15	16	17	18
	14	15	16	17	18	19	20

Table 4b

Observation Categories

		j					
		1	2	3	4	5	6
i	1	A	A	A	B	B	B
	2	A	A	B	B	C	C
	3	B	B	C	C	C	C
	4	C	C	C	D	D	D
	5	D	D	D	D	D	D
	6	D	D	D	E	E	E
i	1	F	F	F	F	F	G
	2	F	F	F	F	G	G
	3	G	G	G	G	G	G
	4	G	G	G	G	G	G
	5	G	G	G	H	H	H
	6	H	H	H	H	H	H

the second replication). The population values are completely connected. From Table 4b you can see that we have introduced two types of systematic observation error. One of the types of systematic error is represented by the fact that the true values have been collapsed into a smaller number of observation categories, whereas the other type of systematic error is inconsistencies (between replications) of the observation categories. However, there is no random error (the true values can be ordered properly by the observation categories in each replication). These types of systematic errors are common types of observational error in practice.

In Figure 1 we plot the parameter estimates obtained by ADDALS ($\lambda = .3366$ in 12 iterations, random initial category values) against the true values (the letters indicate category membership). It is clear that the derived α_i are linearly related to their true values, though the β_j are not. In particular, the derived values of β_1 and β_2 are equal even though the true values are not. This anomaly is due to the fact that the corresponding columns of the observation matrix are identical. We note now that this effect carries through all the analyses of these data which are to be presented, and that a linear relation could be obtained with differing observation columns. Identical columns (or rows) of observations is of some concern, however, and should be treated with caution.

In the remainder of this section we investigate the behavior of ADDALS under the continuous-nominal assumptions. Actually, as noted above, the totally unrestricted form of the continuous-

