TECHNOMETRICS ©, VOL. 21, NO. 4, NOVEMBER 1979

Lower Rank Approximation of Matrices by Least Squares With Any Choice of Weights

K. Ruben Gabriel

Department of Statistics
University of Rochester
Rochester, NY 14627

S. Zamir

Department of Statistics
Hebrew University
Jerusalem, Israel

Reduced rank approximation of matrices has hitherto been possible only by unweighted least squares. This paper presents iterative techniques for obtaining such approximations when weights are introduced. The techniques involve criss-cross regressions with careful initialization. Possible applications of the approximation are in modelling, biplotting, contingency table analysis, fitting of missing values, checking outliers, etc.

KEY WORDS

Reduced rank approximation
Least squares
Criss-cross regression
Householder-Young theorem
Biplot
Contingency table
Outliers

1. INTRODUCTION

Approximation of matrices by other matrices of lower rank plays a useful role in fitting models to data (Mandel [15], [16]; Bradu and Gabriel [1]), in graphical representation of data by means of biplots (Gabriel [4], [5]), in principal component analysis (Whittle [24]) and in other mutivariate techniques. (In fact the underlying approach of S. N. Roy and his students [19], [20], has been that of studying the rank one approximation of the data matrix). The method of approximation used in all these applications is least squares, with the solution due to Householder and Young [13] (and earlier on to Fisher and Mackenzie [3]) for which a variety of special computational routines are available (Golub and Reinsch [11]). An alternative method of approximation is an iterative procedure in which row and column weights are inversely dependent on row and column sums of squared residuals, and weighted least squares are used in each iterative step (McNeil and Tukey [17]). This is presumably more resistant to outliers.

The need for approximation by weighted least squares also arises frequently. For example, a table of means based on samples of widely varying sizes should be fitted with weights proportional to sample sizes. In the extreme case of zero size samples, an "entry" should play no role in fitting. This would also take care of missing values by assigning zero weights.

This paper considers iterative methods of fitting lower rank least squares approximations for a general choice of weights. For an $(n \times m)$ matrix Y of elements $y_{i,j}$ it considers least squares fitting subject to weights $w_{i,j}$. Fitting by a matrix of rank ρ or less is equivalent to fitting by a matrix product AB' where A and B are $n \times \rho$ and $m \times \rho$, respectively (Gabriel [6]). The minimizing criterion can therefore be written as

$$\Phi(A, B) = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m} w_{i,j}(y_{i,j} - \mathbf{a}_i\mathbf{b}_j)^2 \right\}, \qquad (1.1)$$

where $\mathbf{a}_i$ and $\mathbf{b}_j$ denote rows of A and B, respectively.

2. THE CASE OF EQUAL WEIGHTS

Householder and Young [13] dealt with equal weights, $w_{i,j} = 1$, and minimized $\Phi = ||Y - AB'||^2$, the Euclidean norm of the matrix of residuals. A convenient method of solution (see, e.g., Good [9], page 827) deals with the columns of A, and the corresponding columns of B, one at a time. The solutions for $\mathbf{a}_{,r}$ and $\mathbf{b}_{,r}$—the r-th columns of A and B, respectively—are obtained after solutions for $\mathbf{a}_{,1}, \cdots, \mathbf{a}_{,r-1}$ and $\mathbf{b}_{,1}, \cdots, \mathbf{b}_{,r-1}$ are available and subtracted out of Y to give residuals

$$Y^{(r-1)} = Y - \Sigma_{t=1}^{r-1} \mathbf{a}_{,t}\mathbf{b}_{,t}'. \qquad (2.1)$$

The equations determining $\mathbf{a}_{,r}$ and $\mathbf{b}_{,r}$ are

$$(\Sigma_i a_{i,r}^2)^{1/2} b_{j,r} = \Sigma_i a_{i,r} y_{i,j}^{(r-1)} \qquad (2.2)$$

