RB 001-79

PRINCIPAL COMPONENT ANALYSIS OF THREE-MODE DATA BY MEANS OF ALTERNATING LEAST SQUARES ALGORITHMS INCLUDING PROOFS

Pieter M. Kroonenberg

Department of Pedagogy

Jan de Leeuw

Department of Data Theory

January 1979

DEPARTMENT OF DATATHEORY FACULTY OF SOCIAL SCIENCES

UNIVERSITY OF LEYDEN-THE NETHERLANDS

ERRATUM bij Principal component analysis of three-mode data by means of alternating least squares algorithms including proofs
Pieter M. Kroonenberg & Jan de Leeuw

Wherever is referred to Kroonenberg & De Leeuw (1979)
read Part II of this report.

PART I

PRINCIPAL COMPONENT ANALYSIS OF THREE-MODE DATA BY MEANS OF ALTERNATING LEAST-SQUARES ALGORITHMS

(preprint)

Summary

A new method to estimate the parameters of Tucker's three-mode principal component model is discussed, and the convergence properties of the alternating least-squares algorithm to solve the estimation problem are outlined. A special case of the general Tucker model, in which only over two of the three modes principal component analysis is performed, is briefly outlined as well. The Miller & Nicely data on the confusion of English consonants are used to illustrate the programmes TUCKALS3 and TUCKALS2 which incorporate the algorithms for the two models respectively.

Keywords: three-mode principal component analysis, alternating least-squares method, INDSCAL, successive block algorithm, simultaneous iteration method, point-of-view analysis, factor analysis, confusion of consonants

1. Three-mode models and their solutions

The three-mode model - here referred to as the Tucker3 model - was first formulated by Tucker (1963), and in subsequent articles Tucker (1964,1966) and Levin (1963) extended especially the mathematical description and programming aspects of it

In the multidimensional scaling context references to this model occur frequently (e.g. Harshman (1970), Jennrich (1972), Carroll & Chang (1972), and Takane, Young & De Leeuw (1977)), as the Tucker3 model is the general model comprising various individual differences models. A discussion of the relationships between multidimensional scaling and three-mode principal component analysis can be found in Tucker (1972), Carroll & Wish (1974), Takane, Young & De Leeuw (1977), and Sands (1978).

As far as we have been able to trace the algorithms developed by Tucker (1966) are used to solve the three-mode model in all but one case. References to computer programmes based on these algorithms are Wainer et al. (1971), Wainer et al. (1974), Walsch (1964), Walsch & Walsch (1976), and one such programme is embodied in the statistical package SOUPAC developed at the University of Illinois. Numerous similar programmes have been written, and they are mostly referred to in passing in applied articles. The one exception is the programme ALSCOMP3 developed by Sands (1978), who is using - just as we are - an alternating least-squares technique.

In his 1966 article Tucker remarks that his procedures "do not produce a least-squares approximation to the data. Investigations of the mathematics of a least-squares fit for three-mode factor analysis indicate a need for an involved series of successive approximations." The procedures described in the present article are designed to provide least-squares estimates of the parameters in the three-mode model. The alternating least-squares approach used can also be extended towards other levels of measurement as has been recently demonstrated by Sands(1978).

2. Description of the Tucker3 model

definition three-mode matrix

A $l \times m \times n$ three-mode matrix $Z \in R^{l \times m \times n}$ is defined as the collection of elements:

$$\{z_{ijk} \mid i=1,\ldots,l;\ j=1,\ldots,m;\ k=1,\ldots,n\}$$

These elements can be thought to be placed in a three dimensional block with the index i running along the vertical axis, the index j along the horizontal axis, and the index k along the 'depth' axis. We will use the word "mode" to indicate a collection of indices by which the data can be classified. For instance, in semantic differential studies (Osgood , Tannenbaum & Suci,1957) one collects scores of a number of persons on a set of bipolar scales for a collection of attributes. These data can be classified by persons, scales, and attributes, each of these therefore determines a mode of the data.

formulation Tucker3 model

The formal formulation of the Tucker3 model is as follows:

The Tucker3model is the factorization of the three-mode matrix $Z = \{z_{ijk}\}$, $Z \in R^{l \times m \times n}$ such that

$$z_{ijk} = \sum_{\alpha=1}^{s} \sum_{\beta=1}^{t} \sum_{\gamma=1}^{u} g_{i\alpha} h_{j\beta} e_{k\gamma} c_{\alpha\beta\gamma} \quad \text{for } i=1,\ldots,l;\ j=1,\ldots,m;\ k=1,\ldots,n,$$

where the coefficients $g_{i\alpha}$,$h_{j\beta}$, and $e_{k\gamma}$ are the elements of the columnwise orthonormal matrices $G \in K^{l \times s}$, $H \in K^{m \times t}$, $E \in K^{n \times u}$ respectively, and the $c_{\alpha\beta\gamma}$ are the elements of the so-called three-mode core matrix $C \in R^{s \times t \times u}$.

A matrix formulation of the model is

$$Z = GC(H' \otimes E')$$

where $Z \in R^{k \times nm}$ and $C \in R^{s \times tu}$ are now ordinary two-mode matrices by making usė of so-called combination modes (Tucker,1966, 281), and $\otimes$ denotes the Kronecker product (Tucker,1966, 283ff). By symmetry there are two other matrix formulations (see also section 4). We will not introduce special notation to distinguish between the two-mode and three-mode versions of Z and C, as the appropriate one is indicated by the real space where it is an element of.

approximation to the Tucker3 model

If we would compute all the principal components, thus s=l, t=m, and u=n, then we could decompose any data matrix exactly in its components. However, in practical applications one is just interested in the two, three or four first principal components. This precludes in general finding an exact factorisation of Z in G, H, E, and C. One,therefore, has to be satisfied with an approximation, i.e. finding G, H, E, and C such that the difference between the model and the data is minimal according to some loss function, or in slightly different terms, we have to look for an best approximate factorization of the matrix Z into G, H, E, and C according to the Tucker3 model.

In our case, as in many similar situations,we define the loss function to be the mean squared one, and propose to search for those G, H, E, and C such that

$$f(G,H,E,C) = ||Z - GC(H' \otimes E')||^2$$

is minimal. where $||.||$ denotes the Euclidean norm. The minimization has to be carried out under the restrictions of the model, i.e. that G, H, and E are columnwise orthonormal matrices.

3. Existence of a best approximate solution

In this section we will show that there exists a best approximate solution to the Tucker3 model.

Minimization problem

Be Z a three-mode matrix with elements z_{ijk}, and $\tilde{Z}$ an approximate factorization with elements

$$\tilde{z}_{ijk} = \sum_{\alpha=1}^{s} \sum_{\beta=1}^{t} \sum_{\gamma=1}^{u} g_{i\alpha} h_{j\beta} e_{k\gamma} c_{\alpha\beta\gamma} \tag{3.1}$$

where $g_{i\alpha}$, $h_{j\beta}$, and $e_{k\gamma}$ are the elements of columnwise orthonormal matrices G, H, and E respectively, and the $c_{\alpha\beta\gamma}$ are the elements of the three-mode core matrix C.

The best approximate factorization of the Tucker3 model is the solution of the minimization of:

$$\begin{aligned} f(G,H,E,C) &= \sum_{i=1}^{l} \sum_{j=1}^{m} \sum_{k=1}^{n} (z_{ijk} - \tilde{z}_{ijk})^2 \\ &= \sum_{i,j,k} (z_{ijk} - \sum_{\alpha,\beta,\gamma} g_{i\alpha} h_{j\beta} e_{k\gamma} c_{\alpha\beta\gamma})^2 \end{aligned} \tag{3.2}$$

The $\tilde{Z}$ for which f attains its minimum will be designated as $\hat{Z}$ with elements:

$$\hat{z}_{ijk} = \sum_{\alpha,\beta,\gamma} \hat{g}_{i\alpha} \hat{h}_{j\beta} \hat{e}_{k\gamma} \hat{c}_{\alpha\beta\gamma} \tag{3.3}$$

where the variables with a caret are the least-squares estimators of the model parameters.

We will now show that there always exist some G, H, E, and C such that f attains its (global) minimum. The proof has two parts:

1. solve the minimization of f for C with fixed G, H, and E;
2. using the resulting $\hat{C}$ show that f attains its minimum for some G, H, and E.

We first show that there exists a unique best C, called $\hat{C}$ such that for fixed G, H, and E f attains its minimum for C which has as its elements:

$$\hat{c}_{\alpha\beta\gamma} = \sum_{\delta=1}^{l} \sum_{\varepsilon=1}^{m} \sum_{\iota=1}^{n} g_{\delta\alpha} h_{\varepsilon\beta} e_{\iota\gamma} z_{\delta\varepsilon\iota} \tag{3.4}$$

To prove the above assertion we use a simplified version of a lemma by Penrose (1955), which is presented as lemma 3.2 in Kroonenberg & De Leeuw (1977). This lemma states that there exists a unique $\hat{C}$ such that the function

$$h(C) = ||Z - \tilde{Z}||^2 = ||Z - GCF'||^2$$

is as small as possible, this $\hat{C}$ is equal to G'ZF, and the absolute minimum, i.e. 0, is reached if and only if $Z = GG'ZFF'$.
To apply this lemma, let us write $\tilde{Z}$ as the matrix product

$$\tilde{Z} = GC(H' \otimes E'). \tag{3.5}$$

In the same manner $\hat{C}$ can be written as

$$\hat{C} = G'Z(H \otimes E). \tag{3.6}$$

If we write F for $H \otimes E$ in both (3.5) and (3.6) we may conclude that $\hat{C}$ as in (3.4) minimizes F for fixed G, H, and E, and that

$$h(\hat{C}) = 0 \quad \text{if and only if} \quad Z = GG'(H \otimes E)(H' \otimes E') \tag{3.7}$$

The further minimization is now only dependent upon G, H, and E. Once we have found the appropriate $\hat{G}$, $\hat{H}$, and $\hat{E}$, we can reconstruct $\hat{C}$ via equation (3.4).

To proceed with the minimization of f we substitute (3.4) into (3.2), and call the rewritten function g:

$$g(G,H,E) = \sum_{i,j,k} \{ z_{ijk} - \sum_{\alpha,\beta,\gamma} g_{i\alpha} h_{j\beta} e_{k\gamma} (\sum_{\delta,\varepsilon,\iota} g_{\delta\alpha} h_{\varepsilon\beta} e_{\iota\gamma} z_{\delta\varepsilon\iota}) \}^2 \tag{3.8}$$

A less unwieldy formula results if we use again the matrix notation:

$$\begin{aligned} g(G,H,E) = ||Z - \tilde{Z}||^2 &= ||Z - GG'Z(H \otimes E)(H' \otimes E')||^2 \\ &= ||Z - GG'Z(HH' \otimes EE')||^2 \end{aligned} \tag{3.9}$$

Let S be defined as:

$$S = \{ s \mid s = (G,H,E),\ G \in K^{l \times s},\ H \in K^{m \times t},\ E \in K^{n \times u} \}, \tag{3.10}$$

then S is a compact subset in a finite dimensional real space.

Noting that g is a continuous bounded function on S $(0 < g < ||Z||^2)$, it may be concluded that there exists some triplet $(\hat{G},\hat{H},\hat{E})$ in S such that g attains its minimum. In other words the minimum problem always has a solution.

4. A solution of the minimization problem

In this section we will give the details of a solution of the minimization problem (3.2). First we will convert the minimization problem into a maximization problem, and subsequently we will present a theorem stating that the component matrices $\hat{G}$, $\hat{H}$, and $\hat{E}$ are nothing but the eigenvectors corresponding to the largest eigenvalues of suitably constructed cross products of the data matrix Z and the other component matrices.

Let us first convert the minimization problem into a maximization problem. Hereto we rewrite (3.8) with traces instead of norms, and manipulate the various terms somewhat:

$$g(G,H,E) = tr(Z - \tilde{Z})(Z - \tilde{Z})' = tr\,(ZZ' - \tilde{Z}Z' - Z\tilde{Z}' - \tilde{Z}\tilde{Z}')$$

$$= trZZ' - 2tr\tilde{Z}Z' - tr\tilde{Z}\tilde{Z}' \qquad (4.1)$$

Expanding each term in turn, and adding them as in (4.1), we get:

$$-2tr\tilde{Z}Z' = -2trGG'Z(HH' \otimes EE')'Z'$$

$$tr\tilde{Z}\tilde{Z}' = trGG'Z(HH' \otimes EE')(HH' \otimes EE')Z'GG'$$

$$= trG'Z(HH' \otimes EE')Z'G$$

$$trZZ' = trZZ'$$

$$+$$

$$g(G,H,E) = trZZ' - trG'Z(HH' \otimes EE')Z'G \qquad (4.2)$$

We define p to be equal to the last term on the right hand side of (4.2):

$$p(G,H,E) = tr\ G'Z(HH' \otimes EE')Z'G \qquad (4.3)$$

Clearly the minimization of g comes down to the maximization of p, as both are bounded. It is advantageous to rewrite p a bit further:

$$p(G,H,E) = tr\ G'\{Z(HH' \otimes EE')Z'\}G \qquad (4.4)$$

$$= tr\ G'PG \qquad \text{with} \qquad (4.5)$$

$$P = P(H,E) = Z(HH' \otimes EE')Z' \qquad (4.6)$$

So far we have always placed H and E in the Kronecker product term, but we could equally well have done so with G and E, or G and H. Such substitutions entail only a change in notation, and not in the model itself. The model is indifferent to such notational changes as can be clearly seen from (3.1).