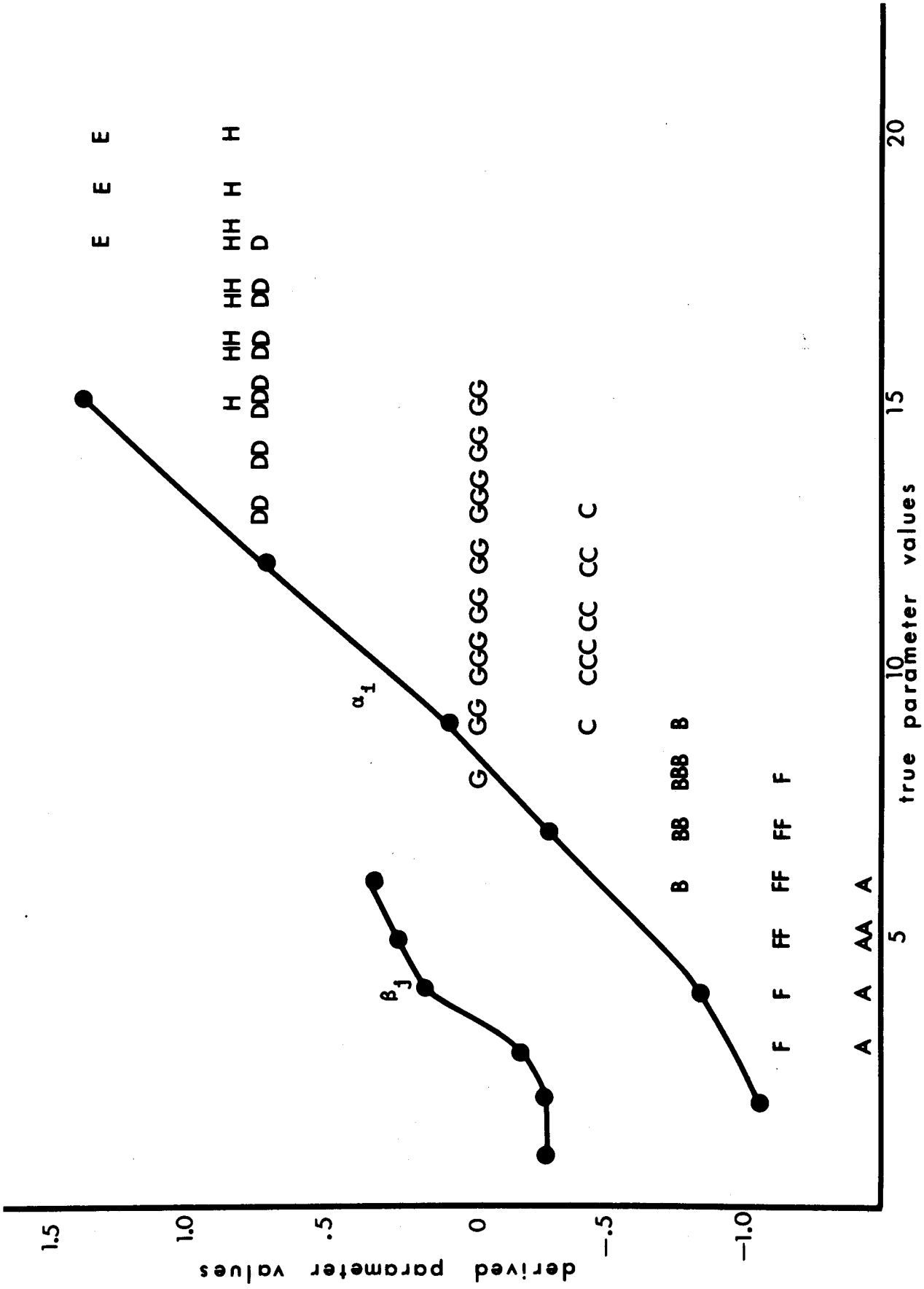


Figure 1. Artificial discrete-nominal example.

nominal assumptions are meaningless, so we impose the additional pseudo-ordinal restrictions discussed above, and then reanalyze the data in Table 4 under these restrictions. The plot of the parameter estimates versus the population values is presented in Figure 2. The solution ($\lambda = .1196$, 21 iterations, random initial category values) has ordered the categories in precisely the correct manner, and the solution is generally the same as that in Figure 1. It deserves emphasis, however, that a perfect solution has not been found, and that this is due to the nature of the systematic error. More specifically, if we look carefully at the inconsistencies between the 5 observation categories for the first replication, and the 3 observation categories for the second replication, we note that there is no order of all 8 observation categories which will permit a perfect solution, since observation category G corresponds with the true values ranging from 8 through 15, whereas category B has observations which correspond to true values as large as 9, and category D has corresponding true values as small as 13. Thus, we see that we must define a partial order of the categories in order to obtain a perfect fit ($\lambda=0$), the partial order being

$$\begin{aligned} A &\leq B \leq C \leq D \leq E, \\ G &\leq G \leq H, \\ F &\leq C, \\ G &\leq E. \end{aligned}$$

Since as stated, the pseudo-ordinal and ordinal assumptions do not permit partial orders, we cannot perfectly fit these data. Thus, if we were to now use the ordinal information developed by the pseudo-ordinal analysis to order all 8 categories, and

then use this information as the basis of a continuous-ordinal analysis, we should still arrive at precisely the same imperfectly fitting solution. Of course it would be relatively trivial to extend the notions of the pseudo-ordinal and ordinal types of measurement to include (pseudo) partial orders, and in fact we have done so in some other closely related work (Young, 1973; de Leeuw, Young and Takane, 1975). If we then reanalyze these data under the assumption that they represent a pseudo partial order, with the prior knowledge that the pseudo partial order consists of two partial orders (one for the first replication, and one for the second), then we should certainly obtain a perfect fitting solution, with the only question being the nature of the relationship of the solution to the true values. We have performed such an analysis using the multiple optimal regression by alternating least squares (MORALS) technique reported by de Leeuw, Young and Takane (1975) (which for orthogonal ANOVA designs is precisely equivalent to ADDALS, except for the ability of MORALS to handle partial orders). The procedure obtained a perfect fit (2 iterations, random start), with the derived parameter values being plotted versus the true values in Figure 3. We note that the figure indicates that the dependent variable and the values of α_i are essentially linear in their relationship to the true values, and that β_j still displays the same nonlinearities as before, but more mildly. The usefulness of such a procedure might be questioned since it assumes that we have prior knowledge about the nature of the partial order (it consists of two sub-orders). However, it is often the case that the observation categories in one replication of the experiment bear no simple relationship to the observation categories in another

