NOTE: This manuscript was originally published in 1972 and is reproduced here to make it more accessible to interested scholars. The original reference is
Harshman, R. A. (1972). PARAFAC2: Mathematical and technical notes. *UCLA Working Papers in Phonetics*, *22*, 30-44. (University Microfilms, Ann Arbor, Michigan, No. 10,085).

PARAFAC2: Mathematical and Technical Notes

Richard A. Harshman*

0. Abstract

The mathematical model of PARAFAC is reviewed, and an examination is made of its application to cross-product matrices (e.g. covariance matrices,. scalar product matrices, etc.). It is shown that PARAFAC1 can correctly describe both orthogonal (uncorrelated) and oblique (correlated) factors in system-variation data matrices, but that it can only correctly describe orthogonal factors when applied to the *cross-product* matrices computed from that data.

A similar examination is made of the INDSCAL-PARAFACl model for three-mode multidimensional scaling. It is shown that this model is restricted to descriptions of stimuli relationships in terms of orthogonal perceptual dimensions, (whereas traditional two-mode multidimensional scaling has no such restriction).

A three-mode model is developed to deal specifically with sets of cross-product matrices. This model, called PARAFAC2, can describe cross-product matrices in terms of orthogonal or oblique factors, whichever best fits the data. Yet it retains PARAFAC1's highly desirable characteristics of providing unique "explanatory" solutions. Along with factor loadings on variables, it provides factor correlations, and loadings for each occasion of measurement. For multidimensional scaling, PARAFAC2 will recover the projections of the stimuli on oblique (or orthogonal) dimensions, give the angles between these dimensions, and provide a set of dimension weights for each person.

The derivation and interpretation of the PARAFAC2 model is discussed, and a precise definition is given of the type of uniqueness which it provides. No formal proof of this uniqueness has as yet been discovered, but the uniqueness has been demonstrated empirically by computer analysis of synthetic data.

An additional important advantage of PARAFAC2 for factor analytic applications is its greater generality: it is not restricted to analysis of system-variation data. This is shown by re-deriving the PARAFAC2 description of cross-products without invoking the restricted system-variation model for data from which the cross-products are computed. Instead, a more general model of the data is used. This model applies to almost all three mode cases which might be approached factor-analytically. System-variation is one special case of this more general model.

Current work on algorithms for fitting the PARAFAC2 model to data is reviewed. A "hybrid" approach is under development which combines general optimization techniques with estimation procedures based on the mathematical

* The author wishes to express his indebtedness to Robert Jennrich, Joseph Kruskal, and Dale Terbeek, for the important part that they played in developing the ideas underlying PARAFAC2.

properties of the PARAFAC2 model.

A method for circumventing the "communalities" problem is proposed. It consists of a technique for ignoring the diagonal values of the cross-product matrices and estimating the parameters of PARAFAC2 by optimizing the least square fit to the off-diagonal elements. The differences between analysis of covariance and correlation matrices is made explicit by developing the appropriate modifications of the PARAFAC2 model which would be necessary to analyze correlations.

PARAFAC2 is compared with more general models which lack unique solutions, but which allow different correlations between a given pair of factors from one cross-product matrix to the next. The possible use of a routine such as Carroll's IDIOSCAL program in conjunction with PARAFAC2 is discussed.

1. The factor model of PARAFAC1 (PARAFAC)

Conventional factor analysis describes a two-way set of data in terms of a small number of underlying factors. Let us think of this two-way set of data as a set of values for *m* persons on *n* tests, x_{jk} is the value for the *jth* test on the *kth* person. Factor analysis writes

(la) $$x_{jk} = a_{j1}b_{1k} + a_{j2}b_{2k} + \ldots + a_{j1}b_{1k} + E_{jk}$$

where a_{j1} is the loading of the first factor on test *j* and b_{1k} is the loading of the first factor on person *k*, and such products are formed for the *l* factors presumed to underlie the data. E_{jk} is an error term, which is necessary as the model should not be expected to fit real data exactly. (E_{jk} is often written U_{jk} and called "uniqueness" term. It is thought of as consisting of components unique to that person and that test, plus error.)

Let us neglect error terms until we consider estimation procedures. Using summation notation, we can write

(1b) $$x_{jk} = \sum_{r=1}^{l} a_{jr} b_{rk} \quad .$$

In matrix notation, we can consider A, an *n* by *l* matrix giving the loadings of *n* tests on *l* factors; and B an *m* by *l* matrix giving the loadings of *m* persons on *l* factors. We can then write the classical factor model (neglecting error) as

(lc) $$X = AB'$$

where X is an *n* by *m* matrix of values for *n* tests obtained by *m* persons.