In the following we will also need the other forms, and thus all three are given below:

$$1.\ p(G,H,E) = \text{tr } G'PG \quad \text{with} \tag{4.5}$$

$$P = P(H,E) = Z(HH' \otimes EE')Z' \text{ and } Z \in R^{l\times mn} \tag{4.6}$$

$$2.\ p(G,H,E) = \text{tr } H'QH \quad \text{with} \tag{4.7}$$

$$Q = Q(E,G) = Z(EE' \otimes GG')Z' \text{ and } Z \in R^{m\times nl} \tag{4.8}$$

$$3.\ p(G,H,E) = \text{tr } E'RE \quad \text{with} \tag{4.9}$$

$$R = R(G,H) = Z(GG' \otimes HH')Z' \text{ and } Z \in R^{n\times lm} \tag{4.10}$$

The maximization of p is not unconstrained, but restricted to the set S. We can incorporate the constraints in the maximization problem by using Lagrange multiplier matrices L, M, and N.

$$\check{p}(G,H,E,L,M,N) = p(G,H,E) - \text{tr } L(G'G - I_s) - \text{tr } M(H'H - I_m) - \text{tr } N(E'E - I_n) \tag{4.11}$$

The maximum of p follows from the requirement that the first order partial derivatives of $\check{p}$ are simultaneously zero at the maximum of p, and that the Hessian is negative.
We will here only state the exact nature of the solution as Theorem 1, but refer the reader for a proof to Kroonenberg & De Leeuw (1979).

Theorem 1

Be Z a three-mode data matrix, and be p, P, Q, R, and S defined as in (4.3), (4.6), (4.8), (4.9) and (3.10) respectively, and finally be U, V, and W defined as follows:

U is an eigenvector matrix of P;
V is an eigenvector matrix of Q;
W is an eigenvector matrix of R;

and $(U,V,W) \in S$.

Then:

a. $(\hat{G},\hat{H},\hat{E}) \in S$ is a stationary point of p if and only if $\hat{G} = U$, $\hat{H} = V$, and $\hat{E} = W$, or some orthonormal rotation thereof;

b. $(\hat{G},\hat{H},\hat{E}) \in S$ maximizes p if and only if their columns are eigenvectors corresponding to the largest s, t, and u eigenvalues of $P(\hat{H},\hat{E})$, $Q(\hat{E},\hat{G})$, and $R(\hat{G},\hat{H})$ respectively, or orthonormal rotations thereof.

The following theorem provides the necessary and sufficient conditions for the existence of an exact solution to the maximization problem (3.2), and provides such a solution. The proof of part A and part B(2) follow directly from the definitions, and the proof of part B(1) can be found in Kroonenberg & De Leeuw (1979).

Theorem 2

A: Be Z a three-mode data matrix, and let f, g, and p be defined as above. Furthermore let (G,H,E,C) satisfy the constraints of (3.1). Then the following statements are equivalent:

$$(1)\ f(G,H,E,C) = 0$$

$$(2)\ g(G,H,E) = 0$$

$$(3)\ p(G,H,E) = \text{tr } ZZ' \qquad Z \in R^{l\times mn}$$

$$(4)\ Z = GG'Z(HH' \otimes EE') \qquad Z \in R^{l\times mn} \tag{4.12}$$

Such a (G,H,E,C) is called an exact solution.

B(1): Let (G,H,E,C) be an exact solution of the minimization problem. Then:

G is the eigenvector matrix (or an orthonormal rotation thereof) associated with the p non-zero eigenvalues of ZZ' with $Z \in R^{l\times mn}$; (4.13)

H is the eigenvector matrix (or an orthonormal rotation thereof) associated with the q non-zero eigenvalues of ZZ' with $Z \in R^{m\times nl}$; (4.14)

E is the eigenvector matrix (or an orthonormal rotation thereof) associated with the r non-zero eigenvalues of ZZ' with $Z \in R^{n\times lm}$; (4.15)

$$C = G'Z(H \otimes E) \tag{4.16}$$

B(2): On the other hand if:

(1) G, H, E, and C are defined as in (4.13) through (4.16),

(2) the eigenvalues associated with G, H, and E are different for each matrix separately, and

(3) (4.12) is satisfied,

then (G,H,E,C) is an exact unique solution.

It should be noted that statement B(2) is not as strong as one would like to have it, as any (G,H,E,C) which satisfies (4.12) determine an exact solution. A more satisfactory statement, however, has not been found yet.

5. Towards an algorithm for the solution of the Tucker3 model

Obviously we would like to construct an algorithm for the maximization of p that converges to a global maximum of p Unfortunately p is the cross-product term of a multivariate polynomial of the sixth degree, and in general it is not possible to prove that methods to solve such non-linear problems attain a global maximum. In the present case this also seems to be true. We will have to be satisfied with proving that the algorithm outlined below will converge to some stationary point, which is not a minimum, rather than a global maximum.

The method to be described utilizes the so-called alternating least-squares (ALS) technique. The essential feature of the ALS approach is that in solving optimization problems with more than one set of parameters, each set is estimated in turn by applying least-squares procedures holding the other sets fixed. After all sets have been estimated once the procedure is repeated again and again until convergence. Early applications of this technique include solutions of regression problems with autocorrelated error terms (Cochrane & Orcutt,1949). Further details and references to applications of the ALS appraoch can, for instance, be found in Young, De Leeuw, and Takane (1979).

In order to see how the ALS approach can be applied in the present context, let us return briefly to (3.9):

$$g(G,H,E) = || Z - GG'Z(HH' \otimes EE')||^2 \qquad \text{(3.9 repeated)}$$

Clearly the sets of parameters are here G,H, and E. Minimizing g over G holding H and E fixed is identical to solving one least-squares problem , minimizing over H with E and G fixed, and minimizing over E with G and H fixed are the two others. That we are in practice maximizing p does not

prevent the problem from being an ALS one.

From the above discussion a rough outline for an algorithm is readily deduced. First choose an arbitrary H_o and E_o, maximize p over G with these fixed H_o and E_o yielding a new G_1 , maximize subsequently over H with the just computed G_1 and E_o fixed yielding a new H_1 , and finally maximize p over E with G_1 and H_1 fixed yielding a new E_1 , and iterate this procedure until - one hopes - convergence. According to theorem 1 the maximizations are essentially identical to searches for eigenvectors and eigenvalues of matrices of the order l,m, and n respectively. As l,m, and n can be quite large, while s,t, and u are typically very small, say 2,3, or 4, we want to use a technique for solving the eigenvector-eigenvalue problem (or eigenproblem for short) which is particularly efficient in finding just a few eigenvectors.

A very appropriate technique in this situation is the so-called simultaneous iteration method (or Treppen (staircase)iteration) of Bauer-Rutishauser (Rutishauser,1969). For further details on this method see section 6.

The maximization of p consists thus of an in principle infinite iteration process, in which at each step three eigenproblems have to be solved. Clearly solving these eigenproblems by yet another in principle infinite iteration process has its drawbacks. The whole procedure is likely to become computationally burdensome. In order to avoid this we perform only one single step towards the solution of the eigenproblems, instead of the complete iterations. A similar approach has been applied by De Leeuw c.s. in a number of cases when using an ALS technique. Their experience has been that carrying out the complete iteration for solving the eigenproblem only serves to decrease the overall efficiency of the procedure, while it has no effect on the eventual convergence point if one uses only one step (cf. Young, De Leeuw, and Takane, 1979). They suggest that the reason for this behaviour might be found in the same reason that often cause relaxation procedures to be more efficient than non-relaxation procedures.

6. The Bauer-Rutishauser method

As the algorithm is based on the method of Bauer-Rutishauser for computing eigenvectors and eigenvalues it seems in order to describe this method in some detail. In addition, some of the formulations developed here will be used in the rest of the paper.

Let $A \in R^{n\times n}$ be a symmetric positive definite matrix, and p the desired number of eigenvectors. Furthermore let $X \in R^{n\times p}$ be defined as the matrix which has as its columns the iteration vectors. If we write X after k iterations as X_k then the method of Bauer-Rutishauser is described as:

a. Choose an arbitrary orthonormal X_o

b. $Y_k = AX_k$

c. $B_k = Y_k'Y_k$

d. Solve the eigenproblem for B_k, i.e.
determine an orthonormal T_k, and a diagonal L_k with $l_1^k \geqslant l_2^k \geqslant \cdots \geqslant l_p^k$, such that $T_k'B_kT_k = L_k$, and T_k is the eigenvector matrix of B_k, and L_k is the eigenvalue matrix of B_k.

e. $X_{k+1} = Y_kT_kL_k^{-\frac{1}{2}}T_k'$

Schwartz et al.(1968) show that for $k \to \infty$ $L_k^{-\frac{1}{2}}$ converges to the matrix with the largest p eigenvalues of A on the diagonal, and the columns of X_k converge to the associated eigenvectors, provided A is positive definite, and the columns of X are not orthogonal to one or more of the eigenvectors concerned, and in addition the pth and (p+1)st eigenvalues are different.

We may write b. through e. somewhat more concisely as:

$$X_{k+1} = Y_kT_kL_k^{-\frac{1}{2}}T_k' = AX_kB_k^{-\frac{1}{2}} = AX_k(X_k'A^2X_k)^{-\frac{1}{2}}.$$

With a view to what follows it will be convenient to define the function :

$$\phi(X_k) = AX_k(X_k'A^2X_k)^{-\frac{1}{2}} \qquad (6.1)$$

When we use in the sequel a recursive formula like (6.1) we mean to say that X_{k+1} can be computed by carrying out one step of the Bauer-Rutishauser method.

It should be noted that the inverse square root of $X'A^2X$ exists and is uniquely defined, if the expression is positive definite. This implies that ϕ is well-defined, and it can be proven that ϕ is continuous as well (vide Kroonenberg & De Leeuw,1979). As will be shown in section 8 rather strong convergence theorems can be used for the algorithm to be described if ϕ is continuous. It seems therefore worthwhile to take measures in constructing the algorithm to ensure the positive definiteness of $X'A^2X$. An inspection of the method to arrive at (6.1) shows that in fact only the inverse square root is taken of the eigenvalues $l_1, l_2,\ldots,l_p$ of B_k. One only has to check in each iteration step if all eigenvalues are larger than zero, or in pratice larger than some very small number. If one of the eigenvalues is too small, one can restart the iteration procedure with a smaller number of eigenvectors. There is, however, no guarantee that this will solve the singularity problem. On the other hand if no singularities have occurred, one knows that at each step ϕ must have been uniquely defined and continuous on $R^{n\times p}$. As we have taken the above precautions we may from now on assume that expressions like $X'A^2X$ are positive definite.

7. The TUCKALS3 algorithm

Be Z a 1×m×n three-mode matrix, and let s,t, and u be the desired number of components for the three component matrices. Furthermore let $g_1, g_2, \ldots, g_s$, $h_1, h_2, \ldots, h_t$, and $e_1, e_2, \ldots, e_u$ be systems of orthonormal iteration vectors which are combined into the orthonormal matrices $G \in R^{1\times s}$, $H \in R^{m\times t}$, and $E \in R^{n\times u}$ respectively.

If we write the matrices G, H, and E as they are after i iteration steps as G_i, H_i, and E_i, then one main iteration step of the TUCKALS3 algorithm can be described by (7.1) through (7.6):

G substep

$$P_i = Z(H_iH_i' \otimes E_iE_i')Z' \qquad (Z \in R^{1\times mn}) \tag{7.1}$$

$$G_{i+1} = \phi_1(G_i) = P_iG_i(G_i'P_i^2G_i)^{-\frac{1}{2}} \tag{7.2}$$

H substep

$$Q_i = Z(E_iE_i' \otimes G_{i+1}G_{i+1}')Z' \qquad (Z \in R^{m\times nl}) \tag{7.3}$$

$$H_{i+1} = \phi_2(H_i) = Q_iH_i(H_i'Q_i^2H_i)^{-\frac{1}{2}} \tag{7.4}$$

E substep

$$R_i = Z(G_{i+1}G_{i+1}' \otimes H_{i+1}H_{i+1}')Z' \qquad (Z \in R^{n\times lm}) \tag{7.5}$$

$$E_{i+1} = \phi_3(E_i) = R_iE_i(E_i'R_i^2E_i)^{-\frac{1}{2}} \tag{7.6}$$

As mentioned before each G, H, and E substep are one step of an 'inner' iteration to find the eigenvectors of P, Q, and R respectively, and together they define one step of the main iteration.