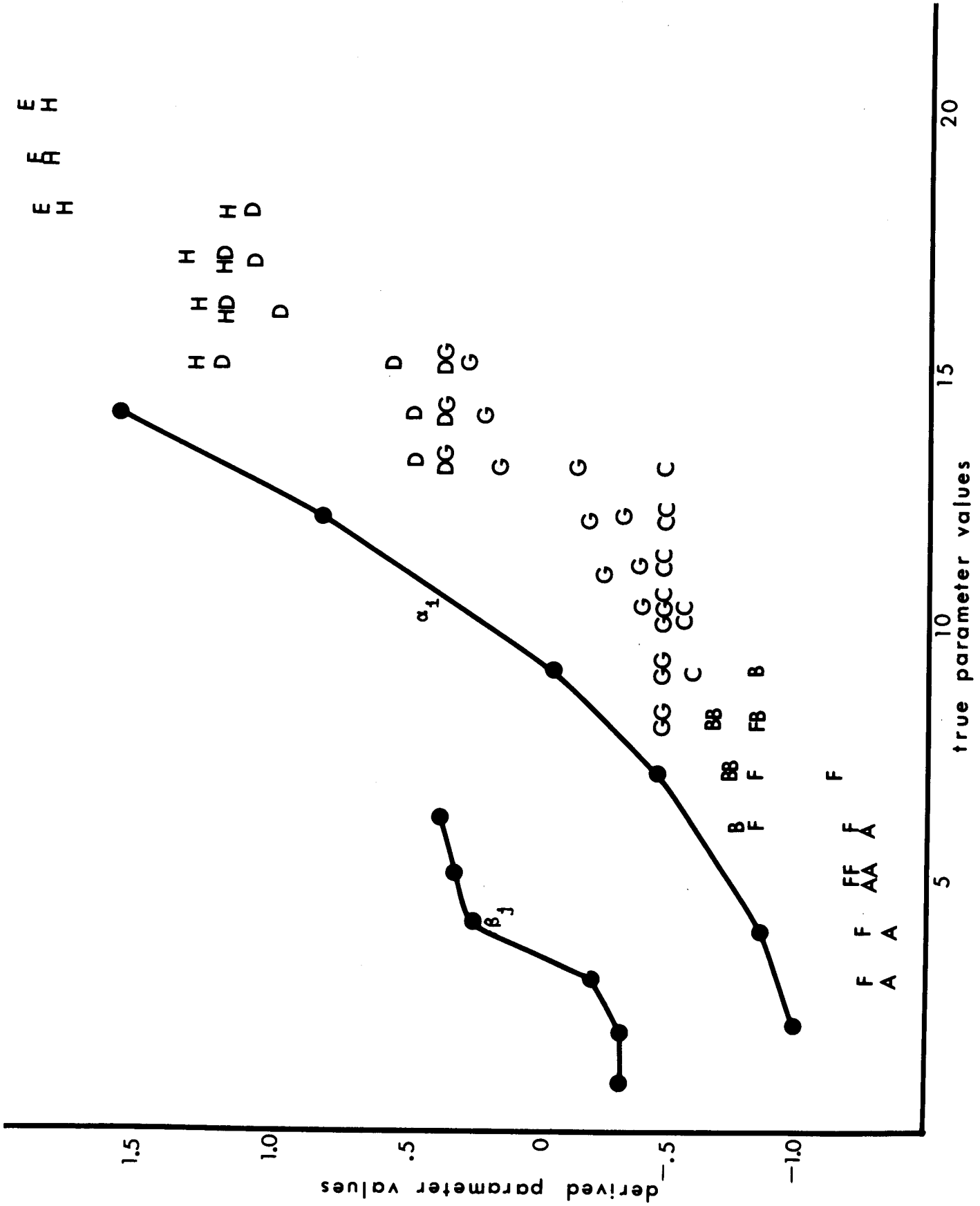


Figure 2. Artificial pseudo-ordinal example.