The PARAFAC (hereafter called PARAFACl) generalization of this model (Harshman, 1970) can be expressed by considering a number of two-way data matrices simultaneously (thus obtaining a three-way data set). The values for the factor loadings of the tests (or alternatively, the persons, it does not matter which) are altered from one occasion to the next by a proportional change as follows:

(1d) $$x_{jk}^{(i)} = a_{j1}c_1^{(i)}b_{1k} + a_{j2}c_2^{(i)}b_{2k} + \ldots + a_{jl}c_l^{(i)}b_{lk}$$

where $x_{jk}^{(i)}$ represents the *ith* two-way matrix x_{jk}. We will here consider it the matrix for the *ith* "occasion", just as we considered x_{jk} to be values for "tests" obtained by "persons", although all these names are only for convenience in visualizing a particular application of the model.

In summation notation we can write

(1e) $$x_{jk}^{(i)} = \sum_{r=1}^{l} a_{jr}c_r^{(i)}b_{rk}$$

and in matrix notation this becomes

(1f) $$X_i = AD_iB'$$

where X_i is the two-way matrix x_{jk} for the *ith* occasion, A and B are the matrices of factor loadings for tests and persons, as before, and D_i is an l by l diagonal matrix, whose diagonal cells give proportional changes in the loadings for the factors on the *ith* occasion.*

In this model of data, there are no restrictions on the patterns of loadings which occur in A and B. In particular, no assumptions are made that the columns of A or B are orthogonal to one another.

If there are a sufficient number of different X_i in a given set of data matrices, the decomposition described by (lf) is unique for a small enough number of factors (l), except for trivial differences which do not affect interpretation. Conditions of data adequacy and of uniqueness of the extracted factors, have been described in Harshman, 1970 (pp. 20-23, 35-44, 61-62).

Because of this uniqueness, and the minimal assumptions under which it is obtained, an argument can be made that factors extracted in this way from "system-variation" data (Harshman 1970, pp. 20-23) have a greater likelihood of explanatory validity than those selected by traditional factor analysis using some rotation principle such as Varimax or simple structure. (For the argument, see Harshman 1970, pp. 1-26.)

2. Representation of summed-cross-product matrices

Now let us consider the problem of representing, in terms of the factors of X, values derived from X by taking sums of cross-products. If the means of the columns of X are zero, the summed-cross-products for two variables will be proportional to their covariances. If, in addition, the variances of the columns of X are unity, then the cross-products for two variables will be equal to the correlation coefficient for those two variables.

For tests j and j', the covariance across the m persons is

(2a) $$c_{jj'} = \frac{1}{m}\sum_{k=1}^{m} x_{jk}x_{j'k} \quad . \qquad (\text{if } \bar{x}_j = \bar{x}_{j'} = 0)$$

---

* In other words, if $d_{rs}^{(i)}$ is an entry of D_i, then $d_{rs}^{(i)}=0$ if $r \neq s$, and $d_{rs}^{(i)} = c_r^{(i)}$ if $r=s$ where $c_r^{(i)}$ is used as in equation (1e).

This is simply 1/*m* times our general summed-cross-product. The full matrix of summed-cross-products C, can be represented in matrix notation by

(2b) $C = XX'$.

To represent these summed-cross-products (e. g,. unscaled covariances) in terms of factors, rather than data values, we can substitute our theoretical factor model of the data matrix (from equation (lf)) into our description of the summed-cross-products (2b). This will give us, for the cross-product matrix derived from the *ith* data matrix

(2c) $C_i = (AD_iB')(AD_iB')'$.

Since the transpose of (AD_iB') is the transpose of the components taken in reverse order we can rewrite $(AD_iB')' = B'' D_i'A'$. Further, since $B'' = B$ and $D_i' = D_i$ (since D_i is diagonal), we can rewrite (2c) as

(2d) $C_i = (AD_iB')(BD_iA')$

regrouping gives

(2e) $C_i = AD_i(B'B)D_iA'$

or, combining terms

(2f) $C_i = AD_iWD_iA'$ (where $W = B'B$).

This describes our *ith* summed-cross-product or covariance matrix in terms of A, our *n* by *l* matrix of loadings of the *n* tests on the *l* factors; D_i, our diagonal matrix giving the weights of the *l* factors on the *ith* occasion, and W an *l* by *l* matrix related to the *factor* cross-products. If the columns of B have zero means, then $(1/l)W$ is the matrix of covariances of the factors. If the cells of W are scaled so that the diagonals are equal to one*, then w_{rs}, a cell of W after appropriate scaling, gives the correlation between factor *r* and factor *s*. This is interpretable geometrically as the cosine of the angle between factor *r* and *s* in the space spanned by the factors.

In summation notation, this model (2f) is written

(2g) $$c_{jk}^{(i)} = \sum_{s=1}^{l} \sum_{r=1}^{l} a_{jr} d_{rr}^{(i)} w_{rs} d_{ss}^{(i)} a_{ks} \quad .$$

3. Orthogonal factors

Now if the factors are uncorrelated with one another in their pattern of loadings across people, i.e. if the columns of B are uncorrelated (have average cross-products of zero) then

(3a) $W = B'B = \Delta$

---

*(i.e. w_{rs} is divided by $\sqrt{(w_{rr})(w_{ss})}$, and the inverse scaling is made on all the D matrices to compensate)

where Δ is some diagonal matrix. This gives zero cosines for the angles between different factors, and thus the factors are said to be orthogonal. Now in this case, the summed cross-products can be represented

(3b) $C_i = AD_i\Delta D_iA'$.