Received April 1977; revised October 1978

and

$$(\Sigma_j b_{j,r}^2)^{1/2} a_{i,r} = \Sigma_j b_{j,r} y_{i,j}^{(r-1)}. \quad (2.3)$$

These are iterated, from some initial $\mathbf{a}_{\cdot r}$ until they converge. Equivalently, one could omit the square roots in both equations. The method then becomes one of criss-cross regression of columns of $Y^{(r-1)}$ onto $\mathbf{a}_{\cdot r}$ to obtain $\mathbf{b}_{\cdot r}$ as coefficients, and regression of rows of $Y^{(r-1)}$ onto $\mathbf{b}_{\cdot r}$ to obtain $\mathbf{a}_{\cdot r}$ as coefficients.

This unweighted least squares fit is seen to proceed by dyadic (i.e., rank one) steps, from fitting dyadic $\mathbf{a}_{\cdot 1}\mathbf{b}_{\cdot 1}'$ to Y, through fitting dyadic $\mathbf{a}_{\cdot 2}\mathbf{b}_{\cdot 2}'$ to $Y^{(1)} = Y - Y_{(1)}$ (where $Y_{(1)} = \mathbf{a}_{\cdot 1}\mathbf{b}_{\cdot 1}'$) and on to fitting $\mathbf{a}_{\cdot \rho}\mathbf{b}_{\cdot \rho}'$ to $Y^{(\rho-1)}$. At each step the sum of the dyadic residual fits gives the overall fit of that rank, that is,

$$Y_{(\rho)} = \sum_{r=1}^{\rho} \mathbf{a}_{\cdot r}\mathbf{b}_{\cdot r}'. \quad (2.4)$$

This stepwise fitting is possible because successive $\mathbf{a}_{\cdot r}$'s, and also successive $\mathbf{b}_{\cdot r}$'s, are orthogonal (or, in the case of multiple eigenvalues of $Y'Y$, they can be chosen so as to be orthogonal).

3. CRISS-CROSS REGRESSIONS AND SUCCESSIVE DYADIC FITS

The method of criss-cross regressions of columns and rows of $Y^{(r-1)}$ onto, respectively, $\mathbf{a}_{\cdot r}$ and $\mathbf{b}_{\cdot r}$ is readily generalized to arbitrarily weighted least squares. (For the special use in which the weights are all 0 or 1 see also A. Ruhe [21].) The iteration equations generalize to

$$(\Sigma_i w_{i,j} a_{i,r}^2) b_{j,r} = \Sigma_i w_{i,j} a_{i,r} y_{i,j}^{(r-1)} \quad (3.1)$$

and

$$(\Sigma_j w_{i,j} b_{j,r}^2) a_{i,r} = \Sigma_j w_{i,j} b_{j,r} y_{i,j}^{(r-1)}. \quad (3.2)$$

Successive solution of these equations for $r = 1, \cdots, \rho$ has been referred to as the NIPALS procedure (Wold [27]; Wold and Lyttkens [26]).

There are two mathematical differences between the case of equal weights and the case of general weights, and these are very crucial to the applicability of the above-mentioned generalization:

(i) For general weights, criss-cross regressions, which solve equations (3.1) and (3.2) iteratively, may converge to some local minimum which *is not* the desired closest fit. The fact that this may happen even in the simple case of 0 and 1 weights was apparently overlooked in the existing literature. Judging from our own experience, the phenomenon of convergence to a "wrong" minimum is not at all unlikely when some weights are 0 (i.e., missing observations). This is further discussed in Section 4, below.

(ii) For $\rho > 1$ and general weights, the strategy of stepwise dyadic fits to residuals *does not* usually lead to the closest overall fit of rank ρ—though it does so when the weights are all equal. This is due to the fact that the successive $\mathbf{a}_{\cdot r}$'s (as well as the $\mathbf{b}_{\cdot r}$'s) are not necessarily orthogonal except when the weights are equal.