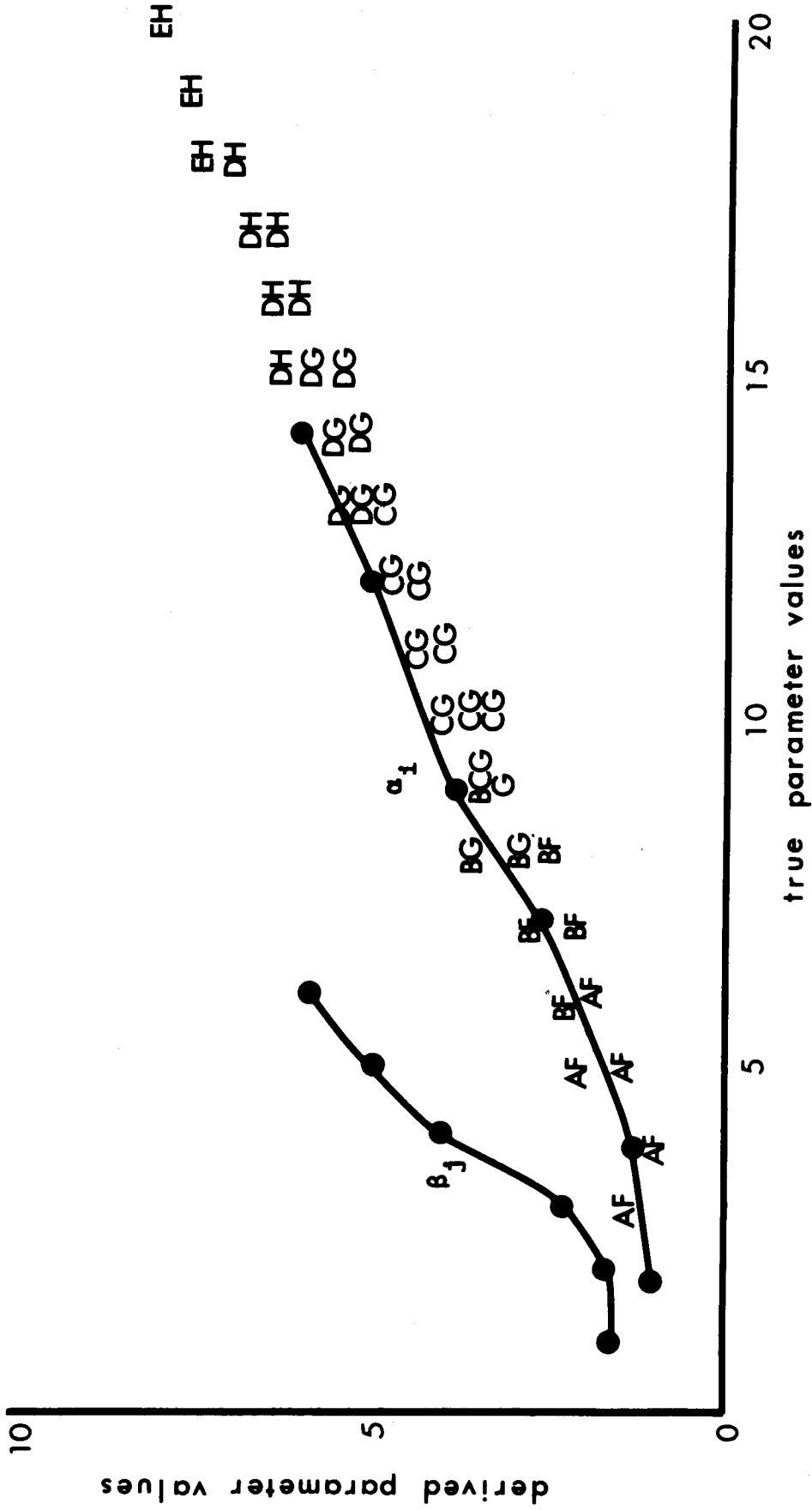


Figure 3. Artificial pseudo-partial-ordinal example.

replication. In such a situation the (pseudo) order really consists of several sub-orders, one for each replication.

We conclude, then, that under the appropriate conditions ADDALS can yield quantitative analyses of nominal data. It seems clear that one necessary condition is that all rows (columns) be connected by common categories, and it is probably the case that the number of observations should be large relative to the number of categories. For the latter reason it is desirable to have as many replications as possible. Finally, some care should be exercised when a) two or more (columns) are identical since this necessarily means the parameter estimates will be equal: and b) the data are pseudoordinal since the parameter restrictions are so weak.

4:2 Ordinal data

Our first ordinal example utilizes an artificial example discussed by Kruskal (1965) in his paper concerning MONANOVA. His 3 x 3 data are the squares of the "true" values obtained by the simple addition of the population row and column values. Thus, his data contain only systematic error. Furthermore, his population values have completely connected rows and columns. The ADDALS analysis of these data obtained a solution with $\lambda = .0000$ in 5 iterations (the discrete-ordinal assumption was used). Since this result might have been an artifact of the "rational start" (i.e., the observations were used to initialize the algorithm) we repeated it with a random start, obtaining $\lambda = .0000$ in 8 iterations. Both solutions are indistinguishable and are perfectly related to the underlying structure.

We felt that the results reported in the previous paragraph might be due to the strong connectedness of the data (and the assumption of discrete observations) so we analyzed a second set of 3 x 3 artificial discrete-ordinal data which have one unconnected column. The results of this analysis were essentially identical to those of the first analysis ($\lambda = .0000$, in 5 iterations from a rational start and 12 iterations from a random start, estimates perfectly related to true values). We pushed this notion even further by analyzing a third set of identical 3 x 3 discrete-ordinal data for which one row and one column are unconnected. In this case the analysis suffered, with the underlying structure not perfectly recovered (although $\lambda = .0000$ in 4 iterations for rational start). So, again, it is important to have connected rows and columns, especially for unreplicated matrices as small as the ones analyzed here. Of course, if we had assumed the data were continuous-ordinal our results would have been less encouraging for these 3 x 3 matrices, since this effectively disconnects any connections which may be present in the data. (We also performed all the previous analyses with Kruskal's MONANOVA and obtained indistinguishable results.)