This can be rewritten

(3c) $C_i = A\tilde{D}_iA'$

where $\tilde{d}^i_r$, the *rth* diagonal element of $\tilde{D}_i$, is equal to $(d^i_r)^2(k_r)$, where k_r is the *rth* diagonal element of Δ, and relates to the "size" of factor *r* (i.e. the sum of the squared loadings of factor *r* in B). If the mean loading for factor *r* were zero, k_r would be proportional to the variance of factor *r* in B.

Suppose we took a set of summed-cross-product or covariance* matrices C_i and analyzed them by the PARAFAC1 procedure. Then we would be decomposing them according to the model of (lf), namely

(3d) $C_i = \overset{*}{A}\overset{*}{D}_i\overset{*}{B}'$.

Now this would yield a perfect fit and a "unique" solution (see Harshman 1970), when

(i) model (lf), or the more general model of section 7, is appropriate for the data from which the C_i were computed;

(ii) we had guessed the number of factors correctly (guessed the correct value for *l*);

(iii) $B'B = \Delta$, that is, the factors were uncorrelated or "orthogonal" across B in the original data from which the cross-products were computed.

Further, we would discover the $\overset{*}{B}$ was proportional to $\overset{*}{A}$. Thus the solution could be represented (by choosing the right $\overset{*}{D}_i$) as

(3e) $C_i = \overset{*}{A}\overset{*}{D}_i\overset{*}{A}'$.

By comparing the PARAFACl solution in this form with our equation (3c), we can interpret it as follows. $\overset{*}{A}$ would equal A, the matrix of loadings of the tests on factors which underlie the original data from which the covariances were computed. $\overset{*}{D}_i$ would equal $(D_i)^2\Delta$, the squares of the weights of the factors for the *ith* occasion, multiplied by coefficients proportional to the "size" of the factors in B. We note that D_i should therefore have all elements greater than or equal to zero. If PARAFAC gave a negative value in $\overset{*}{D}_i$ it would indicate a failure of the data to correspond to the model.

4. Generalized model for oblique factors

We have seen that PARAFAC1 provides an adequate model for analysis of covariance or other summed-cross-product matrices only when the factors are orthogonal in the data. This restricts the application of the model, since

---

* Covariance matrices can be analyzed by PARAFAC2 only if the means of the variables do not change appreciably from one C_i matrix to the next. If they do, cross products or average cross products must be used.

it is not usually possible to know whether or not this would be the case for any given set of covariance matrices.

Therefore, we adopt the more general expression derived in (2f) as a general model for three-mode factor analysis of matrices composed of summed-cross-products:

(4a) $C_i = AD_iWD_iA'$

We shall call this model PARAFAC2. This model describes a set of C_i cross-product matrices in terms of a common set of factors A, and a common set of angles between the factors, W. The occasions differ only in the weights given the factors, described for the *ith* occasion (the *ith* matrix of summed-cross-products) by D_i.

5. Derivation of PARAFAC2 for scalar product matrices derived from distance matrices

In multidimensional scaling, and in particular in the Carroll and Chang three-mode model (Carroll and Chang 1970) we start with a set of matrices giving the distances between stimuli. These are then converted to the scalar products of vectors by application of a trigonometric relationship which holds for oblique or orthogonal coordinates (see Torgerson 1958, pp. 255, 258). The resulting scalar products represent something similar to covariances between the stimuli. Each scalar product gives the "common strength" of two stimuli by multiplying the length of the first stimulus vector from some origin by the size of the projection of the second stimulus vector upon the first.

Now scalar products can also be represented by the sum of products of the projections of the two stimuli onto a set of *orthogonal* axes as follows: Let A be an *n* by *l* matrix of projections of *n* stimuli onto *l* orthogonal axes, then

(5a) $S = AA'$,

where S is an *n* by *n* matrix of scalar products of the stimulus vectors.

Now Carroll and Chang generalized this model to a set of parallel scalar product matrices for the same *n* stimuli, allowing individual axes to be expanded or contracted in each set (Carroll and Chang 1970, pp. 284-5). This gives

(5b) $S_i = (AD_i)(AD_i)'$

which can be rewritten

(5c) $S_i = AD_i^2A'$.

This is obviously closely related to our model (3e) for summed-cross-products with orthogonal factors. Now the scalar products as derived from the distances between stimuli can only be represented by $S_i = AD_i^2A'$ when the axes onto which the stimuli are projected are orthogonal.