Because we want to discuss the properties of the TUCKALS3 algorithm in the sequel it is useful to introduce some notation.

-- $F: S \to S$ is a function on S, F defines a complete step of the main iteration, and S is defined as in (3.10)

-- $F = F_3 \cdot F_2 \cdot F_1$ with $F_i: S \to S$ i=1,2,3 s.t.

$$F_1(G_i, H_i, E_i) = (\phi_1(G_i), H_i, E_i) = (G_{i+1}, H_i, E_i)$$

$$F_2(G_{i+1}, H_i, E_i) = (G_{i+1}, \phi_2(H_i), E_i) = (G_{i+1}, H_{i+1}, E_i)$$

$$F_3(G_{i+1}, H_{i+1}, E_i) = (G_{i+1}, H_{i+1}, \phi_3(E_i)) = (G_{i+1}, H_{i+1}, E_{i+1})$$

$$\text{thus } F(s_i) = F(G_i, H_i, E_i) = (G_{i+1}, H_{i+1}, E_{i+1}) = s_{i+1}$$

In section 6 we remarked that ϕ as defined in (6.1) was a continuous function and thus ϕ_1, ϕ_2, and ϕ_3 are continuous functions, and because F is the composite of continuous functions F is continuous as well.

It can be shown that both at each step of the main iteration and at each substep the value is increased. (For a proof that at each substep p is maximized see Kroonenberg & De Leeuw (1979)). Thus $p(F(s_i)) = p(s_{i+1}) \geq p(s_i)$. If p is not increased strictly, i.e. $p(F(s_i)) = p(s_i)$, the algorithm stops. In that case (G,H,E) satisfies the necessary conditions of lemma 3. Consequently we can assume without loss of generality that the algorithm generates an infinite sequence with $p(F(s_i)) > p(s_i)$.

Obviously we need some G_o, H_o, and E_o to intialize the procedure. It seems sensible to choose them in such a way that they are optimal in some sense. We chose such an initialization that it would solve the maximization problem exactly if the problem had such a solution. In other words the eigenvector matrices mentioned in theorem 2 point 4 were used as initializations. Comparing this with the method 1 of Tucker (1966, p. 297) we note that the initialization is nothing but the final Tucker solution. In practice we do not need to know the eigenvectors exactly as they are only used to intialize,and therefore we made only five iteration steps of again the Bauer-Rutishauser method towards their solution.

8. Discussion of the convergence of the TUCKALS3 algorithm

It is of course of prime importance to show that the algorithm outlined in (7.1) - (7.6) converges, and moreover that it converges to a maximum of p, or at least not a minimum.

The algorithm considered here is a type of algorithm that has been described in the non-linear programming literature, and in that field various theorems about the convergence of algorithm such as ours exist. The most appropriate one in our case is the following "fixed point" lemma described and proven by d'Esopo(1959):

Lemma 3

Let F,p,S satisfy the following conditions:

1. a. S is a subset of a finite dimensional real space

   b. F is a continuous transformation of S to S

   c. p is a real function defined and continuous for all $s \in S$

2. $p(F(s)) \geq p(s)$

3. if $p(F(s)) = p(s)$, then $F(s) = s$

4. if the sequence $s_0, s_1, s_2, \ldots$ satisfies $p(s_{i+1}) \geq p(s_i)$ with $s_i \in S$

Then for every limit point $\bar{s}$ of $s_0, s_1, s_2, \ldots$ $F(\bar{s}) = \bar{s}$.

In section 3 we noted and discussed properties 1a,c, and in section 7 we did the same for 1b,2,3, and 4.

We may therefore conclude that the lemma applies to the TUCKALS3 algorithm.

As S is a bounded real subspace, any infinite sequence $s_0, s_1, s_2, \ldots$ is bounded, and thus the sequences generated by the algorithm are bounded as well. A theorem due to Weierstrass shows that such sequences have at least one limit point.

Kroonenberg & De Leeuw (1979) show that every point $\bar{s}$ such that $F(\bar{s}) = \bar{s}$ is a stationary point of p, and because we know that at every step p increases, we know that the stationary points will not be minima.

As has been shown by Ostrowski (1966) the set of limit points of $\{s_i\}$ consists either of a single point or a continuum. The latter case is, however, a very unlikely one in practical applications, as is the occurrence of equal eigenvalues in real data.

The above results imply that from any arbitrary starting point s_o the algorithm converges to a stationary point of p, but the algorithm "like all numerical methods based on local searches for solutions, can at best be expected to yield local minima [here: maxima] (...). Global minimality [here: maximality] could be assured only by exhaustive searches over succesively finer grids" (Meyer, 1970, p.45).

9. Special cases: the Tucker2 and the Tucker1 model

Tucker2 model

An important special case of the Tucker3 model is obtained if the matrix E in the formula (3.5) is taken to be the identity matrix. This defines the Tucker2 model, which can consequently be written as

$$z_{ijk} = \sum_{\alpha=1}^{s} \sum_{\beta=1}^{t} g_{i\alpha} h_{j\beta} c_{\alpha\beta k} \quad i=1,\ldots,l;\ j=1,\ldots,m;\ k=1,\ldots,n \tag{9.1}$$

with the same meaning of, and restrictions on G and H as before.
The matrix notation is

$$Z = GCH' \tag{9.2}$$

where Z and C are three-mode matrices.
Instead of specifying principal components of all three modes, the Tucker2 model only specifies them for two (say the first two) of the three modes. In other words, the third mode is not condensed, and remains intact. This will enable one to study the interrelationships between the components of the other two modes for each element (variable, person, moment in time) of the third mode. The Tucker2 model has been independently formulated by Israelsson (1969), Carroll & Chang (1970), and Jennrich (1970).
The Tucker2 model has three important fields of application:

1. In those analyses of data for which no natural condensation of the third mode can be defined. An obvious example would be the multivariate analysis of time series. In general no useful meaning can be attached to the components of a time mode. In certain other applications one is interested in persons are replications, and one does not want to investigate person components, but rather the interrelationships between the other two modes for each person.
2. In individual differences scaling with asymmetric similarity matrices. A typical example would be the analysis of the confusion matrices of the now famous Miller & Nicely (1955) data. (See section 10).
3. In testing the appropriateness of various individual differences models in multidimensional scaling, such as INDSCAL, IDIOSCAL, and PARAFAC. These models can be seen as a special case of the Tucker2 model. For instance in the INDSCAL model it is assumed that the core matrix is diagonal in each of its frontal planes, i.e.

$$c_{\alpha\beta k} = 0, \text{ if } \alpha \neq \beta$$

If such an assumption is true for the data at hand remains to be seen in many cases.

Technically the estimation of the parameters of the Tucker2 model poses no problem; in the algorithm outlined in section 7 one simply leaves out the E substep, and inserts the identity matrix for E in the other substeps. Computationally it is, however, more efficient to solve the model directly by the analogon of the TUCKALS3 algorithm, than solving the model through the TUCKALS3 algorithm itself. Because of the analogy the proofs of the properties of the TUCKALS2 algorithm are exactly the same as in the TUCKALS3 case. Details are given in Kroonenberg & De Leeuw (1977).

Tucker1 model

Instead of performing a principal component analysis over two or three modes, it is very well feasible to perform such an analysis over just one mode of the data. This would give the Tucker1 model:

$$z_{ijk} = \sum_{\alpha=1}^{s} g_{i\alpha} c_{\alpha jk} \qquad i=1,\ldots,l;\ j=1,\ldots,m;\ k=1,\ldots,n \tag{9.3}$$

with the standard meaning of, and restriction on G.
The matrix formulation is:

$$Z = GC, \tag{9.4}$$

where Z and C are three-mode matrices.
For the case that the horizontal planes of the data matrix Z are similarity matrices, the principal component analysis of the Tucker1 model is identical to the procedure proposed by Tucker & Messick (1963,336ff.). The Tucker1 model has, of course, wider application as it does not restrict the horizontal planes to be similarity matrices.
There is no need to write separate programmes to solve the Tucker1 model as the analysis can be carried out with any principle component analysis programme by properly organizing the data input.

10. Example: the Miller & Nicely data

data

The data from a classical study of confusions of English consonants were used as an example to show a number of the possible features of the TUCKALS programmes. The data consist of confusions among the 16 most used English consonants under 17 different degrading conditions (Miller & Nicely, 1955). Five North American female subjects serves as talkers and listening crew; when one talked, the other four listened. One syllable stimuli consisting of |a| (as in father) preceded by one of 16 consonants were spoken - |p|, |t|, |k|, |f|, |θ| (as in thought), |s|, |ʃ| (as in should), |b|, |d|, |g|, |v|, |ð| (as in that), |z|, |ʒ| (as in vision), |m|, |n|.
The consonants spoken were fed through a transmission circuit which was degraded each of the 17 times in a different way. Notably there were differences in signal-to-noise ratio, low-pass filtering, and high-pass filtering, some details of which are listed in Table 1.

Table 1 here

In each condition tested some 4000 observations were collected, be it that not each consonant was spoken equally often. In our analysis we first corrected for this by dividing each entry by its row total, for each row corresponds to the consonant spoken, while each column corresponds to the consonant heard. The entries in the matrix therefore indicate for that particular condition, the number of times each of the consonants was heard, when the consonant associated with that row was spoken.
In our analysis we added two more matrices to provide zero-point references, i.e. a matrix with perfect discrimination (only entries on the main diagonal), and a matrix with total uniform confusion (equal entries in all cells). Strictly speaking the former matrix does not belong to any of the series of degrading conditions, as perfect discrimination would probably require increasing both the signal-to-noise ratio above 12 db, and extending the frequency range on the high and the low side. With regard to the latter matrix, we could interpret it as referring to noise coming from just one frequency wave band for any signal-to-noise ratio, or as coming from any frequency band with very low signal-to-noise ratio. It therefore would fit into any degradation series.

TABLE 1

Degradation conditions in the Miller & Nicely study of confusions among consonants

Degradation condition	Signal-to-noise ratio	Frequency response in hz	Amount of information per matrix
Masking			
REF1	-	-	4.00
N1	12 db	200-6500	3.55
N2	6 db	200-6500	3.23
N3	0 db	200-6500	2.81
N4	- 6 db	200-6500	1.84
N5	-12 db	200-6500	0.96
N6	-18 db	200-6500	0.06
REF0	-	-	0.00
Low-pass filtering			
REF1	-	-	4.00
L1 (=N1)	12 db	200-6500	3.55
L2 (=H1)	12 db	200-5000	3.20
L3	12 db	200-2500	2.83
L4	12 db	200-1200	2.38
L5	12 db	200- 600	2.18
L6	12 db	200- 400	1.67
L7	12 db	200- 300	1.15
REF0	-	-	0.00
High-pass filtering			
H1 (=L2)	12 db	200-5000	3.20
H2	12 db	1000-5000	2.67
H3	12 db	2000-5000	1.59
H4	12 db	2500-5000	1.07
H5	12 db	3000-5000	0.62
H6	12 db	4500-5000	0.44
REF0	-	-	0.00

Based on Miller & Nicely (1955), and adapted from Carroll & Wish (1974)
Higher numbers denote more severe degradations.
REF1= perfect intelligibility; REF0= total uniform confusion

The Miller & Nicely data have been extensively used both in the field of phonetics as support or disproof of the distinctive feature theory, and as demonstration material for various scaling procedures. In the latter class fall most notably Shepard (1972,1974), Wish(1970), Carroll & Wish(1974), Smith (1973), and Smith & Jones(1975).
Substantially with respect to the structure of the consonant space we have not much to add to the very detailed and thoughtful analysis of Shepard(1972). We give, however, a new interpretation of the dimensions in the noise-condition space. We want to emphasize that our primary aim in presenting the analysis of these data is to demonstrate the developed computer programmes, rather than provide a substantial contribution to theory in other domains. At the same time it should be realized that it is impossible in the present context to do full justice to all the various aspect of the two programmes.

stimulus spaces

Inspection of the Tucker2 and Tucker3 models shows that the principal component matrices G of the first mode and H of the second mode, respectively consonants spoken and consonants heard are treated independently. There is therefore the possibility to compare these two configurations, based on the same 16 consonants. It turned out that only small differences were present - an interesting phenomenon in itself - therefore we will discuss for the moment the stimulus spaces, as if they were one, and come back to their difference later on.
In principal component analysis the number of components to retain is a primary problem, essentially it amounts to deciding how much distortion one allows of the original stimulus space. In three-mode principal component analysis the situation is the more problematic, because of the interwovenness of the three modes in the estimation procedure. Changing the number of components in one mode implies immediately a different (whether substantial or not) solution of the other modes. It was to us, however, surprising how stable many solutions were to such changes.
The usual criterion based on the amount of variance explained by the components is again problematic, as the components are eigenvectors not of the original inproducts of the data, like the eigenvectors G, H, and E in theorem 2 but they are eigenvectors of P, Q, and R (see section 4) which all are functions of the

other modes as well. At this moment we have not worked out a satisfactory guide yet, and in the case of the Miller & Nicely data we have used both the inspection-of-eigenvalues criterion and the alltoo fallible interpretation criterion.
In interpreting the output of analyses such as ours one can look for homogeneous group of variables (here: groups of consonants which are very often confused) or search for meaningful directions (axes) in the stimulus space. We have used both approaches simultaneously.Especially the consonant stimulus space has both very clear interpretational axes and homogeneous subgroups, as can be seen from table 2 and figure 3.