Kruskal (1965) used several sets of real data to evaluate his procedure. We reanalyzed two of these sets to further evaluate ADDALS (both of these sets have also been analyzed by Box and Cox, 1964). The first of these two sets of data concern the strength of yarns (in terms of the number of cycles before failure) when the amount of load placed on the yarn, the amplitude of the load cycle, and the length of the piece of yarn are varied. Each of the three variables had three levels, and one observation was

obtained in each cell. Thus, this is a balanced, unreplicated 3 x 3 x 3 design. In keeping with Kruskal's analysis, we assume that the observations are continuous-ordinal and the experimental conditions are nominal. These data were submitted to ADDALS and to Kruskal's MONANOVA procedure. After 7 iterations, ADDALS had converged to a value of $\lambda = .071$, and after 8 iterations MONANOVA had converged to the same value. Both procedures obtained solutions identical up to a linear transformation.

The second set of Box & Cox data analyzed by Kruskal concern the survival time of animals subjected to one of three poisons and one of four treatments. These data were obtained from four animals in each condition, thus the experiment is a balanced, 3 x 4 design with four replications. The results of our analysis, which assumed that the observations were continuous and that the experimental variables were nominal, were compared with the results of Kruskal's analysis (which made the same assumptions) (ADDALS $\lambda = .3064$ on the sixth iteration; MONANOVA $\lambda = .3064$ on the eighth). Again, the results are virtually identical.

By removing some of the observations from these data we obtain an unbalanced design whose analysis can be compared with the analysis of the balanced design. Thus, we removed four of the 48 observations, one from each of the three cells involving the fourth level of the treatment variable, and one from cell 1,1. This leaves us with an unbalanced 3 x 4 design with four replications in eight of the 21 cells and three replications in each of the remaining four cells. When we compare the results of this analysis with those of the previous one, we see that the estimates have changed somewhat. We also note that the value of

(.2751 in 5 iterations) has decreased some from the balanced case, suggesting that its value is a function of the number of observations (as is the case in a closely related situation discussed by Young, 1970. Finally, we note the observations have been removed from the balanced design in such a way that two columns have no observations removed, one has one removed, and one has three removed. The number of observations removed is related to the degree of change in the corresponding parameter's estimate. Specifically, the column parameter estimate which changed the most is the one with the largest number of observations removed.

We now turn to two examples involving ordinal constraints on the experimental variables. Roskam (1968), in demonstrating his ADDIT procedure (which is nearly identical to Kruskal's MONANOVA) utilized a set of data gathered by Ekman (1965) concerning the average ratings of unpleasantness of an electrical shock whose intensity and duration was varied, involving 12 levels of each variable. We analyzed these data assuming that the experimental variables were ordinal and the measurement process was continuous-ordinal. When we compared our results ($\lambda = .0100$ in 9 iterations) with Roskam's (who was unable to assume ordinal effects, so treated them as nominal) we concluded that the two analyses were highly similar (all α_i and β_j were identical for both analyses except two values whose order was "incorrect" for the unrestricted analyses). This implies that the assumption of ordinal effects was appropriate, though unnecessary, and that it had no deleterious effects on the analysis.

As a second example of imposing ordinal constraints on the experimental variables, we analyzed data gathered by Kempler (1972) concerning the number of times each of 100 rectangles was judged to be either large or small by each of several subjects. The variables are the height and width of the rectangles, with 10 levels of each variable. We analyzed these data both with and without the ordinal constraints on the two experimental variables. Without ordinal constraints we (and Kempler) discovered a few inversions from the expected order. We note that the value of λ increased from .1558 for the unconstrained analysis (5 iterations) to .1565 for the constrained analysis (also 5 iterations), a very slight increase due to the restraints. Thus, this aspect of ADDALS allows us to observe that the best fitting constrained estimates (and their overall descriptive adequacy) are nearly as adequate as the free estimates.

Finally, we reanalyzed the artificial data in Table 4 under the assumption that the categories were continuous-ordinal, with the ordinal information being derived from the pseudo-ordinal analysis. The results were identical to those of the pseudo-ordinal analysis ($\lambda = .1196$, all parameters the same to four decimal places) with the only difference being that less iterations were required, due, apparently, to the non-random initial category values. Thus, this lends some credence to the pseudo-ordinal procedure. We also analyzed these data under the partial order assumptions discussed above, and obtained precisely the same solution as obtained with the pseudo partial order assumptions.