The problem with oblique axes does not arise in traditional two-mode multidimensional scaling. Since the orthogonal axes extracted by two-mode scaling are arbitrary, they can always be rotated, without loss of fit, into a "true" oblique position. But the INDSCAL – PARAFAC1 unique three-mode solution will be the best fitting pair of orthogonal axes "common to all persons." If the true axes are oblique, experiments with synthetic data show that two different distortions will occur. Not only will the orientations of the INDSCAL – PARAFAC1 axes be incorrect, but the recovered stimulus configuration will be distorted as well. (But it is interesting to note that the distorted stimulus configuration with the orthogonal axes will predict the data almost as well as the correct configuration with oblique axes.)

Let us derive a model for three-mode multidimensional scaling which allows oblique, as well as orthogonal axes. We can express the scalar products in terms of oblique axis projections of the stimuli as follows.

Let A be an n by l matrix giving the projections of n stimuli onto a set of l oblique (or orthogonal) axes, the "underlying perceptual dimensions".
Let D_i be a diagonal l by l matrix giving the expansions or contractions of the axes of A for the *ith* person.
Let T be an l by l matrix describing the projections of the axes of A onto a set of orthogonal axes (so that AT describes the projections of the stimuli onto a set of orthogonal axes).
Then:

(5d) $S_i = (AD_iT)(AD_iT)'$

(5e) $S_i = (AD_iT)(T'D_iA')$.

(5f) $S_i = AD_i(TT')D_iA'$

(5g) $S_i = AD_iWD_iA'$.

Thus we get the same model as developed earlier for summed-cross-products, with the same interpretation for all the constituent matrices.

A gives the stimulus projections, D_i the relative scale or importance of the dimensions for the *ith* person, and W the angles between the dimensions.

6. Uniqueness of PARAFAC2 Solutions

No proof of the uniqueness or of the conditions for non-uniqueness has yet been discovered for PARAFAC2. Nonetheless, computer experiments with PARAFAC2 and synthetic data have established empirically that it will give unique solutions, if the number of factors extracted is not greater than the number "actually in the data" (used to create the synthetic data), and if there are a sufficient number of independent C_i matrices. When too many factors are extracted however, the solution is not unique. The behavior of the

model and the conditions of uniqueness seem to be similar to those of PARAFAC1, which are described in detail in Harshman (1970).

The term "unique" needs some clarification. There is a trivial non-uniqueness of scale for PARAFAC2, in that all the loadings on a given factor can be doubled or otherwise changed in scale, as long as a compensatory inverse scaling is performed on all the person weights for that factor (or, on the W matrix). Such changes do not affect the pattern of loadings across measures or across persons, and since this pattern is what is interpreted, such changes do not alter the interpretation that would be given to a particular factor. The columnar order of the factors is also arbitrary. Factor I on a given analysis may come out as factor III on a different analysis, but the same total set of factors will be found to be present in both cases.

Mathematically, these two trivial types of indeterminacy can be expressed by the following equalities. Let S be an *l* by *l* diagonal matrix changing the scale of a set of *l* factors. Let S be another such diagonal scale-changing matrix. Let P be an *l* by *l* permutation matrix changing the order of the *l* factors. We can see that if

(6a) $C_i = AD_iWD_iA'$

then we can insert *l* by *l* identity matrices into this equation without altering it, as follows

(6b) $C_i = A(I)D_i(I)W(I)D_i(I)A'$

now since $(SS^{-1})=I$, and $\left(\overset{*}{S}\overset{*}{S}^{-1}\right)=I$, we can substitute

(6c) $C_i = A\left(SS^{-1}\right)D_i\left(\overset{*}{S}\overset{*}{S}^{-1}\right)W\left(\overset{*}{S}^{-1}\overset{*}{S}\right)D_i\left(S^{-1}S\right)A'$.

In a similar fashion, since $(PP^{-1})=(I)$, we can write

(6d) $C_i = A(PP^{-1})D_i(PP^{-1})W(PP^{-1})D_i(PP^{-1})A'$

combining these substitutions, we can write

(6e) $C_i = A\left(P\left[SS^{-1}\right]P^{-1}\right)D_i\left(P\left[\overset{*}{S}\overset{*}{S}^{-1}\right]P^{-1}\right)W\left(P\left[\overset{*}{S}^{-1}\overset{*}{S}\right]P^{-1}\right)D_i\left(P\left[S^{-1}S\right]P^{-1}\right)A'.$

By regrouping parentheses we can display the indeterminacies of both scale and permutation that the A, D and W matrices are subject to, and the compensatory inverse transformations that are necessary to preserve the equality:

(6f) $C_i = \left(APS\right)\left(S^{-1}P^{-1}D_iP\overset{*}{S}\right)\left(\overset{*}{S}^{-1}P^{-1}WP\overset{*}{S}^{-1}\right)\left(\overset{*}{S}P^{-1}D_iPS^{-1}\right)\left(SP^{-1}A'\right)$

However, there are other special indeterminacies of the *signs* in the D_i matrices which aren't as obvious. These indeterminacies can be expressed by the following rule:

(Rule 6a) The sign of a single person's loading on any given factor can be reversed, provided that his signs

are also reversed for loadings on all factors that are oblique to (i.e. have a non-zero cross-product with) the given factor.