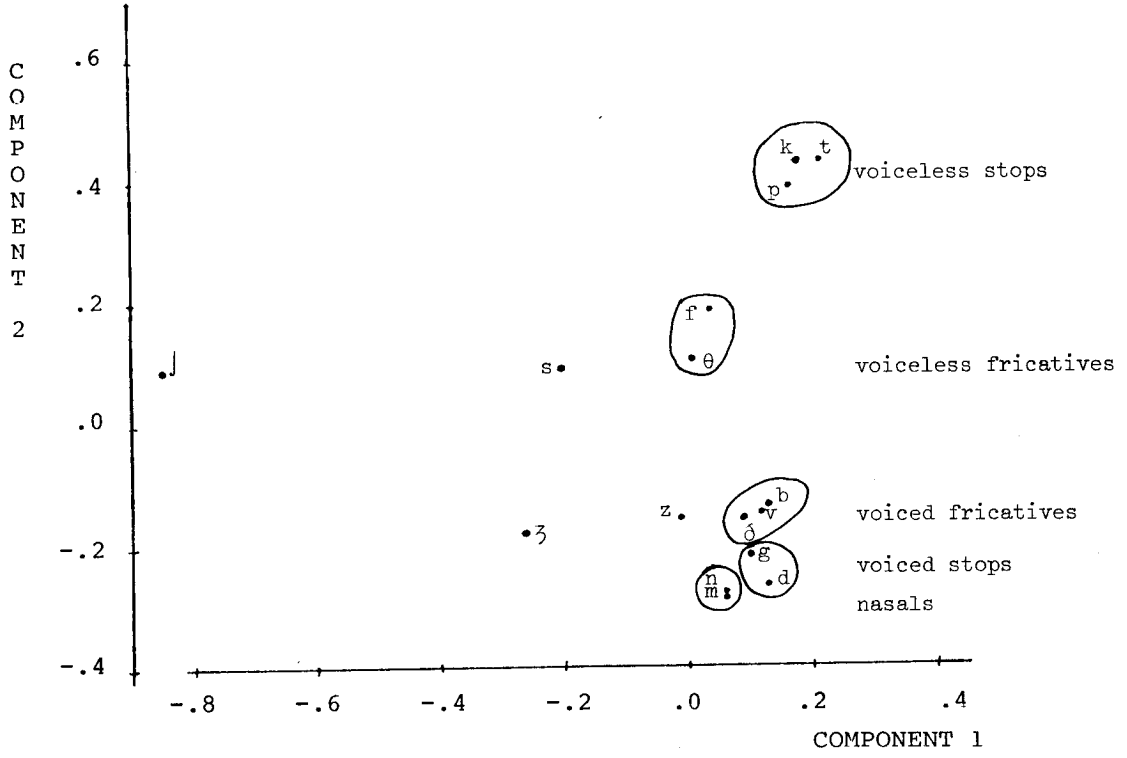
Table 2 and figure 3 here

Figure 3 shows the rotated solution in three dimensions. The first component roughly corresponds to the energy content of the various components, the |ʃ| being the most energetic consonant, while the voiceless stops are the least energetic; we will refer to this component as "energy". The second axis separates the voiceless stops, voiceless fricatives, voiced fricatives and voiceless stops (except for |b|): we will refer to this axis as "voice", although it is clear that this is an oversimplification. The third axis, finally, serves to set the nasals off from the rest, and we will refer to it as "nasality". Furthermore we can easily identify a number of homogeneous subgroups as is indicated in figure 3. A four dimensional solution did not give any additional information about the group structure, it only served to set off the voiceless stops and voiceless fricatives against each other, a distinction already contained in the three dimensional configuration. A problem is, of course, to separate the group of consonants more formally. If a some type of average similarity matrix would have been available an appropriate cluster analysis could be called to assistance (see e.g. Shepard (1972) for the use of such a procedure on the same data after the symmetrization of the matrices). In our case we have used another feature of the TUCKALS programmes as a rough guide. The programmes generate an "average" matrix between the elements of each two of the three modes on the basis of the components. In table 4 this "average" confusion matrix - GCH' - is given.

Table 4 here

A visual analysis indicates four major, partially overlapping clusters, and some

FIG. 3 STIMULUS SPACE FOR MODE 1 (CONSONANTS SPOKEN) - rotated



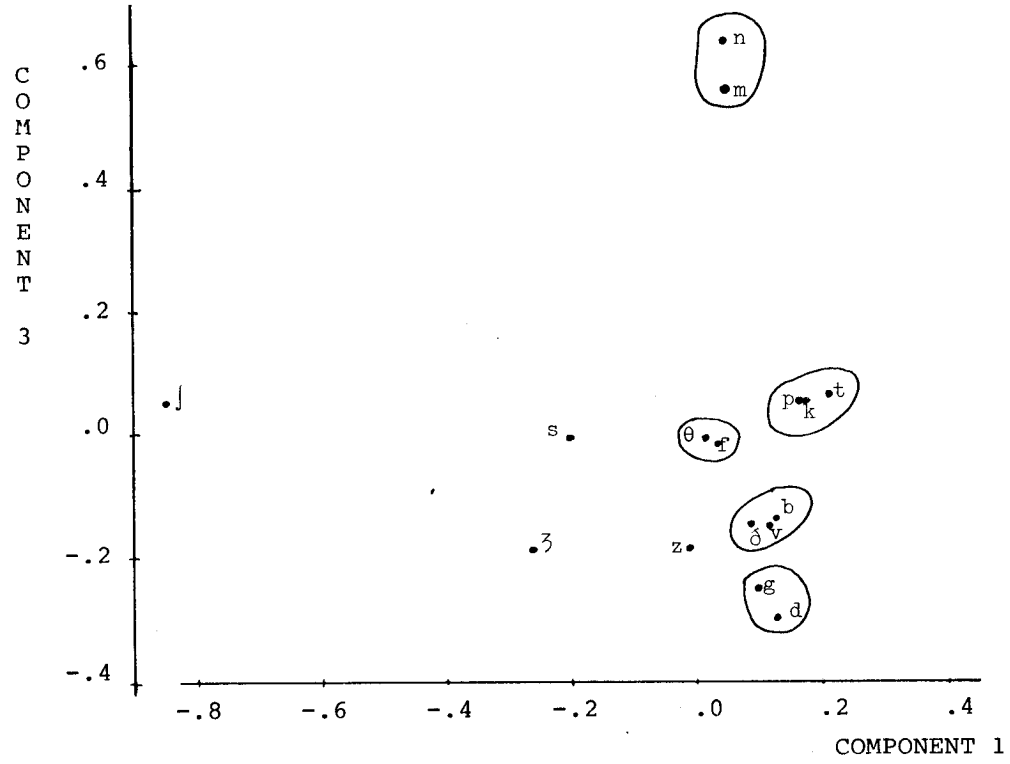


TABLE 4

"average" confusion matrix

consonants heard

consonants spoken	t	k	p	f	θ	s	ʃ	ʒ	z	ð	v	b	g	d	n	m
t	11	11	10	4	3	1	-5	-7	-4	-3	-3	-2	-4	-6	-4	-4
k	11	11	10	4	3	1	-4	-6	-4	-2	-3	-2	-4	-5	-4	-4
p	10	10	9	4	3	1	-4	-6	-4	-2	-2	-2	-4	-5	-3	-4
f	4	4	4	2	1	1	-0	-2	-3	-1	-1	-1	-1	-2	-3	-3
θ	3	3	3	1	1	1	1	-1	-3	-1	-1	-1	-1	-2	-2	-3
s	1	1	1	1	1	3	9	1	-1	-2	-2	-2	-2	-3	-3	-3
ʃ	-5	-4	-4	-0	2	9	33	8	-1	-4	-6	-6	-5	-8	-4	-4
ʒ	-7	-7	-6	-2	-1	2	8	6	3	1	1	1	3	4	-3	-4
z	-4	-4	-4	-1	-1	-1	-1	3	3	2	3	2	4	5	-3	-3
ð	-3	-3	-3	-1	-1	-2	-5	2	2	2	3	3	3	5	-1	-2
v	-2	-3	-2	-1	-1	-2	-6	1	2	2	3	3	4	5	-1	-2
b	-2	-2	-2	-1	-1	-2	-6	1	2	2	3	3	3	5	-1	-2
g	-4	-5	-4	-2	-1	-2	-5	3	4	3	4	4	5	7	-3	-4
d	-5	-5	-5	-2	-2	-3	-7	3	4	4	5	5	6	9	-4	-5
n	-4	-4	-3	-3	-2	-3	-4	-3	-3	-1	-2	-1	-3	-4	18	21
m	-4	-4	-3	-3	-2	-3	-4	-4	-3	-2	-2	-2	-3	-5	20	24

Notes:- the "average" confusion matrix is constructed on the basis of the two component matrices G and H, and the average frontal plane of the core matrix, i.e. $G\bar{C}H'$

- each entry indicates the weighted product of the row stimulus and the column stimulus. High positive values indicate that the row and column stimulus are often confused, and large negative values indicate that they are very seldom confused. High values on the main diagonal indicate that the consonant is very distinct, and is seldom confused with other consonants in nearly all degradation conditions.
- the order of the consonants is different from the one in Miller & Nicely.
- decimal points omitted

further distinction within the major clusters. At the same time we can see that the second component "voice" corresponds very close to the entries on the main diagonal of table 4 (disregarding the | s |, | ʃ | , | m |, | n |).
The above analysis has been carried out on per noise condition double centred matrices (i.e. $\hat{z}^k_{ij} = z^k_{ij} - z^k_{.j} - z^k_{.i} + z^k_{..}$), and tap therefore the main structure of the data.
Another way of looking at the same data using the same programme, is to centre the data for each consonant heard- consonant spoken combination over all noise conditions (a procedure for instance used by Tucker & Messick (1963) in their point-of-view analysis). This will bring to light if the various noise conditions treat the consonants differently. For instance, a consonant which behaves more or less similarly in all conditions, will now be located close to the centre of the configuration; on the other hand a consonant (such as | t |) which is treated differently for high-pass filters from low-pass filters will have a high loading on some component. Figure 5 illustrates this clearly for the Miller & Nicely data.

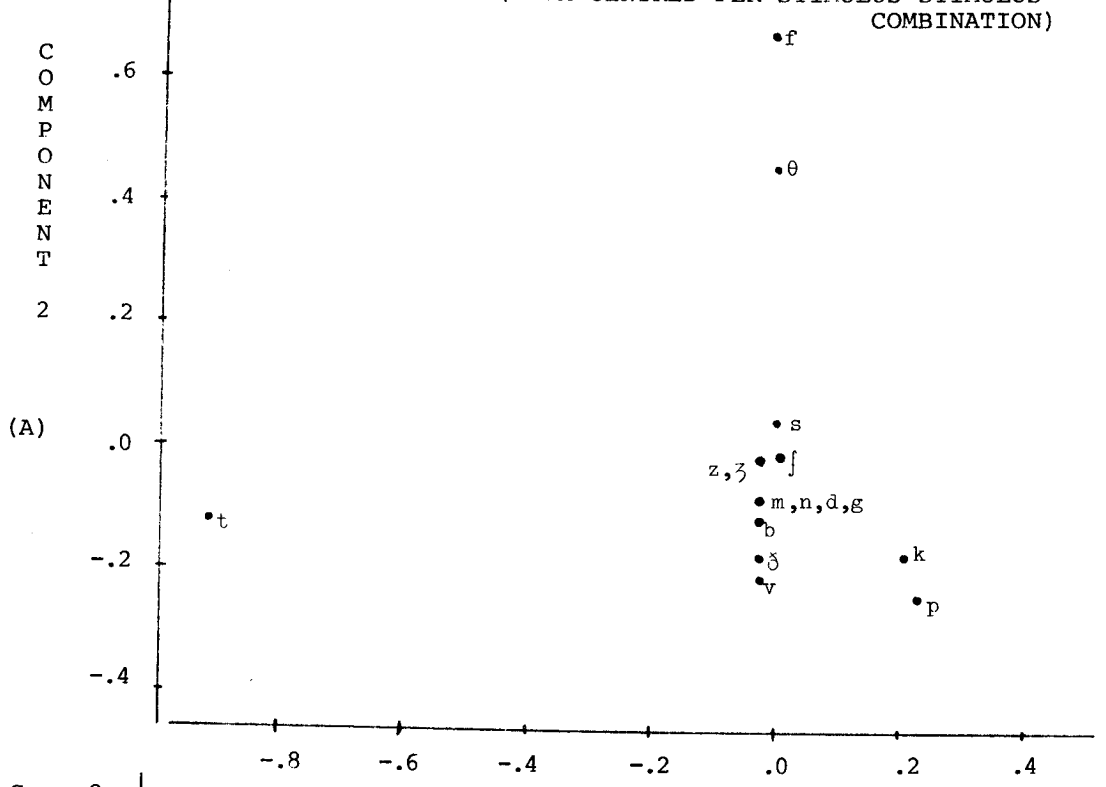
Figure 5 here

Figure 6 is a joint plot of the component matrices for the first and the second mode of the first analysis showing how well the correspondence is between the understood and the spoken consonants. (For details on the method to produce these plots, see Kroonenberg & De Leeuw, 1977).