4:3 Numerical data

It is unnecessary, of course, to give an example of ADDALS applied to discrete-numerical data, since ADDALS reduces to computing row and column means of the data matrix in this case. Furthermore, with discrete numerical data which have missing observations ADDALS is equivalent to the iterative missing data technique proposed by Yates (1933), and there are many examples analyzed by this technique in the analysis of variance literature. Thus we do not discuss the discrete-numerical case, but turn instead to the continuous-numerical case.

We cannot compare our method with previous ones in the continuous-numerical case since we know of none, so we evaluate this case by analyzing a set of artificial data (in Table 5a we present the population values, in Table 5b the observation categories, and the category constraints). This example contains errors of observation similar to those in Table 4 (there are fewer observation categories than population values), but the range constraints are such that the population values constitute a perfect solution. Note that this example is quite strong in that all rows and columns of the population matrix are connected.

The parameter estimates obtained by the ADDALS analysis of these data are plotted against the population values in Figure 4. We observe that the estimates of the four row parameter α_i are, essentially, a perfect linear transformation of their population values, and we also observe that the estimates of the six column parameters β_j are related by the same linear transformation to

Table 5a
Population Values

		β_j					
		1	7	11	14	16	17
α_i	1	2	8	12	15	17	18
	2	3	9	13	16	18	19
	4	5	11	15	18	20	21
	8	9	15	19	22	24	25