To overcome difficulty (i) we propose in Section 4 an initialization of the criss-cross regression method which prevents the most frequent type (and so far the only type known to us) of "wrong convergence."

For difficulty (ii) we propose two alternative strategies: One is a sort of repeated NIPALS procedure which is based only on successive dyadic fits—Section 5. The other is based on criss-cross multiple regressions—Section 6.

4. INITIALIZATION FOR DYADIC FITS

As mentioned above, in the case of unequal weights the iterative solutions of (3.1) and (3.2) may not converge to the closest fit. In particular, when some weights were zero, these iterations occasionally converged well away from the least squares fit. If $w_{i,j} = 0$ for some given (i, j) and the dyadic fit iteration reached approximately

$$\mathbf{a}_{\cdot} \doteq \frac{1}{\alpha}(y_{1,j}, y_{2,j}, \cdots, y_{i-1,j}, \alpha\beta, y_{i+1,j}, \cdots, y_{n,j})' \quad (4.1)$$

and

$$\mathbf{b}_{\cdot} \doteq \frac{1}{\beta}(y_{i,1}, y_{i,2}, \cdots, y_{i,j-1}, \alpha\beta, y_{i,j+1}, \cdots, y_{i,m}) \quad (4.2)$$

for some constants α and β, then it converged to these vectors with infinitely increasing α and β. This provided perfect fit in all cells of row i and of column j, except cell (i, j) whose fitted value $\alpha\beta$ increased indefinitely (which did not affect the goodness of fit since $w_{i,j} = 0$). The fit outside these columns could be extremely poor, as each fitted value decreased indefinitely to zero. The sum of squared deviations therefore converged to

$$\Phi_{i,j}^* = \sum_e \sum_g w_{e,g} y_{e,g}^2 - \sum_e w_{e,j} y_{e,j}^2 - \sum_g w_{i,g} y_{i,g}^2. \quad (4.3)$$

In all the numerical experiments that we ran with various initial vectors $\mathbf{a}_{(0)}$, this was *the only type* of above-minimal convergence that we came across. In view of this, we tried to eliminate this undesirable phenomenon by a suitable choice of the initial vector $\mathbf{a}_{(0)\cdot}$.

Note that on the i-th row and j-th column the same sum of squared deviations, namely zero, is obtained for any values α and β in the above vectors $\mathbf{a}_{\cdot}$ and $\mathbf{b}_{\cdot}$. These two values may therefore be chosen so as to minimize the deviations outside the i-th row and j-th

column; Φ will then be reduced below $\Phi_{i,j}$*—except in the special case when it remains equal to $\Phi_{i,j}$* because all values in Y outside the i-th row and j-th column are zero.

An obvious way to choose α and β with this purpose is to regress the values $(y_{e,g};\ e \neq i,\ g \neq j)$ onto the products $a_e b_g;\ e \neq i,\ g \neq j)$. Thus, one may solve

$$\left(\sum_{e\neq i}\sum_{g\neq j} w_{e,g} y_{e,j}^2 y_{i,g}^2\right)\Big/\alpha\beta = \sum_{e\neq i}\sum_{g\neq i} w_{e,g}\, y_{e,j}\, y_{i,g}\, y_{e,g} \quad (4.4)$$

for α and β, where either of these can be given any arbitrary nonzero value. Putting $\alpha = 1$, the regression coefficient in (4.4) is $1/\beta$ and the initial column becomes

$$\mathbf{a}_{(0),} = (y_{1,j} \cdots, y_{i-1,j}, \beta, y_{i+1,j}, \cdots, y_{n,j})'. \quad (4.5)$$

Unless $1/\beta = 0$, the initial fit must be at least as good as, and in non-trivial cases strictly better than, that of (4.1) and (4.2) and the process must converge to a sum of squares below $\Phi_{i,j}$*.