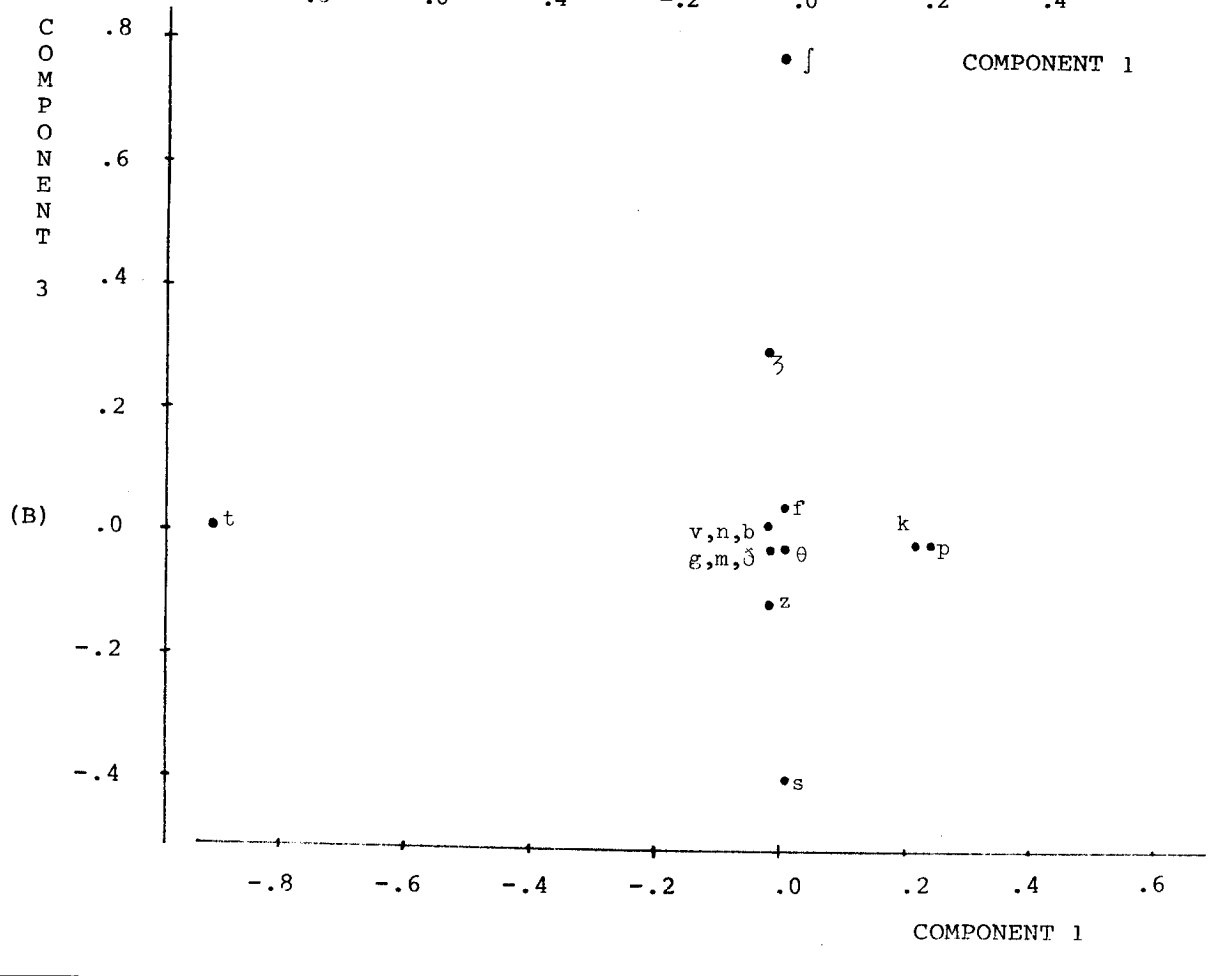
Figure 6 here

In passing it should be noted that for the same number of components the TUCKALS2 and TUCKALS3 solutions were virtually identical. Finally it should be remarked that our substantive results resemble those of Shepard (1972) quite closely. The INDSCAL analysis of Carroll & Wish (1974) seems to us to involve far too many dimensions, and our impression is that they need six dimensions to separate the various groups, where we need only three. The fact that they performed the analysis on the raw data might have something to do with this.

COMBINATION)



(A)



(B)

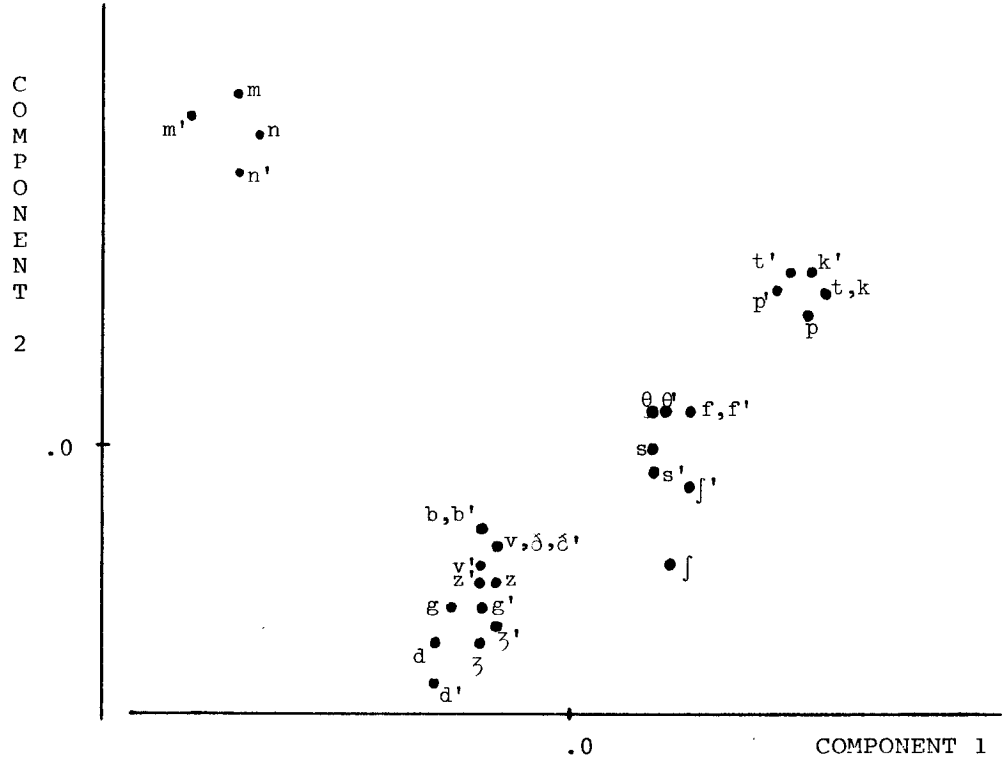


FIG. 6 JOINT PLOT OF THE STIMULUS SPACES OF CONSONANTS SPOKEN AND CONSONANTS HEARD (indicated with apostrophe)

Note: third component not shown

noise condition

As is shown in table 1 we can define three large groups of noise conditions - varying signal-to noise ratios, filtering the high frequencies, and filtering the low frequencies respectively. As mentioned by Miller & Nicely (1955) and confirmed by Shepard(1972) low-pass filters and low signal-to-noise conditions look somewhat alike, and are both different from high-pass filters. A two dimensional rotated solution is given in table 7, and figure 8.

Table 7 and Figure 8 here

The stimulus space for the noise conditions has been rotated in such a way that one of the axes passes through the two reference points, i.e. uniform total confusion and perfect intelligibility.
In the original publication Miller & Nicely use a measure of covariance between input and output to classify the various noise conditions. In particular this measure is

$$T(x,y) = -\sum_{i,j} p_{ij} \log \frac{p_i p_j}{p_{ij}} ,$$

where T is often referred to as the amount of information transmitted from input variable x to the output variable y in bits per stimulus, and where it is assumed that x takes on discrete values (here consonants) i=1,2,..,k with probability p_i, and similarly y takes on the values j=1,2,...,k with probability p_j, and p_{ij} is the probability of the joint occurrence of input i and output j. We have recalculated the values of T for the confusion matrices based on proportions, and these values are listed in table 1.
By a heuristic method a direction in the noise condition space can be found which corresponds to (a nonlinear transformation of) the amount of information contained in each matrix (see figure 9B). This direction is indicated in figure 8. One would have preferrred this direction to be one of the axes of figure 8, but this is unfortunately not the case, as can be clearly seen from figure 9A.

Figure 9 here

It is , however, interesting to note that the higher

TABLE 7

Stimulus space of the third mode (noise conditions)

	masking			low-pass filtering			high-pass filtering	
	1	2		1	2		1	2
REF1	30	0	REF1	30	0			
N1	30	0	L1	30	0	H1	30	1
N2	30	1	12	30	1	H2	30	- 7
N3	29	3	L3	28	6	H3	26	-20
N4	24	9	L4	18	32	H4	24	-28
N5	13	25	L5	18	33	H5	21	-40
N6	1	7	L6	18	27	H6	18	-42
REF0	0	0	L7	9	44	REF0	0	0
			REF0	0	0			

Notes:-decimal points omitted

-component weights : .31 and .01 (unrotated)
-stimulus space has been rotated, such that the first component runs through the reference points, REF0 and REF1
-sum of all unrotated weights is equal to one