This rule can be understood by examining equation 2g. The individual person loadings are always taken two at a time. If , whenever a sign is reversed for one element of such a pair of d_{rr}, d_{ss}, it is also reversed for the other, then the value of the product is not changed. Now if two factors are orthogonal to one another, then the w_{rs} value for that pair is zero, and those products in which that pair of d_{rr} , d_{ss} values would have occurred will all vanish. Therefore reversal of the sign of d_{rr} to compensate for a reversal of d_{ss}, (or vice versa) is not necessary when factors *r* and *s* are orthogonal. A consequence of these rules is that for an orthogonal solution, the signs of all the "d" values are arbitrary. Only the squares of these values would enter into the solution.

In many applications of PARAFAC2, it seems reasonable to put certain constraints on the D_i matrices. It seems reasonable in multidimensional scaling, for example, to demand that all elements of D_i , (i.e. all the person weights for the dimensions), be non-negative. This can be interpreted as requiring that all persons see a given dimension. as going in the same direction through the space. It prohibits a description in which a given individual uses a dimension as if it is "reversed in direction" compared to the way others use it. An option for constraining the elements of the D_i to non-negative values has been implemented in the current PARAFAC2 computer program.

7. Analysis of Non-System-Variation Data with PARAFAC2

Only certain types of three-mode data are analyzable with PARAFAC1. The data model underlying PARAFAC1 requires certain strict correspondences between the data matrices from one occasion to the next. First, it requires that the data consist of measurements of the same objects (or persons) on the same variables, on every occasion. Second, it requires that the patterns of variation of the factor influences from one occasion to the next must have a special, restricted form. Specifically, it requires that the changes in one person's loadings on any given factor be proportional to the changes in every other person's loadings on that factor. (See equation ld or lf. For a more detailed discussion of system-variation, and other types of data variation, see Harshman, 1970, pp. 18-26.) The requirement of system-variation data places a substantial undesirable restriction on the areas of application of the original PARAFAC1 model. It is therefore of great practical, as well as theoretical significance to note that PARAFAC2 is not restricted to analysis of system-variation data. It is possible to derive the same PARAFAC2 model for cross-products, using a much more general model for data than was used in section 2.

Let us consider the case where we have no particular correspondence between the columns of our data matrices from one occasion to the next. We shall study a set of *n* common measures, as was done in section 2. But instead of repeatedly measuring the same set of persons, we shall measure a different set of persons on each occasion. The number of persons will

vary from occasion to occasion. On the *ith* occasion, our data matrix X_i will have the dimensions *n* by m_i, where m_i is the number of persons measured on that occasion.

Since the same set of *n* measures is used on each occasion, we hypothesize that the same set of common factors is used by the m_i persons to respond to the *n* measures. Let A represent the *n* by *l* matrix of factor loadings on measures. Let B_i be an m_i by *l* matrix giving the loadings of the m_i persons in the *ith* subpopulation on the *l* common factors underlying the measures. We can then describe our *i* data matrices as follows

(7a) $X_i = AB_i'$

Now if the correlations between the factors are caused by some basic relationship between the influences which those factors represent, then it will often be reasonable to assume that these correlations remain constant from one subpopulation to another, and that only the size of a given factor's influence will vary across subpopulations. If B_i and B_j are the factor loading matrices for any two of our subpopulations, we can express the consistent "correlations" (cross-products) of the factors by writing

(7b) $B_j' B_j = D_i B_i' B_i D_i$

where D_i is a diagonal matrix with the same interpretation as in the system-variation model. It describes the changes in "size" of the factors from one occasion to the next.

Now let us derive a representation of the cross-products derived from data which is describable by (7a) and (7b). If C_i is the cross-product matrix derived from X_i, then

(7c) $C_i = (X_i)(X_i)'$.

But by substitution of our definition (or model) for X_i given earlier in (7a) we get

(7d) $C_i = (AB_i')(AB_i')'$

By applying the rule that the transpose of a product is the product of the transposes in reverse order, we get

(7e) $C_i = (AB_i')(B_i''A') = (AB_i')(B_iA')$

or, regrouping terms,

(7f) $C_i = A(B_i'B_i)A'$.

Now by applying the second part of our model, (7b), which postulates the same "correlations" among the factors in the different subpopulations, but differing sizes as described by the D_i diagonal matrices, we can write

(7g) $C_i = A(D_i(B_j'B_j)D_i)A'$

for some other jth subpopulation. In general if we write

(7h) $W = B_j'B_j$

we can then describe the cross-products in terms of the model of PARAFAC2, namely

(7i) $C_i = AD_iWD_iA'$

with the same interpretations of the constituent matrices as in the derivation from a system variation model, in section 2.

Two special cases of this general model should be noted. In the first special case, all the B_i could be taken from the same people. The size of the B_i matrices would all be the same, but the columns of B_i would not necessarily be proportional to the corresponding columns of B_j . This case corresponds to the example of personality test "object-variation" type data described elsewhere (Harshman, 1970, pp.22-23).