Motivated by this observation, we start the initialization with the calculation of $\Phi_{i,j}$* for each (i, j) with $w_{i,j} = 0$ (and also for (i, j)'s with weights which are close to zero). The (i, j) with the highest

$$\Psi_{i,j} = \sum_{e,g} w_{e,g}\, y_{e,g}^2 - \Phi_{i,j}^* = \sum_{e} w_{e,j} y_{e,j}^2 + \sum_{g} w_{i,g} y_{i,g}^2 \quad (4.6)$$

is chosen and the appropriate vector $\mathbf{a}_{(0),}$ calculated from (4.4) and (4.5). Clearly, the iteration must converge to a lower sum of squares than that of any pair $\mathbf{a}$, and $\mathbf{b}$, of (4.1) and (4.2). All those above-minimum convergences are therefore improved upon.

If no $w_{i,j}$ is zero or very small, our initialization consists simply of choosing the column with the longest weighted norm

$$\theta_j = \sum_i w_{i,j} y_{i,j}^2 \quad (4.7)$$

and putting

$$\mathbf{a}_{(0),} = (y_{1,j}, \cdots, y_{i,j}, \cdots, y_{n,j})'. \quad (4.8)$$

This initialization has converged to the true minimum of Φ in all the examples we have tried. We conjecture that it does so for all except perhaps some highly pathological cases.

5. FITS OF RANKS HIGHER THAN ONE

It was noted above that with unequal weights, the NIPALS procedure, i.e., stepwise dyadic residual fitting, may fail to provide the best overall fit of rank ρ ($\rho > 1$). However, it is still possible to obtain that overall fit by means of a succession of dyadic fits. Thus one would repeat each step of the NIPALS on residuals from fits of all other columns until the entire matrices A and B converged. This procedure will be referred to as successive dyadic fits. Several alternative programs of such procedures are given in Section 6. They all have the advantage of using only dyadic fits and thus require only the repeated solution of equations (3.1) and (3.2). Relying entirely on dyadic fitting makes these procedures relatively simple and also gives us confidence that they avoid "wrong" convergences by using proper initialization.

6. CRISS-CROSS MULTIPLE REGRESSIONS

A more direct approach to higher rank fits deals with the entire factor matrices $A_{(n\times\rho)}$ and $B_{(m\times\rho)}$, rather than separately with each column. Thus, for a given matrix A one would obtain B as coefficients of the weighted multiple regressions of the columns of Y onto those of A. Similarly, for given B, one would obtain A as coefficients of the weighted multiple regression of the rows of Y onto the columns of B. The equations are

$$((\Sigma_i w_{i,j}\, a_{i,g} a_{i,g'}))\mathbf{b}_j = ((\Sigma_i w_{i,j} a_{i,g}\, y_{i,j})) \quad (6.1)$$

and

$$((\Sigma_j w_{i,j} b_{j,e} b_{j,e'}))\mathbf{a}_i = ((\Sigma_i w_{i,j} b_{j,e} y_{i,j})) \quad (6.2)$$

for the rows of B and of A, respectively.

Since neither A nor B are given, one starts with some initial guess $A_{(0)}$ and iterates from A to B then from B to A, etc.

The least squares properties of multiple regression can be used to show that this method of iteration must converge, though it does not prove that it converges to the minimum value of Φ. However, the convergence point of the iterations, say (A^*, B^*), satisfies

$$\min_A \Phi(A, B^*) = \min_B \Phi(A^*, B) = \Phi(A^*, B^*) \quad (6.3)$$

which is a necessary, though not sufficient, condition for (A^*, B^*)'s being the minimum point of Φ.