FIG. 8 NOISE CONDITION SPACE - rotated

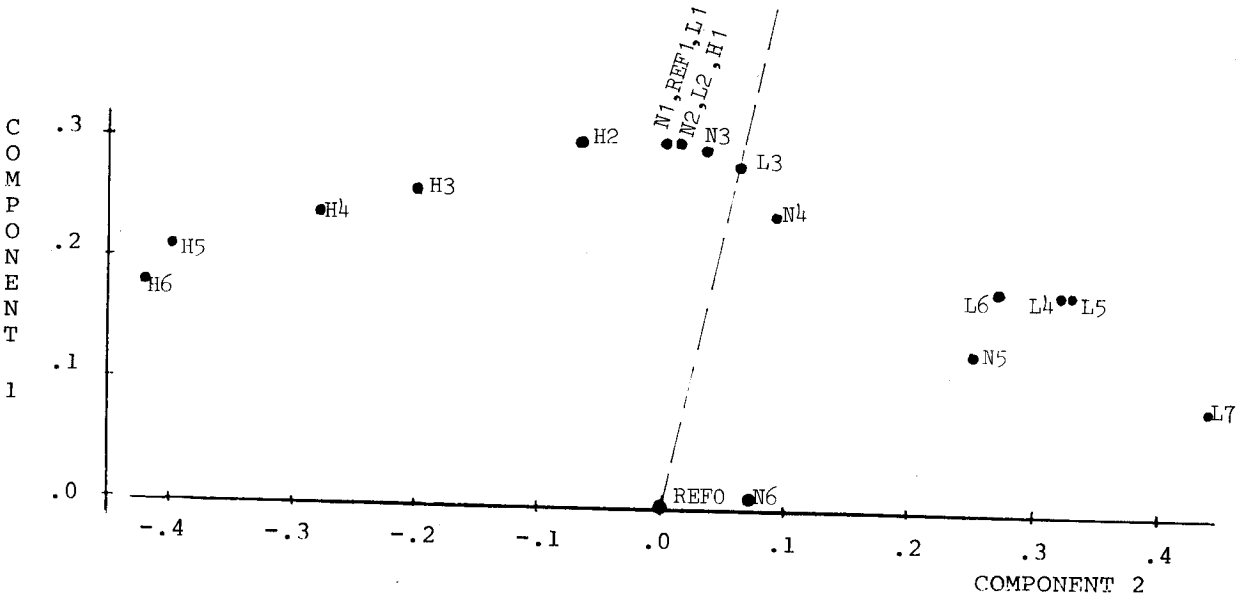
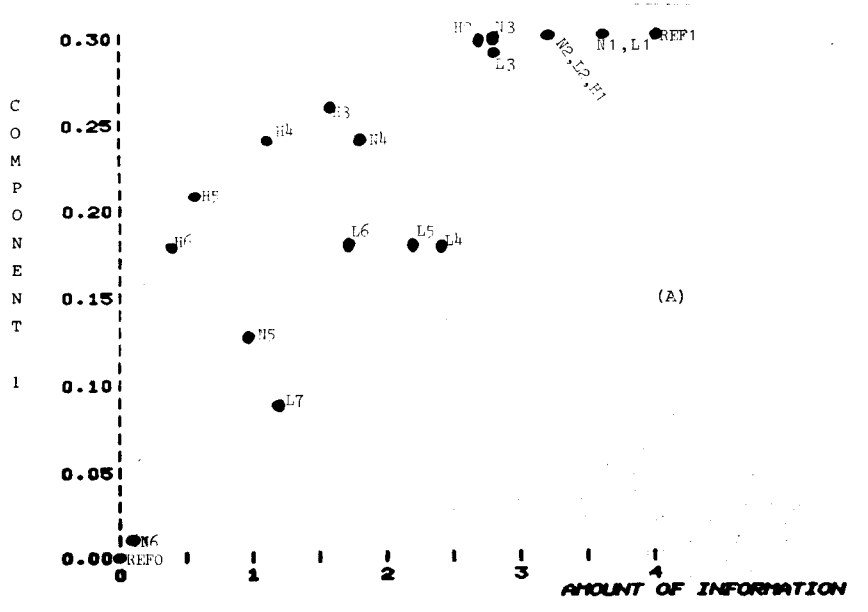
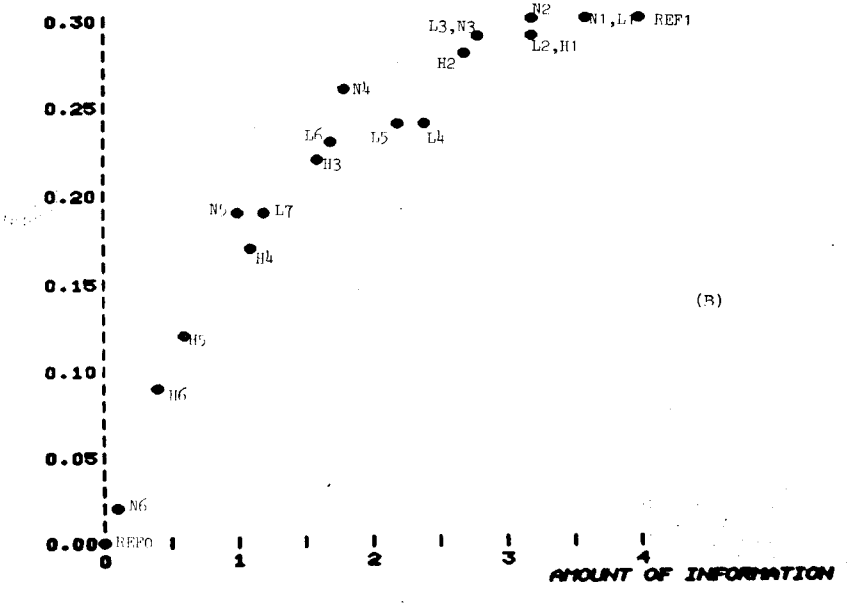


FIG. 9 RELATION BETWEEN THE AMOUNT OF INFORMATION AND DIRECTIONS IN THE NOISE CONDITION SPACE





points in figure 10 A are all high-pass filtering conditions. A particular feature distinguishes these filter conditions from the other conditions, i.e. the amount of information in the confusion matrix is transmitted by just a few stimuli (|ʃ|, |ʒ|, |s|, and |t|). Inspection of the core matrix (see below) shows that, in addition, the "energy"-axis of the consonant space stands out in the amount-of-information plane of the core matrix. Conceiveably, therefore, this feature might have caused their different behaviour.
As far as the other component of the (rotated) noise condition space is concerned, it reflects something like the average frequency of the filtering or masking but a proper measure to account for the numerical values is not known to us.

core matrices

Finally we should say something about the core matrices, both of the TUCKALS3 and TUCKALS2 analysis. As mentioned above we have in the former analysis performed a number of rotations. The stimulus spaces for the consonants were rotated in such a way that the frontal planes of the core matrix were far more diagonal than before. At the same time this improved the interpretability of the axes of the stimulus spaces of the consonants. The above mentioned rotations of the noise condition space - here through point L3 - is of course also compensated by a counter rotation of the core matrix. The final effect of these three rotations on the core matrix is shown in figure 10, where the frontal planes are shown, and where, as far as possible, the appropriate labels of the components have been added.

Figure 10 here

The main pattern of the frontal planes is that each of the components of the first mode (consonants spoken)is predominantly related to the same one of the second mode (consonants heard). Secondly all components have their largest loadings on the "average frequency of unfiltered band" frontal plane, and thirdly that the "energy" component seems to be the only one substantially contributing to the amount-of-information distinctions, while all three components nearly play an equal role in the frequency distinction. We are unfortunately not versed enough in the substantial theory of filtering to further interpret these findings. The frontal planes of the core matrix in the TUCKALS2 analysis provide us with the relations between the components of the first and second mode for each of the degradation conditions. In a sense these planes provide a summary of

how the noise conditions affect the three major components of the stimulus space of the consonants. In table 11 some typical examples are given.

TABLE 10

Frontal planes of core matrix from the TUCKALS3 analysis

"amount of information" (component 1 noise conditions)

		energy 1	voice 2	nasality 3
energy	1	13	- 0	- 1
voice	2	1	5	0
nasality	3	1	- 0	4

"average frequency of unfiltered band" (component 2 noise conditions)

	energy 1	voice 2	nasality 3
1	18	- 2	- 2
2	2	23	1
3	2	- 2	23

TABLE 11 Frontal planes of core matrix from the TUCKALS2 analysis - rotated

	masking			low-pass filtering			high-pass filtering			reference matrices		
	N6 -18db;200-6500			L7 12db; 200- 300			H6 12db;4500-5000			REF1 perfect intelligibility		
	e 1	v 2	n 3	e 1	v 2	n 3	e 1	v 2	n 3	e 1	v 2	n 3
energy 1	1	- 1	1	15	- 5	2	45	- 2	0	67	1	- 0
voice 2	- 2	9	- 2	- 3	57	- 3	- 2	10	1	- 1	67	0
nasality 3	0	2	7	1	- 3	57	2	1	2	0	- 0	67
	N5 -12db;200-6500			L4 12db; 200-1200			H3 12db;2000-5000			REF0 total uniform confusion		
energy 1	26	- 4	3	36	2	3	60	1	4	0	0	0
voice 2	- 1	54	- 4	5	65	- 1	1	47	- 4	0	0	0
nasality 3	3	- 4	47	1	- 1	66	2	0	37	0	0	0

Note : - all values have been multiplied by 10

REFERENCE NOTES

Carroll, J. D. & Chang, J. J. IDIOSCAL: A generalization of INDSCAL allowing IDIOsyncratic reference systems as well as an analytic approximation to INDSCAL. Paper presented at the Spring Meeting of the Psychometric Society, Princeton, N.J., March, 1972

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

Jennrich, R. A generalization of the multidimensional scaling model of Carroll & Chang. (Working Papers in Phonetics No.22). Los Angelos: University of California, 1972

Kroonenberg, P. M. & De Leeuw, J. TUCKALS2: A principal component analysis of three mode data. (Research Bulletin, RB 001-77). Leiden: Department of Data Theory, University of Leiden, 1977

Kroonenberg, P. M. & De Leeuw, J. Principal component analysis of three-mode data by means of alternating least-squares algorithms including proofs. (Research Bulletin, RB 001-79). Leiden: Department of Data Theory, University of Leiden, 1979

Levin, J. Three mode factor analysis. (Unpublished doctoral thesis). Urbana, Ill.: University of Illinois, 1963

Sands, R. Component models for three-way data: ALSCOMP3, an alternating least-squares algorithm with optimal scaling features. (Unpublished master's thesis). Chapel Hill, N. C.: University of North Carolina, 1978

Wish, M. An INDSCAL analysis of the Miller & Nicely consonant confusion data. Paper presented at meetings of the Acoustical Society of America. Houston, November, 1970

REFERENCES

Carroll, J. D. & Chang, J. J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika, 1970, 35, 283-320

Carroll, J. D. & Wish, M. Models and methods for three-way multidimensional scaling. In D. H. Krantz et al. (Eds.) Contemporary Developments in Mathematical Psychology, Vol. II. San Francisco: W. H. Freeman & Cy, 1974

Cochrane, D. & Orcutt, G. H. Application of least squares regression to relationships containing autocorrelated error terms. Journal of the American Statistical Association, 1969, 44, 32-61

d'Esopo, D. A. A convex programming procedure. Naval Research Logistics Quarterly, 1959, 11, 33-42

Israelsson, A. Three-way (or second order) component analysis. In H. Wold & E. Lyttkens (Eds.) Nonlinear iterative partial least-squares (NIPALS) estimation procedures. Bulletin of the International Statistical Institute, 1969, 43, 29-51

Meyer, R. R. The validity of a family of optimization methods. SIAM Journal of Control and Optimization, 1970, 15, 699-715

Miller, G. A. & Nicely, P. E. An analysis of perceptual confusion among some English consonants. Journal of the Acoustical Society of America, 1955, 338-352

Osgood, C. E. , Suci, G. J. & Tannenbaum, T. H. The measurement of meaning. Urbana, Ill.: University of Illinois Press, 1957

Ostrowski, A. M. Solution of equations and systems of equations. New York: Academic Press, 1966

Penrose, R. On the best approximate solutions of linear matrix equations. Proceedings of the Cambridge Philosophical Society, 1955, 51, 406-413

Rutishauser, H. Computational aspects of F. L. Bauer's simultaneous iteration method. Numerische Mathematik, 1969, 13, 4-13

Schwartz, H. R., Rutishauser, H. & Stiefel, E. Numerik. Symmetrischer Matrizen. Stuttgart: Teubner, 1968

Shepard, R. N. Psychological representation of speech sounds. In E. E. David & P. B. Denes (Eds.) Human Communication. A unified view. New York: McGraw Hill, 1972

Shepard, R.N. Representation of structure in similarity data: problems and prospects. Psychometrika, 1974, 39, 373-421.

Smith, P.T. Feature-testing models and their application to perception and memory for speech. Quarterly Journal of Experimental Psychology, 1973, 25, 511-534.

Smith, P.T. & Jones, K.F. Some hierarchical scaling methods for confusion matrix analysis II. Applications to large matrices. British Journal of Mathematical and Statistical Psychology, 1975, 28, 30-45.

Takane, Y., Young, F. & De Leeuw, J. Non-metric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. Psychometrika, 1977, 42, 7-67.

Tucker, L.R. Implications of factor analysis of three-way matrices for measurement of change. In: C.W. Harris (Ed.) Problems in measuring change. Madison, Wis.: Univ. Wisconsin Press, 1963, 122-137.

Tucker, L.R. The extension of factor analysis to three-dimensional matrices. In: Gulliksen, H. and Frederiksen, N. (Eds.) Contributions to Mathematical Psychology, New York: Holt, Rinehardt & Winston, 1964.

Tucker, L.R. Some mathematical notes on three-mode factor analysis. Psychometrika, 1966, 31, 279-311.

Tucker, L.R. Relations between multidimensional scaling and three-mode factor analysis. Psychometrika, 1972, 37, 3-27.

Tucker, L.R. & Messick, S. An individual difference model for multidimensional scaling. Psychometrika, 1963, 28, 333-367.

Wainer, H. Gruvaeus, G. & Blair, M. TREBIG: A 360/75 FORTRAN program for three-mode factor analysis for big data sets. Behavioral Research Methods and Instrumentation, 1974, 6, 53-54.

Wainer, H. Gruvaeus, G. & Snijder, F. TREMOD: A 360/75 program for three mode factor analysis. Behavioral Science, 1971, 16, 421-422.

Walsh, J.A. An IBM 709 program for factor analyzing three-mode matrices. Educational and Psychological Measurement, 1964, 24, 669-773.

Walsh, J.A. & Walsh, R. A revised Fortran program for three-mode factor analysis. Educational & Psychological Measurement, 1976, 36, 169-170.