An even more restricted special case is where $B_j = B_iD_i$, the columns of B_i are proportional to the columns of B_j for all *i*, *j*, under study. This would be a case of system-variation. This reveals how the system-variation data model is a special case of the more general PARAFAC2 model.

8. Estimation Algorithms

Work is currently in progress to find the most efficient way of fitting the PARAFAC2 model to a set of data. A detailed report will be forthcoming in a later issue of W.P.P., but a brief glimpse of current developments is appropriate here.

An algorithm for PARAFAC2 has been devised by Robert Jennrich (UCLA Dept. of Mathematics) and implemented by the author as a FORTRAN IV program on the IBM 360/91 computer. The algorithm uses an approach similar to the "quick algorithm" for PARAFAC1. It estimates values for one of the three basic sets of parameters (A, D_j or W) while holding the other two sets of parameters fixed. A complete iteration consists of first estimating the D_i 's, then W, and finally A. The program iterates this process until the values converge on a stable least-squares solution.

This procedure generally works, but it is expensive, and sometimes slow to converge. Even worse, from certain bad starting positions local optima or extremely slow convergence can be encountered. We are accumulating experience, however, which would show how, by putting certain constraints on starting positions and values of the parameters, these deadlocks can largely be avoided.

A basic modification to this special algorithm is currently under study. An attempt is being made to combine Jennrich's algorithm with a generalized optimization routine called TBU (Reichenbach, 1962, 1969). The combined program alternates between two phases: first, it observes how the parameters

change during an iteration of Jennrich's algorithm. It computes or updates trends for the direction and amount of change in each parameter. In the second phase, it moves all the parameters simultaneously an optimal (hopefully large) distance in the direction of their respective trends. The author has been working with Hans Reichenbach on this approach and has found that it can rapidly accelerate the convergence of PARAFAC1 in certain difficult cases. The application of this technique to PARAFAC2 is the next step to be implemented. Preliminary results with PARAFAC2 have been encouraging.

9. Avoiding the "Communalities" Problem

In deriving our model of the factor structure underlying data and cross-products, we have ignored the fact that there would be a certain amount of error or "noise" in the data. For PARAFAC1, and its estimation procedure, this approach was justified. We are looking for a least squares best estimate of the common-factor loadings, or the A, D_i and B values. The process of directly optimizing the fit of these parameters to the raw data provides the least square estimates that we desire. The problem becomes more complicated, however, when we consider cross-product matrices.

Let us take the following model of "noisy" data:

(9a) $$\tilde{X}_i = X_i + E_i = AD_iB_i' + E_i$$

where E_i is a matrix of the same size as X_i, containing random "error terms". (In traditional E_i actually represents "unique variance" composed of both "specific" and "error" variance.) Since X_i represents *all* the common-factor part of the data matrix $\tilde{X}_i$, then what remains when we subtract it from X_i should be a matrix of uncorrelated random errors. The rows of E_i should be uncorrelated both with the rows of X_i and with each other. The means of these random errors should be zero.

We can develop a model for the cross-products from such "noisy" data, as follows:

(9b) $$\tilde{C}_i = \tilde{X}_i\tilde{X}_i'$$

By substituting from (9a) into (9b) we get

(9c) $$\tilde{C}_i = (X_i + E_i)(X_i + E_i)'$$

(9d) $$\tilde{C}_i = (X_i + E_i)(X_i' + E_i')$$

(9e) $$\tilde{C}_i = X_iX_i' + E_iX_i' + X_iE_i' + E_iE_i'$$

now, since the rows of E_i are uncorrelated with the rows of X_i, we can write

(9f) $$\tilde{C}_i = X_iX_i' + 0 + 0 + E_iE_i'$$

(9g) $$\tilde{C}_i = C_i + E_iE_i'$$

Although the rows of E_i are uncorrelated with one another, each row is of course correlated with itself. Thus $(E_i)(E_i')$ will yield a diagonal matrix whose diagonal elements represent the sum of the squared errors for that

row of the data matrix (i.e. a measure of the amount of noise in that variable). This diagonal matrix is added to the matrix predicted by the common factors, and therefore increases the size of the diagonal elements of the cross-product matrix beyond what would be predicted by the factors.

An attempt to achieve a least squares best fit for all the values of the covariance matrix, including the diagonals, would distort the solution, since the diagonals are larger than would be predicted by the factor loading matrices.

This problem is analogous to the standard "communalities" problem of traditional two-mode factor analysis. We solve this problem in the following way. In the algorithm which performs the fitting of the model to the data (see section 8 of this article) we perform an extra step at the end of each iteration. In this step, we replace the diagonal entries of the cross-product matrices with the values for those diagonal entries which are predicted by the model at that iteration. This will cause the values in the diagonals to "float" one iteration behind the predictions of the model. As the algorithm proceeds, the difference between these diagonal values (from the last iteration) and the models prediction (from this iteration) will soon become very small, and therefore will not influence the process of convergence. The diagonal values will, in effect, be ignored and only the off diagonal values will be 'fit' by the program. This should provide the best possible least squares estimate of the A, D_i and W parameter matrices.