7. ROUTINES AND PROGRAMS

Programs for weighted least squares approximation were built up from a small number of routines which carry out the computations described in the preceding sections. The basic iterative routines are: $FR(Y;\ \mathbf{a}_{(0),}) \rightarrow (\mathbf{a};\ \mathbf{b})$, which uses initial $\mathbf{a}_{(0),}$ to fit Y dyadically by $\mathbf{ab}'$; FRINT(Y) $\rightarrow$ ($\mathbf{a}$; $\mathbf{b}$), which uses the initialization discussed in Section 4 to produce dyadic fit $\mathbf{ab}'$ to Y; GENERT(Y; $A_{(0)}$) $\rightarrow$ (A; B), which uses alternate row and column multiple regressions to fit Y by AB' from initial $A_{(0)}$. It will be noticed that FR(Y; $\mathbf{a}_{(0),}$) $\rightarrow$ ($\mathbf{a}$; $\mathbf{b}$) is a special case of GENERT(Y; $A_{(0),}$) $\rightarrow$ (A; B) in which $A_{(0)}$, A and B

DISPLAY 1. Building blocks for routines.

(1.) Read data Y and weights

FRINIT(Y) → ($\underline{a}_{,1}$; $\underline{b}_{,1}$)

(2.) FRINIT(Y − $\underline{a}_{,1}\underline{b}'_{,1}$) → ($\underline{a}_{,2}$; $\underline{b}_{,2}$)

(2d) (dyadic)
FR(Y − $\underline{a}_{,2}\underline{b}'_{,2}$; $\underline{a}_{,1}$) → ($\underline{a}_{,1}$; $\underline{b}_{,1}$)
FR(Y − $\underline{a}_{,1}\underline{b}'_{,1}$; $\underline{a}_{,2}$) → ($\underline{a}_{,2}$; $\underline{b}_{,2}$)
Repeat until Φ converges

(2r) (multiple regression)
GENERT(Y; A) → (A, B)
Repeat until Φ converges

At this stage
A = ($\underline{a}_{,1}$, $\underline{b}_{,2}$)
B = ($\underline{b}_{,1}$, $\underline{b}_{,2}$)

(3.) FRINIT(Y − AB') → ($\underline{a}_{,3}$; $\underline{b}_{,3}$)

(3d) (dyadic)
FR(Y − $\underline{a}_{,2}\underline{b}'_{,2}$ − $\underline{a}_{,3}\underline{b}'_{,3}$; $\underline{a}_{,1}$) → ($\underline{a}_{,1}$; $\underline{b}_{,1}$)
FR(Y − $\underline{a}_{,1}\underline{b}'_{,1}$ − $\underline{a}_{,3}\underline{b}'_{,3}$; $\underline{a}_{,2}$) → ($\underline{a}_{,2}$, $\underline{b}_{,2}$)
FR(Y − $\underline{a}_{,1}\underline{b}'_{,1}$ − $\underline{a}_{,2}\underline{b}'_{,2}$; $\underline{a}_{,3}$) → ($\underline{a}_{,3}$; $\underline{b}_{,3}$)
Repeat until Φ converges

(3r) (multiple regression)
GENERT(Y; A) → (A; B)
Repeat until Φ converges

At this stage
A = ($\underline{a}_{,1}$, $\underline{a}_{,2}$, $\underline{a}_{,3}$)
B = ($\underline{b}_{,1}$, $\underline{b}_{,2}$, $\underline{b}_{,3}$)

(3s) (skew dyadic)
FR(Y − $\underline{a}_{,1}\underline{b}'_{,1}$ − $\underline{a}_{,2}\underline{b}'_{,2}$; $\underline{a}_{,3}$) → ($\underline{a}_{,3}$; $\underline{b}_{,3}$)
{ FR(Y − $\underline{a}_{,2}\underline{b}'_{,2}$ − $\underline{a}_{,3}\underline{b}'_{,3}$; $\underline{a}_{,1}$) → ($\underline{a}_{,1}$, $\underline{b}_{,1}$)
FR(Y − $\underline{a}_{,1}\underline{b}'_{,1}$ − $\underline{a}_{,3}\underline{b}'_{,3}$; $\underline{a}_{,2}$) → ($\underline{a}_{,2}$; $\underline{b}_{,2}$)
Repeat until Φ converges }
Repeat until Φ converges

umn. However this simplifies programming so much that the special routine FR is used for that case. These basic iterative routines were used to construct program building blocks as shown in Display 1. Several alterative programs were then built up from these building blocks as shown in Display 2.