Table 5b
Observations

		i					
		1	2	3	4	5	6
j	1	A	A	B	C	C	D
	2	A	A	B	C	D	D
	3	A	B	C	D	D	E
	4	B	C	D	E	E	E

Constraints

$$2 \leq A \leq 9 \leq B \leq 13 \leq C \leq 17 \leq D \leq 20 \leq E \leq 25$$

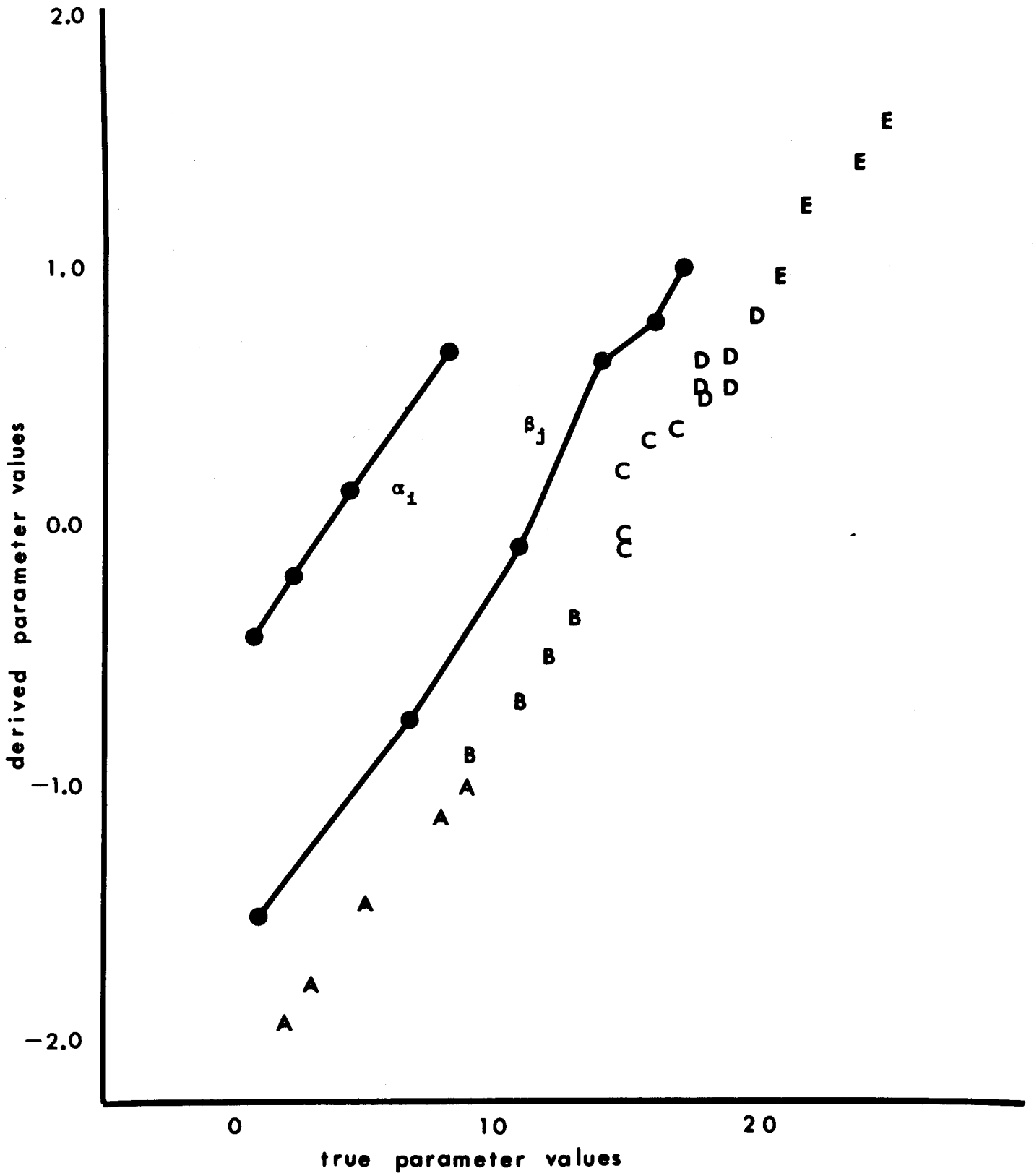


Figure 4. Artificial continuous-numerical example.

their population values, but that this latter relationship is not perfect (of course, when we plot the dependent variable we see the same linear, imperfect relationship). In particular, we note that the fourth largest column estimate is relatively imprecise. We are unsure why this is the case, but we do note that convergence is very slow for this example (38 iterations before the convergence criterion of .00005 was met), and that the solution, at this point, does not yet fit perfectly ($\lambda = .0050$). Perhaps if we had let ADDALS run for more iterations an improved solution would be obtained. We do feel, however, that this example indicates that with continuous numerical data ADDALS can behave in a relatively efficacious manner.

5: CONCLUSIONS

We conclude that the ADDALS approach enables one to quantify qualitative data via the application of the additive model (subject to conditions discussed in the previous section). Furthermore, we conclude that the associated algorithm is simple and efficient, in terms of both speed and size. We note that ADDALS includes, as special cases, the procedure first proposed by Fisher (1938) to analyze discrete-nominal data and the procedure first proposed by Kruskal (1965) to analyze both discrete or continuous-ordinal data. ADDALS can also be used to analyze 'ordinary' discrete-numerical data, and it includes a generalization of the procedure proposed by Yates (1933) for continuous-numerical data. ADDALS also includes the ability to apply the additive model to continuous-nominal data, to analyze data with an additive model which is subject to ordinal constraints on its parameters, and to analyze data when the experimental design is unbalanced. We know of no previous proposals which cover any of these last developments. Thus, we also conclude that ADDALS is a procedure which is much more general and flexible than previous proposals.

Finally, it is fairly simple to generalize the approach to models other than the simple additive model. Research recently completed suggests that the alternating least squares approach can be generalized in a straight-forward manner to other linear models. We have already developed robust (and rapid) ALS procedures to apply the multiple and canonical correlation models to nominal and ordinal variables (de Leeuw, Young, & Takane, 1975). Special cases of this procedure include procrustean rotation, external

unfolding, vector projection, additive models with interaction terms, nonorthogonal models, ADDALS, etc., etc. An ALS procedure has also been developed and evaluated for the bilinear model which includes nonmetric (and, of course, nominal) factor analysis, components analysis, etc., as special cases. At the time of this writing, this development appears to yield a robust and rapid method. Finally, we have extended the ALS methodology to the biquadratic models (the Euclidian and weighted Euclidian models) commonly used in multidimensional scaling (Takane, Young, and de Leeuw, 1975). Although this is considerably more complex than those just mentioned, it does appear to provide a promising alternative to the commonly used procedures. Thus, we find ALS methodology encouraging not only because of its ability to quantify qualitative data via application of the additive model, but also because of its promise to quantify qualitative data via application of a variety of other models.

References

- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. Statistical inference under order restrictions. London: Wiley, 1972.
- Barlow, R. E. & Brunk, H. D. The isotonic regression problem and its dual. Journal of the American Statistical Association, 1972, 67, 140-147.
- Berman, A. & Plemmons, R. J. Cones and iterative methods for best least squares solutions of linear systems. SIAM Journal of Numerical Analysis, 1974, 11, 145-154.
- Bock, R. D. Methods and applications of optimal scaling. University of North Carolina, Psychometric Laboratory Report No. 25, 1960.
- Box, G. E. P. & Cox, D. R. An analysis of transformations. Journal of the Royal Statistical Society, series B, 1964, 26, 211-252.
- Bradley, R. A., Katti, S. K., & Coons, I. J. Optimal scaling for ordered categories, Psychometrika, 1962, 27, 355-374.
- Burt, C. The factorial analysis of qualitative data. British Journal of Statistical Psychology, 1950, 3, 166-185.
- Carroll, J. D. Categorical conjoint measurement. Mimeographed paper, Bell Telephone Laboratories, Murray Hill, New Jersey, 1969.
- Cea, J. & Glowinski, R. Sur des methodes d'optimisation par relaxation. Revue Francaise d'Automatique, Informatique, et Recherche Opérationelle, section R3, 7, 1973, 5-32.