(This same technique for ignoring cells in the data matrices can be used to ignore missing values in PARAFAC1 or PARAFAC2.)

10. Analysis of Correlation Matrices

The technique for analysis of cross-product matrices developed in this article applies directly to covariance matrices, as was pointed out in section 2, but it does not apply directly to the analysis of correlation matrices.

In a correlation matrix, the rows and columns of the covariance matrix have been rescaled so that the maximum value for any entry is 1.0. This is accomplished by dividing each entry by the standard deviation of its column and row variable. In other words,

(10a) $$Corr_{ij} = Cov_{ij} / s_i s_j$$

where s_i, s_j represent the standard deviation of the *ith* and *jth* variables respectively.

In matrix notation, we can represent the relation between the covariance matrix C and the correlation matrix R, as follows

(10b) $$C = \underline{D} R \underline{D}$$

where $\underline{D}$ is a diagonal matrix whose *ith* diagonal element is the standard deviation of the *ith* variable. Therefore

(10c) $R = \underline{D}^{-1}C\underline{D}^{-1}$

and if we substitute our PARAFAC2 model for the covariance into equation (10c) we get

(10d) $R_i = \underline{D}_i^{-1}AD_iWD_iA'\underline{D}_i^{-1}$.

This expression is more complicated than PARAFAC2, and its mathematical properties have not been investigated. It serves to demonstrate, however, why correlation matrices cannot be directly analyzed by PARAFAC2. It would seem inefficient to derive a further algorithm for this model (which may not have the same degree of uniqueness) when one can simply compute covariance matrices instead of correlation matrices and analyze them with PARAFAC2.

11. Relation of PARAFAC2 to Jennrich's multidimensional scaling model

Jennrich has developed a generalization of Carroll and Chang's INDSCAL multidimensional scaling model which leads to a decomposition of matrices of summed-cross-products which is even less restricted than PARAFAC2. His derivation is based on a generalized Euclidean distance metric, and allows each person to use his own metric in judging inter-stimulus distances. His derivation is presented in an accompanying article (Jennrich 1972).

It is possible to reach Jennrich's model by a slightly different path, namely generalizing PARAFAC2 so that the correlations among the factors underlying the data can vary from one occasion to the next.

Start with the model of PARAFAC2, namely

(lla) $C_i = AD_iWD_iA'$.

Now let each occasion have its factors $\overset{*}{A}$ related to A by some oblique or orthogonal rotation. Let T_i be the l by l matrix describing this rotation by giving the projections of the axes of A onto the axes of $\overset{*}{A}$.

If we let $\overset{*}{A}_i$ be the axes which are expanded or contracted on each occasion (or by each individual), then

(11b) $\overset{*}{A}_i = AT_i$

(llc) $C_i = (AT_i)D_iWD_i(T_i'A')$

(lld) $C_i = A(T_iD_iWD_iT_i')A'$

(lle) $C_i = AW_iA'$

This model would not give a unique solution since, for example, we can always let $\tilde{A} = AR_i$ and $W_i = (R_i^{-1}W(R_i')^{-1})$. It might be useful, however, to choose R_i such that $\tilde{A}$ approaches A of PARAFAC2 (2f) as closely as possible. Then the following research strategy might seem reasonable. First, analyze with PARAFAC2. Then use Jennrich's model (lle) to analyze the same data. If the improvement of fit gained with the more general model is

significant enough (by some criterion) to lead the investigator to suspect that there were "real" changes in obliqueness across occasions, then rotate the solution of (lle) into maximal agreement with (2f) as just described above. The model of (2f) (PARAFAC2) describes the common space, and (lle) describes how each occasion (or individual) distorts that common space by rotating the axes, in addition to simply expanding or contracting them.

Carroll (personal communication) has recently pointed out that this (lle) model is mathematically equivalent to one which he discusses in conjunction with INDSCAL (Carroll, 1970). The equivalence is easily missed, because Carroll interprets his model in terms of a common set of axes which each person subjects to a different orthogonal rotation in addition to contraction or expansion. The interpretation developed above by Jennrich and the author is one of each person imposing an oblique transformation on the common axes in addition to his expansion or contraction of the different axes.

Carroll and some others have raised questions about the interpretation of oblique perceptual axes. This interesting question will not be treated here. The author believes that interesting and plausible interpretations do exist, and notes, more importantly, that PARAFAC2 analysis of real perceptual data does in fact sometimes reveal oblique perceptual dimensions. (see Terbeek and Harshman, 1972.)

Carroll is developing a program to analyze data in terms of the more general model of (lle). It will be called IDIOSCAL (for IDIOsyncratic SCALing). If we applied to IDIOSCAL the interpretation developed above for (lle) and rotated it into maximum agreement with PARAFAC2, it might by a very useful partner to PARAFAC2 in multidimensional scaling.