Programs I and II have to be run separately for the fit of each rank except 1. Thus, if one requires approximations of ranks 2 and 3, one has to run either of the programs twice, i.e, I_2 and I_3 or II_2 and II_3 of Display 2.

It was felt that programs might run more efficiently if they used the rank $(\rho - 1)$ fit in initializing for the rank ρ fit. In other words, the rank ρ iteration was to begin with

$$A_{(0)} = (\mathbf{a}_{(k),1}, \mathbf{a}_{(k),2}, \cdots, \mathbf{a}_{(k),\rho-1}, \mathbf{a}_{\rho}), \qquad (7.1)$$

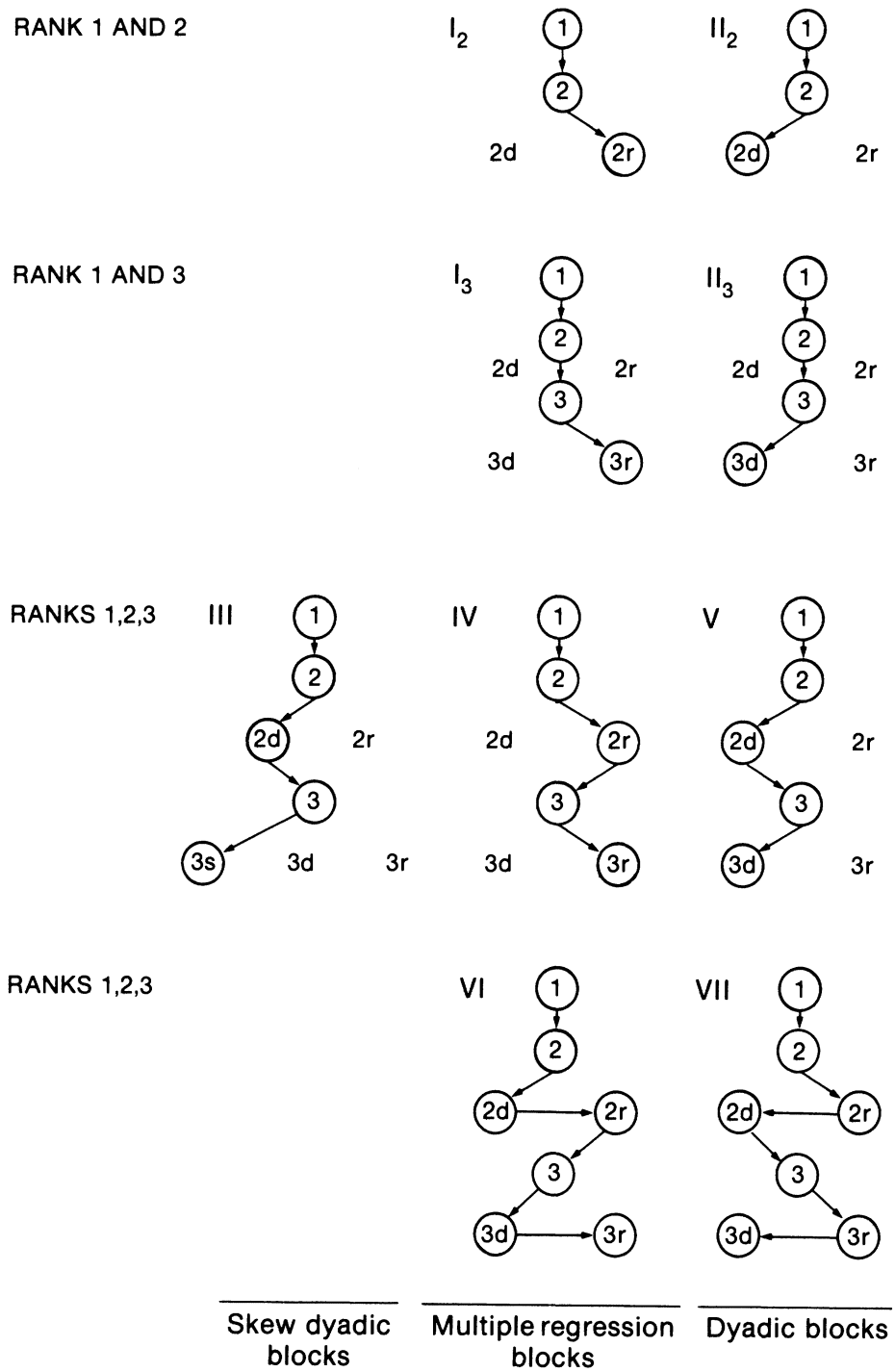
where k is the number of iterations at which the rank $(\rho - 1)$ fit was said to converge and $\mathbf{a}_{\rho}$ is the dyadic fit to