Young, F. W., De Leeuw, J. & Takane, Y. Quantifying qualitative data. In H. Feger (Ed.) Similarity and choice. New York: Academic Press, 1979 (in press)

PART II

PROOFS

Introduction

In part II of this report proofs are given of a number of statements made in part I.
In order to make part II slightly more readable a number of definitions which also appear in part I have been repeated in sections 1 and 3. For the rationale behind the definitions we refer the reader to the first part of this report.

1. Notation and terminology

1.1 Matrices

All matrices used in this report are real, and in general the number of rows will be larger than or equal to the number of columns.
The following classes of matrices will be used throughout:

$R^{n\times m}$: the class of real $n\times m$ matrices

$K^{n\times m}$: the class of all columnwise orthonormal matrices, i.e. if $U \in K^{n\times m}$ then $U'U = I_m$ if and only if $n=m$; the rank of any $U \in K^{n\times m}$ is m

$D^{n\times m}$: the class of all $n\times m$ diagonal matrices; $E \in D^{n\times m}$ is a 'truly' diagonal matrix if $n=m$; if $n < m$ then $E = (F\ O)$ with $F \in D^{m\times m}$ and O is a matrix with zeroes; if $n > m$ then $E = \begin{pmatrix} F \\ O \end{pmatrix}$ with $F \in D^{m\times m}$.

I_n : the $n\times n$ identity matrix

A $l\times m\times n$ block matrix Z is defined as the collection of elements:

$$\{z_{ijk} \mid i=1,\ldots,l;\ j=1,\ldots,m;\ k=1,\ldots,n\}\ .$$

These elements can be thought to be placed in a three dimensional block with the index i running along the vertical axis, the index j running along the horizontal axis, and the index k along the 'depth' axis.

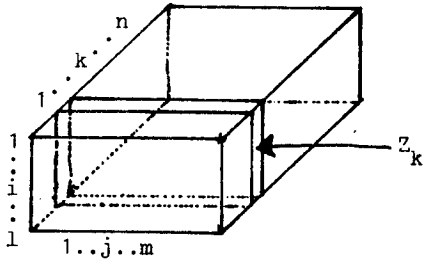


Fig. 1-1
A three-mode matrix

We will also view a block matrix as collections of normal (two-mode) matrices. This can be done in three different ways:

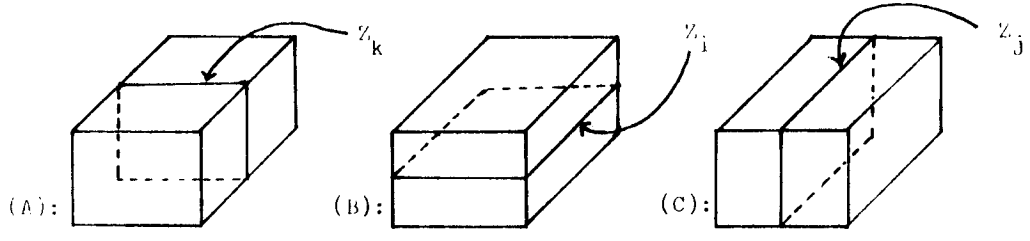


Fig. 1-2
Three different ways to view a three-mode matrix Z as a collection of two-mode matrices.

a: the collection of frontal planes : $Z = \{Z_k\}_{k=1,\ldots,n}$ fig. 1-2A

b. the collection of horizontal planes: $Z = \{Z_i\}_{i=1,\ldots,l}$ fig. 1-2B

c. the collection of lateral planes: $Z = \{Z_j\}_{j=1,\ldots,m}$ fig. 1-2C

1.2 Stationary points

For convenience we will use in this report a more restricted definition of a stationary point of a function than is customary:

Definition:

Let $S = \{(G,H,E) \mid G \in K^{l\times s}, H \in K^{m\times t}, E \in K^{n\times u}\}$, and let p be a real continuous differentiable function on S then $(\hat{G},\hat{H},\hat{E}) \in S$ is a stationary point of p if $(\hat{G},\hat{H},\hat{E})$ is a solution of the stationary equations:

$$\frac{\delta}{\delta X}\left[p(G,H,E) - trL(G'G - I_s) - trM(H'H - I_t) - trN(E'E - I_u)\right] = 0$$

with X = G, H, E, L, M, N respectively, and L, M, and N being matrices of Lagrange multipliers.

2. Analytic solutions of the principal component problem

approximate solution

Theorem 1

Be 1. Z a three-mode matrix,

2. $p(G,H,E) = \text{tr } G'\{Z(HH' \otimes EE')Z'\}G$, (2.1)

3. $P(H,E) = Z(HH' \otimes EE')Z'$ with $Z \in R^{l \times mn}$, (2.2)

$Q(E,G) = Z(EE' \otimes GG')Z'$ with $Z \in R^{m \times nl}$, (2.3)

$R(G,H) = Z(GG' \otimes HH')Z'$ with $Z \in R^{n \times lm}$, (2.4)

4. $S = \{s \mid s = (G,H,E),\ G \in K^{l \times s},\ H \in K^{m \times t},\ E \in K^{n \times u}\}$, (2.5)

5. U is an eigenvector matrix of P with rank s,
V is an eigenvector matrix of Q with rank t,
W is an eigenvector matrix of R with rank u,

Then:

a. $(\hat{G},\hat{H},\hat{E}) \in S$ is a stationary point of p if and only if $\hat{G} = U$, $\hat{H} = V$, and $\hat{E} = W$, or orthonormal rotations thereof.

b. $(\hat{G},\hat{H},\hat{E}) \in S$ maximizes p if and only if their columns are the eigenvectors corresponding to the largest s,t, and u eigenvalues of $P(\hat{H},\hat{E})$, $Q(\hat{E},\hat{G})$, and $R(\hat{G},\hat{H})$ respectively or orthonormal rotations thereof.

Proof:

a. 1. Let us first determine the stationary equations for

$$p(G,H,E) = \text{tr } G'PG = \text{tr } H'QH = \text{tr } E'RE$$

Incorporating the constraints on the parameter space into the function to be maximized, we get:

$$p(G,H,E,L,M,N) = \text{tr } G'PG - \text{tr } L(G'G - I_s) - \text{tr } M(H'H - I_t) - \text{tr } N(E'E - I_u).$$

Differentiating with respect to all the parameter matrices, and setting all the derivatives equal to zero, we obtain the following set of equations, which have to be solved simultaneously for all the parameter matrices:

$$P(\hat{H},\hat{E})\hat{G} = \hat{G}\hat{L} \quad \& \quad \hat{G}'\hat{G} = I_s \tag{2.6}$$

$$Q(\hat{E},\hat{G})\hat{H} = \hat{H}\hat{M} \quad \& \quad \hat{H}'\hat{H} = I_t \tag{2.7}$$

$$R(\hat{G},\hat{H})\hat{E} = \hat{E}\hat{N} \quad \& \quad \hat{E}'\hat{E} = I_u \tag{2.8}$$

To simplify the notation we will drop the carets from now on. Note that L, M, and N are necessarily symmetric, because e.g. the restriction $g_i'g_j = \delta_{ij}$ is identical to the restriction $g_j'g_i = \delta_{ji}$, where g_i is the i-th column of G.

2. As G and L are solutions of (2.6), it follows from PG = GL that L = G'PG. Furthermore L is positive definite, because P is, and because in addition L is symmetric, there exists an $F \in K^{s\times s}$ such that $L = F\Lambda F'$ with $\Lambda \in D^{s\times s}$. Substituting this in (2.6), and postmultiplying with F we get PGF = GFΛ.
By defining U = GF (and thus G = UF') it follows that Λ is an eigenvalue matrix of P, and U is the associated eigenvector matrix.
Analoguously it follows that:
   - if H and M are solutions of (2.7), then there exists an $\bar{F} \in K^{t\times t}$ such that $H = V\bar{F}'$ and $M = \bar{F}\bar{\Lambda}\bar{F}'$; in other words H is an orthonormal transformation of an eigenvector matrix of Q;
   - if E and N are solutions of (2.8), then there exists an $\tilde{F} \in K^{u\times u}$ such that $E = \tilde{W}\tilde{F}'$ and $N = \tilde{F}\tilde{\Lambda}\tilde{F}'$; in other words E is an orthonormal transformation of an eigenvector matrix of R.

3. Conversely, if we let U,V,W, and $\Lambda,\bar{\Lambda},\tilde{\Lambda}$ be eigenvector matrices and eigenvalue matrices of P, Q, and R respectively, then (U,V,W) as well as their orthonormal transformations (G,H,E) with $G = UF'$, $H = V\bar{F}'$, $E = \tilde{W}\tilde{F}'$, $F \in K^{s\times s}$, $\bar{F} \in K^{t\times t}$, and $\tilde{F} \in K^{u\times u}$ satisfy (2.6) through (2.8), and are thus stationary points of p.

b. 1. Define:

$$T = \{t \mid t = (G,H,E);\ G,\ H,\ \text{and } E \text{ are eigenvector matrices of P, Q, and R respectively, or orthonormal transformations thereof}\}$$

We already know that there exists a $(G,H,E) \in S$ such that p attains its maximum (see Part I , section 3). Now we

can state that this maximum will and can only be attained for some $(G,H,E) \in T$.

2. Any $(G,H,E) \in T$ can be written as $(UF', V\bar{F}', W\tilde{F}')$ with $F \in K^{s\times s}$, $\bar{F} \in K^{t\times t}$, and $\tilde{F} \in K^{u\times u}$.

Thus: $p(G,H,E) = p(UF', V\bar{F}', W\tilde{F}') =$

$= \operatorname{tr} FU'Z(V\bar{F}'\bar{F}V' \otimes W\tilde{F}'\tilde{F}W')Z'UF' =$

$= \operatorname{tr} U'Z(VV' \otimes WW')Z'U = p(U,V,W)$

3. Let $(\hat{G},\hat{H},\hat{E}) \in T$ be the point at which p attains its maximum:

$p(\hat{G},\hat{H},\hat{E}) = p(\hat{U},\hat{V},\hat{W}) = \max_U p(U,\hat{V},\hat{W}) = \max_U \operatorname{tr} U'P(\hat{V},\hat{W})U =$

$= \max_\lambda \sum_{i=1}^{s} \lambda_i$ with the maximum taken over all possible ways to combine s of the total of l eigenvalues of $P(\hat{V},\hat{W})$.

$= \sum_{i=1}^{s} \hat{\lambda}_i$ where $\hat{\lambda}_i$ $(i=1,\ldots,s)$ are the s largest eigenvalues of $P(\hat{V},\hat{W})$.

Thus $\hat{U}$ must be the eigenvector matrix associated with these largest eigenvalues. Analoguously $\hat{V}$ and $\hat{W}$ are the eigenvector matrices associated with the largest eigenvalues of $Q(\hat{W},\hat{U})$ and $R(\hat{U},\hat{V})$.

The value of the maximum is:

$$\sum_{i=1}^{s} \hat{\lambda}_i = \sum_{j=1}^{t} \hat{\mu}_j = \sum_{k=1}^{u} \hat{\nu}_k ,$$

where $\hat{\mu}$ and $\hat{\nu}$ are analoguously defined as $\hat{\lambda}$.

4. Conversely, let $\hat{U},\hat{V},\hat{W}$ be the eigenvector matrices associated with the largest s, t, and u eigenvalues of $P(\hat{V},\hat{W})$, $Q(\hat{W},\hat{U})$, and $R(\hat{U},\hat{V})$; $\hat{G} = \hat{U}F'$, $\hat{H} = \hat{V}\bar{F}'$, and $\hat{E} = \hat{W}\tilde{F}'$ with F, $\bar{F}$, and $\tilde{F}$ as above.

Then $p(\hat{U},\hat{V},\hat{W}) = \sum_{i=1}^{s} \hat{\lambda}_i = \sum_{j=1}^{t} \hat{\mu}_j = \sum_{k=1}^{u} \hat{\nu}_k = \max_{(U,V,W)} p(U,V,W)$

As $p(\hat{G},\hat{H},\hat{E}) = p(\hat{U},\hat{V},\hat{W})$, $\max_{(U,V,W)} p(U,V,W) = \max_{(G,H,E)} p(G,H,E)$, thus

if $(\hat{U},\hat{V},\hat{W})$ maximizes p, then any orthonormal rotations of $\hat{U}$, $\hat{V}$, $\hat{W}$ do so as well.

exact solution

Theorem 2(B) (see section 4 of Part I)

Be g defined as:

$$g(G,H,E) = \mathrm{tr}\,(Z - GG'Z(HH' \otimes EE'))'(Z - GG'Z(HH' \otimes EE')) \quad \text{with } Z \in R^{1\times mn}.$$

Let $(\hat{G},\hat{H},\hat{E})$ be an exact solution of g, i.e. $g(\hat{G},\hat{H},\hat{E}) = 0$.