References

Carroll, J.D. and Chang, J.J. (1970), "Analysis of individual differences in multidimensional scaling via an N -way generalization of 'EckartYoung' decomposition," *Psychometrika* 35, 283-319.

Harshman, R.A. (1970), "Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-mode factor analysis," *Working Papers in Phonetics No. 16*, 84pp.

Jennrich, R. (1972), "A generalization of the multidimensional scaling model of Carroll and Chang," *Working Papers in Phonetics No. 22*. [see following pp. in this file]

Reichenbach, H.G. (1962), Routine TBU: routine for optimizing a set of variables. Report 62-21, Institute of Transportation and Traffic Engineering, University of California at Los-Angeles, 1962. (unpublished expanded version dated 1969).

Terbeek, D. and Harshman, R.A. (1972), Is Vowel Perception Non-Euclidean? *Working Papers in Phonetics No. 22*.

Torgerson, W.S. (1958), *Theory and Methods of Scaling*, New York: John Wiley and Sons, Inc.

NOTE: This manuscript was originally published in 1972 and is reproduced here to make it more accessible to interested scholars. The original reference is
Jennrich, R. (1972). A generalization of the multidimensional scaling model of Carroll and Chang. *UCLA Working Papers in Phonetics*, *22*, 45-47. (University Microfilms, Ann Arbor, No. 10,085).

A Generalization of the Multidimensional Scaling Model of Carroll and Chang

Robert Jennrich

(Mathematics Department, UCLA)

1. Carroll and Chang model

Carroll and Chang (1970) assume each stimulus may be represented by a point

$$p_j = (x_{j1}, \ldots, x_{jr}) \quad .$$

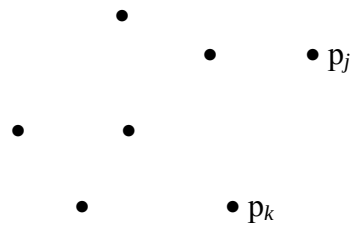
The theoretical distance between the *j* and *kth* stimulus as judged by the *ith* individual is given by

$$d_{jk}^{(i)} = \left(\sum_{t=1}^{r} w_{it} \left(x_{jt} - x_{kt} \right)^2 \right)^{1/2} \quad .$$

For each *i* this defines a norm on the *r*-dimensional space which is the home of the stimulus points p_j. Each individual views the points differently, the difference being reflected in the dependence of the weight w_{it} on individual index *i*.

2. Generalization

In pictures, Carroll and Chang view the stimuli as points in an *N*-dimensional space.



The most general Euclidean distance formula in such a space is

$$d^2(p_j, p_k) = (p_j - p_k)\, W\, (p_j - p_k)'$$

where W is a symmetric positive definite matrix. Carroll and Chang have demanded that W be diagonal. We simply suggest relaxing this restriction. Each individual *i* then has his own completely arbitrary Euclidean metric and the distance between the *jth* and *kth* stimulus as seen by the *ith* individual becomes

$$d_{jk}^{(i)} = \left(\sum_{s=1}^{r} \sum_{t=1}^{r} w_{ist} \left(x_{js} - x_{ks} \right) \left(x_{jt} - x_{kt} \right) \right)$$

3. Estimation

For purposes of estimation the distances $d_{jk}^{(i)}$ are converted to inner products (Torgerson 1958)

$$b_{jk}^{(i)} = -1/2 \left(\left(d^{(i)} \right)_{jk}^{2} - \left(d^{(i)} \right)_{j\cdot}^{2} - \left(d^{(i)} \right)_{\cdot k}^{2} + \left(d^{(i)} \right)_{\cdot\cdot}^{2} \right)$$

where "·" means average as usual. We have now, assuming no errors,

$$b_{jk}^{(i)} = \sum_{s=1}^{r} \sum_{t=1}^{r} w_{ist} x_{js} x_{kt} \quad .$$

Using observed values of $b_{jk}^{(i)}$ we will attempt to fit these by least squares using the model on the right.

4. Algorithm

Let $B_i = \left(b_{jk}^{(i)} \right)$, $W_i = \left(w_{irs} \right)$, $X_i = \left(x_{js} \right)$. Compute

$$Y = \left(\sum_i B_i X W_i \right) \left(\sum_i W_i X' X W_i \right)^{-1}$$

$$\tilde{X} = 1/2 \left(X + Y \right)$$

$$\tilde{W}_i = \left(\tilde{X}' \tilde{X} \right)^{-1} \tilde{X}' B_i \tilde{X} \left(\tilde{X}' \tilde{X} \right)^{-1} \quad .$$

$\tilde{X}$ and $\tilde{W}$ are the new values of X and W. Iterate for convergence.

References

Carroll, J.D. and Chang, J.J. (1970), "Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckhart-Young' decomposition," *Psychometrika* 35, 283-319.

Torgerson, W.S. (1958), *Theory and Methods of Scaling*, New York: John Wiley and Sons, Inc.