$$Y^{(\rho-1)} = Y - \sum_{t=1}^{\rho-1} \mathbf{a}_{(k),t}\mathbf{b}_{(k),t}'. \qquad (7.2)$$

A number of programs were constructed which incorporate this rank-by-rank fitting idea. For the rank ρ fit Program III separately iterates rank $(\rho - 1)$ fits and an additional dyadic fit. Program IV uses the criss-cross multiple regression routine for each rank whereas Program V uses successive dyadic fits. Programs VI and VII combine the latter two, VI first using successive dyadic fits and then criss-cross multiple regression, and VII combines them in the reverse order.

We tried out all these programs because we have no proof that any one of them invariably converges to the true minimum. A comparison of several alternative calculations might therefore have shown which program, if any, was uniformly more reliable. Alternatively, it was hoped that it would show that several of the methods always converged to the same

DISPLAY 2. Seven program flow charts using building blocks.



point—which would make it reasonably certain that all these methods did reach the true minimum.

As such agreement had been found we thought it well to opt for the method which used least computer time. Rank one, two and three fits were computed in each case. In all instances the seven programs converged to the same fit.

This is strong evidence that all the programs work and reach the true least squares approximation. Additional evidence to this effect was obtained by trying weights which factored into column and row components and finding that the present programs arrived at the true minimum as obtained by an extension of the Householder-Young method (Haber and Gabriel [12]). In particular, for equal weights, the programs always yielded the Householder-Young approximation.

When fits of both rank 2 and 3 were calculated, it was found cheapest (fastest) to run program IV, which uses criss-cross multiple regressions. Next fast-

have only one single colest were programs V and then III, which use successive dyadic fits. Double checking programs VI and VII was of course slower, but not as slow as making separate runs for each rank's fit i.e., I_2, *and* I_3 or II_2 *and* II_3.

The advantage in cost of program IV over V was not found to be large enough to warrant an unequivocal recommendation.

8. RELATION TO EXISTING RESULTS

Golub and Pereyra [10] have provided an algorithm for solving the following minimization problem:

$$\left.\begin{array}{l}\text{Find } \boldsymbol{\alpha} \text{ and } \mathbf{a} \text{ which minimize } r(\mathbf{a}, \boldsymbol{\alpha}) \text{ where:} \\ r(\mathbf{a}, \boldsymbol{\alpha}) = \|\mathbf{y} - \mu(\boldsymbol{\alpha})\mathbf{a}\|^2, \\ \mathbf{y} \in R^s, \mathbf{a} \in R^p, \boldsymbol{\alpha} \in R^q \text{ and } \mu(\boldsymbol{\alpha}) \text{ is an } s \times q \\ \text{matrix.}\end{array}\right\} \quad (8.1)$$

Ruhe and Wedin [22] suggested three variations for this algorithm. The third one is very similar to crisscross regression.