- Corsten, L. C. A. & van Eynsbergen, A. C. Multiplicative effects in two-way analysis of variance, Statistica Neerlandica, 1972, 26, 61-68.
- de Leeuw, J. Canonical analysis of categorical data. University of Leiden, Psychological Institute, 1973.
- de Leeuw, J. The linear nonmetric model. University of Leiden, Department of Data Theory, Report RN003-69, 1969.
- de Leeuw, J. Normalized cone regression. Mimeographed paper, Department of Data Theory, University of Leiden, 1975.
- de Leeuw, J., Young, F. W., and Takane, Y. Canonical regression with a mix of qualitative and quantitative variables. Paper presented to the Psychometric Society, May, 1975.
- Ekman, G. The influence of intensity and duration of electrical stimulation on subjective variables. Psychological Laboratory, University of Stockholm, Report 17a, 1965.
- Fisher, R. A. Statistical methods for research workers. Edinburgh: Oliver and Boyd, 1938 (7th printing), 1946 (10th printing).
- Gubin, L. G., Polyak, B. T., & Raik, E. V. The method of projections for finding the common point of convex sets. U.S.S.R. Computational and Mathematical Physics, 1967, 7, 1-24.
- 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.
- Hayashi, C. On the predictions of phenomena from qualitative data and quantifications of qualitative data from the mathematico-statistical point of view. Annals of the Institute of Statistical Mathematics, 1952, III, 69-92.

- Jöreskog, K. G. Some contributions to maximum likelihood factor analysis. Psychometrika, 1967, 32, 443-482.
- Kempler, B. Stimulus correlates of area judgments: a psychological developmental study. Developmental Psychology, 1971, 4, 158-163.
- 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. Nonmetric multidimensional scaling: a numerical method. Psychometrika, 1964, 29, 28-42.
- Kruskal, J. B. and Carmone, F. Use and theory of MONANOVA, a program to analyze factorial experiments by estimating monotone transformations of the data. Mimeographed paper, Bell Telephone Laboratories, Murray Hill, New Jersey, 1968.
- Kruskal, J. B. and Carroll, J. D. Geometric models and badness-of-fit functions. In: P. R. Krishnaiah (Ed.), Multivariate Analysis II. New York: Academic Press, 1969.
- Levitin, E. S. and Polyak, B. T. Methods of minimization under restrictions. U.S.S.R. Computational and Mathematical Physics, 1966, 6, 1-50.
- Lingoes, J. C. The Guttman-Lingoes nonmetric program series. Ann Arbor, Michigan: Mathesis Press, 1973.
- Nishisato, S. Optimal scaling and its generalizations, I: Methods, II: Applications. Department of Measurement and Evaluation. Ontario Institute for Studies in Education, Report I, 1972, Report 2, 1973.
- Oberhofer, W. and Kmenta, J. A general procedure for obtaining maximum likelihood estimates in generalized regression models. Econometrica, 1974, 42, 579-590.

- Puri, M. L. and Sen, P. K. Nonparametric methods in multivariate analysis. New York: Wiley, 1971.
- Roskam, E. E. Ch. I. Metric analysis of ordinal data in psychology. Voorschoten, VAM, 1968.
- Takane, Y., Young, F. W. and de Leeuw, J. Alternating least squares scaling. Paper presented to the Psychometric Society, May, 1975.
- Wilkinson, G. N. Estimation of missing values for the analysis of incomplete data. Biometrics, 1958, 14, 257-286.
- Wold, H. and Lythenns, E. (Eds.). Nonlinear iterative partial least squares (NIPALS) estimation procedures (group report). Bulletin of the International Statistical Institute, 1969, 43, 29-51.
- Yales, F. The analysis of replicated experiments when the field results are incomplete. The Empire Journal of Experimental Agriculture, 1933, I, 129-142.
- Young, F. W. Nonmetric multidimensional scaling: Recovery of metric information, Psychometrika, 1970, 35, 455-473.
- Young, F. W. A model for polynomial conjoint analysis algorithms. In R. N. Shepard, A. K. Romney, and S. Nerlove (Eds.), Multidimensional scaling: Theory and applications in the behavior-sciences. New York: Academic Press, 1972.
- Young, F. W. Conjoint scaling. L. L. Thurstone Psychometric Laboratory Report No. 118, 1973.
- Zangwill, W. I. Convergence conditions for nonlinear programming algorithms. Management Science, 1969, 16, 1-13.