Then:

G is the eigenvector matrix (or an orthonormal rotation thereof) associated with the s non-zero eigenvalues of ZZ' with $Z \in R^{1\times mn}$;

H is the eigenvector matrix (or an orthonormal rotation thereof) associated with the t non-zero eigenvalues of ZZ' with $Z \in R^{m\times nl}$;

E is the eigenvector matrix (or an orthonormal rotation thereof) associated with the u non-zero eigenvalues of ZZ' with $Z \in R^{n\times lm}$.

Proof:

From $g(\hat{G},\hat{H},\hat{E}) = 0$, and the definition of g , it follows that

$$Z = GG'Z(HH' \otimes EE'). \tag{2.6}$$

Theorem 1b states that G, H, and E are the eigenvector matrices associated with the s, t, and u largest eigenvalues of P, Q, and R respectively.

If we define $\Lambda \in K^{s\times s}$ to be the eigenvalue matrix of P associated with G we have:

$$G\Lambda G' = P = Z(HH' \otimes EE')Z' \quad \text{with } Z \in R^{1\times mn} \tag{2.7}$$

Pre- and postmultiplying (2.7) with GG', and subsequently substituting this in (2.6) shows that

$$G\Lambda G' = ZZ' \qquad \text{with } Z \in R^{1\times mn}$$

In other words G is the eigenvector matrix of ZZ', and Λ the associated eigenvalue matrix. Furthermore the rank of G (= s) is equal to that of ZZ', and thus the λ_i (i=1,...,s) are the s non-zero eigenvalues of ZZ'.

The analoguous result holds for H and E.

3. The TUCKALS3 algorithm

Be Z a l×m×n three-mode matrix, and let s,t, and u be the desired number of components for the three component matrices. Furthermore let $g_1, g_2, \ldots, g_s$, $h_1, h_2, \ldots, h_t$, and $e_1, e_2, \ldots, e_u$ be systems of orthonormal iteration vectors which are combined into the orthonormal matrices $G \in R^{l\times s}$, $H \in R^{m\times t}$, and $E \in R^{n\times u}$ respectively.

If we write the matrices G, H, and E as they are after i iteration steps as G_i, H_i, and E_i, then one main iteration step of the TUCKALS3 algorithm can be described by (3.1) through (3.6):

G substep

$$P_i = Z(H_iH_i' \otimes E_iE_i')Z' \qquad (Z \in R^{l\times mn}) \qquad (3.1)$$

$$G_{i+1} = f_1(G_i) = P_iG_i(G_i'P_i^2G_i)^{-\frac{1}{2}} \qquad (3.2)$$

H substep

$$Q_i = Z(E_iE_i' \otimes G_{i+1}G_{i+1}')Z' \qquad (Z \in R^{m\times nl}) \qquad (3.3)$$

$$H_{i+1} = f_2(H_i) = Q_iH_i(H_i'Q_i^2H_i)^{-\frac{1}{2}} \qquad (3.4)$$

E substep

$$R_i = Z(G_{i+1}G_{i+1}' \otimes H_{i+1}H_{i+1}')Z' \qquad (Z \in R^{n\times lm}) \qquad (3.5)$$

$$E_{i+1} = f_3(E_i) = R_iE_i(E_i'R_i^2E_i)^{-\frac{1}{2}} \qquad (3.6)$$

Because we want to discuss the properties of the TUCKALS3 algorithm in the sequel it is useful to introduce some notation.

-- $F: S \to S$ is a function on S, F defines a complete step of the main iteration, and S is defined as in (2.5)

-- $F = F_3 \cdot F_2 \cdot F_1$ with $F_i: S \to S$ i=1,2,3 s.t.

$$F_1(G_i, H_i, E_i) = (f_1(G_i), H_i, E_i) = (G_{i+1}, H_i, E_i)$$

$$F_2(G_{i+1}, H_i, E_i) = (G_{i+1}, f_2(H_i), E_i) = (G_{i+1}, H_{i+1}, E_i)$$

$$F_3(G_{i+1}, H_{i+1}, E_i) = (G_{i+1}, H_{i+1}, f_3(E_i)) = (G_{i+1}, H_{i+1}, E_{i+1})$$

thus $F(s_i) = F(G_i, H_i, E_i) = (G_{i+1}, H_{i+1}, E_{i+1}) = s_{i+1}$

4. F is monotone

From the previous section it can easily be seen that if each F_i (or f_i) is monotone, then F is monotone.

Theorem 3

Let f be of the form:

$$f(X) = AX(X'A^2X)^{-\frac{1}{2}} \quad \text{with } X'A^2X \text{ positive definite,} \tag{4.1}$$

and let p be defined as:

$$p(X,Y) = \text{tr } XAY' \quad \text{with } X,Y \in K^{a\times b}, \text{ and } A \in R^{a\times a} \text{ and symmetric.}$$

If $Y = f(X)$, then (4.2)

$$p(Y,Y) \geq p(X,X), \tag{4.3}$$

with equality if and only if $Y = X$.

Proof:

As $X'A^2X$ is positive definite (see also section 6 of Part I), its inverse exists, and thus f is uniquely defined.

a. We first show that for every $X \in K^{n\times m}$

$$p(Y,X) = \max_{Z \in K^{n\times m}} p(Z,X) \tag{4.4}$$

To do this we incorporate the constraints on Z into the maximization

$$\tilde{p}(Z,X) = p(Z,X) - \text{tr } M(Z'Z - I_b) \tag{4.4}$$

with M a symmetric matrix of Lagrange multipliers.
Differentiating with respect to Z and M, and setting all the partial derivatives equal to zero, we obtain the following set of equations which have to be solved simultaneously:

$$AX = 2ZM \tag{4.5}$$

$$Z'Z = I_b \tag{4.6}$$

Say some $(\hat{Z},\hat{M})$ is the solution of the system. Then by premultiplying (4.5) with its transpose, reminding ourselves of the symmetry of A and $\hat{M}$, and substituting (4.6) into (4.5) we get:

$$X'A^2X = 4\hat{M}, \text{ and}$$

$$\hat{Z} = AX(X'A^2X)^{-\frac{1}{2}} = f(X) = Y \text{ according to (4.1).}$$

Thus for any $X \in K^{a\times b}$ Y maximizes p, or in other words:

$$p(Y,X) = \max_{Z \in K^{a\times b}} p(Z,X) \leq p(X,X) \quad \text{for all } X \in K^{a\times b}$$

b. Next we show that

$$p(Y,X) \leq p(Y,Y)^{\frac{1}{2}}\, p(X,X)^{\frac{1}{2}}.$$

As A is symmetric it may be decomposed into $A = B'B$, where B is a upper triangular matrix. Thus

$$p(Y,X) = \operatorname{tr} Y'AX = \operatorname{tr} (RY)'(RX)$$

The Cauchy-Schwarz inequality can now be applied:

$$\operatorname{tr} Y'AX = \operatorname{tr} (RY)'(RX) \leq \{\operatorname{tr} (RY)'(RY)\}^{\frac{1}{2}}\{\operatorname{tr} (RX)'(RX)\}^{\frac{1}{2}} =$$

$$= (\operatorname{tr} Y'AY)^{\frac{1}{2}}(\operatorname{tr} X'AX)^{\frac{1}{2}}$$

or

$$p(Y,X) \leq p(Y,Y)^{\frac{1}{2}}\, P(X,X)^{\frac{1}{2}}$$

c. Now we can prove inequality (4.3):

$$p(X,X) \leq p(Y,X) \leq p(Y,Y)^{\frac{1}{2}}\, p(X,X)^{\frac{1}{2}},$$

and as p is always non-negative:

$$p(X,X)^{\frac{1}{2}} \leq p(Y,Y)^{\frac{1}{2}} \quad \text{, and thus } \quad p(X,X) \leq p(Y,Y).$$

d. In the Cauchy-Schwarz inequality the equality sign holds if and only if X and Y are proportional, and inspection shows that the only possible proportionality constant is 1.

The extension of the monotonicity to F is straightforward. The equality condition can be seen to hold if one applies Theorem 3 succesively in each substep of the algorithm. Arriving finally at the conclusion that the equality sign holds if and only if $F(s) = s$.

5. F is continuous

From the definitions of F, F_i, and f_i it follows that if all f_i are continuous, all F_i and F will be as well.
It is thus sufficient to show that f as defined in (4.1) is continuous for all $X \in K^{n\times m}$, as all f_i are of the form (4.1).

Theorem 4

Be A given and symmetric, and be $X'A^2X$ positive definite (see also section 6 of Part I).
Let f be such that $f: K^{n\times m} \to K^{n\times m}$ with f defined as in (4.1) for all $X \in K^{n\times m}$, then f is continuous on $K^{n\times m}$.

Proof:

As $X'A^2X$ is positive definite, its inverse exists, and f is uniquely defined.
Let $\bar{X}$ be an arbitrary point in $K^{n\times m}$, and let $X_0, X_1, X_2, \ldots$ be a sequence in $K^{n\times m}$, which converges to $\bar{X}$, such that $X_i \neq \bar{X}$ $(i=0,1,2,..)$
Define $Y_l = f(X_l)$ $l=0,1,2,\ldots$
For each l part a of Theorem 3 shows that

$$\operatorname{tr} Y_l'AX_l \geq \operatorname{tr} Y'AX_l \text{ for all } Y \in K^{n\times m}.$$

Because the sequence $Y_0, Y_1, Y_2, \ldots$ is defined on a compact set, viz. $K^{n\times m}$, there exists at least one limit point, say Y, and $Y \in K^{n\times m}$.
In addition, there exists a subsequence $Y_{l_0}, Y_{l_1}, Y_{l_2}, \ldots$ which converges to Y. For each Y_{l_j} $(j=0,1,2,\ldots)$ of this subsequence it is true that:

$$\operatorname{tr} Y'A\bar{X} = \lim_{j\to\infty} \operatorname{tr} Y_{l_j}' AX_{l_j} \geq \lim_{j\to\infty} \operatorname{tr} Y'AX_{l_j} = \operatorname{tr} Y'A\bar{X} \qquad (5.1)$$

In Theorem 3 part a it was shown that if $\hat{Y} = f(\hat{X})$ for each $\hat{X} \in K^{n\times m}$,

$$\operatorname{tr} \hat{Y}'A\hat{X} = \max_{Y \in K^{n\times m}} \operatorname{tr} Y'A\bar{X}$$

Thus from (5.1) we get that:

$$Y = A\bar{X}(\bar{X}'A^2\bar{X})^{-\frac{1}{2}} = \bar{Y} \text{ , because of the definition of Y.}$$

Thus we have now that every convergent subsequence of $Y_0, Y_1, Y_2, \ldots$ has as its limit $\bar{Y}$, and therefore $\bar{Y}$ is the limit point of $Y_0, Y_1, Y_2, \ldots$ itself. Thus we may conclude that:

> for each $\bar{X} \in K^{n\times m}$, and each $X_0, X_1, X_2, \ldots$ converging to $\bar{X}$, the sequence $Y_0, Y_1, Y_2, \ldots$ converges to the limit point $\bar{Y}$, which means that $f(X_0), f(X_1), f(X_2), \ldots$ converges

to $f(\bar{X})$.

Recalling the definition of the continuity of a function, we see that f f is continuous for each $\bar{X} \in K^{n\times m}$, and therefore f is contiuous on $K^{n\times m}$.

6. Limit points of the algorithm are stationary points of p

Theorem 5

Let Z, G, H, E, P, Q, R, and p be defined as in the previous sections. If s = (G,H,E) is a limit point of the algorithm, then s is a stationary point of p.

Proof:

Let s = (G,H,E) be a limit point of the algorithm, then

$$G = PG(G'P^2G)^{-\frac{1}{2}}$$

$$H = QH(H'Q^2H)^{-\frac{1}{2}}$$

$$E = RE(E'R^2E)^{-\frac{1}{2}} \quad \text{(cf. Lemma 3 of Part I)}$$

First consider G.

Define $L = (G'P^2G)^{\frac{1}{2}}$, then (G,L) is a solution of :

$$\tilde{P}\tilde{G} = \tilde{G}L \qquad (6.1)$$
$$\tilde{G}'\tilde{G} = I_s$$

As L is symmetric, there exists a $F \in K^{s\times s}$, such that $L = F\Lambda F'$ with $\Lambda \in D^{s\times s}$. Substituting this in (6.1) we get

$$PG = GP\Lambda P'$$

which leads to

$$P(GF) = (GF)\Lambda \quad \text{or} \quad P\hat{G} = \hat{G}\Lambda \quad \text{with } \hat{G} = GF$$

Thus $\hat{G}$ is a matrix with eigenvectors of P and G is an orthonormal rotation of $\hat{G}$.

The analoguous result holds for H and E.

Theorem 1, part a, tells us that (G,H,E) is a stationary point of p.