The problem we are concerned with in the present note is the following (See 1.1):

$$\left.\begin{array}{l}\text{Find an } (n \times \rho) \text{ matrix } A \text{ and an } (m \times \rho) \\ \text{matrix } B \text{ which jointly minimize:} \\ \Phi(A, B) = \|W^*(Y - AB')\|^2 \\ \text{where } Y \text{ and } W \text{ are } (n \times m) \text{ matrices and } * \\ \text{denotes the Hadamard (i.e., element by ele-} \\ \text{ment) product.}\end{array}\right\} \quad (8.2)$$

Ruhe [21] considered (8.2) for $\rho = 1$ and all elements of W being 0 or 1. For this case he wrote (8.2) as a special case of (8.1) and applied the methods of Ruhe and Wedin [22] for solving (8.1). (See also Wold [27], Wold and Lyttkens [26], and Christofferson [2].)

Although it is not entirely straightforward it can be shown that what Ruhe did can also be generalized to $\rho > 1$ and general W, i.e., (8.2) can *always* be written as a special case of (8.1). So, in principle, all existing methods for solving (8.1) are applicable to (8.2). However, in the present note we chose to handle (8.2) as it is and not as a special case of (8.1). The main reasons for this approach are the following:

(i) Writing (8.2) as a special case of (8.1) enormously expands the dimension of the problem. More precisely, instead of the original $(n \times m)$ matrices in (8.2), the equivalent (8.1) has a matrix $\mu(\alpha)$ of dimention $(nm \times n\rho)$. Most of the elements of this matrix are zeroes and its rank is far below maximal. Working with such a matrix is not efficient numerically.

(ii) All existing algorithms of (8.1) involve computing generalized inverses and derivatives of matrices. The algorithm proposed in this note for (8.2) is considerably "simpler" and involves solving mainly the trivial equations (3.1) and (3.2). (Compare to (3.4) of Ruhe [21] page 8 for the case $\rho = 1$.)

(i) and (ii) have direct implications for the amount of computation involved in the procedure. As an illustration Ruhe and Wedin [22] found that the number of products and divisions needed in each iteration of their algorithm III (this is the crisscross regression type algorithm on their page 27) is of the order of $s(p^2 + q^2) + z_\alpha - (p^3 + q^3)/3$ where z_α is the number of non-zero elements in $\mu'(\alpha)$. For solving (8.2) with $\rho = 1$ via (8.1) with this algorithm we have $s = mn$, $p = n$, $q = m$. Hence (omitting z_α) we would need the order of $N_1 = n^3(m - 1/3) + m^3(n - 1/3)$ multiplications per iteration. (This does not include the formation of $\mu(\alpha)$ which has to be done at every iteration.) On the other hand, each iteration of (3.1) and (3.2) requires $N_2 = n(4m + 1) + m(4n + 1)$ multiplications (or divisions). So for $n = m$ we have $N_1 \sim n^4$ compared to $N_2 \sim n^2$. For instance if $n = m = 8$ one has $N_1 \sim 7000$ while $N_2 \sim 530$.

(iii) By transforming (8.2) into the form of (8.1) one loses the intuitive meaning of the problem and of the iterative steps that we propose, namely those of simple (or multiple) regression.

(iv) The methods suggested in this note are closely related to the well-known methods of least squares approximations for equal weights. This enables us to point out, and respond to, the new phe-

TABLE 1a—*Ratios of seeded to unseeded precipitation (with S.E.) on various types of days (Gabriel and Baras [7a]).*

Temperature at 700 mb	Precipitable Water in Atmosphere (in mm) 0-11		12-13		14-	
≤ -8	2.265	±.307	0.973	±.136	1.031	±.154
$-7 \leq \leq -3$	1.522	±.254	1.146	±.126	1.327	±.111
$-4 \leq$	0.284	±.690	2.259	±.527	0.971	±.091

TABLE 1b—*Rank two approximation to data of Table 1a—fitted by least squares weighted inversely to variance (i.e. squared S.E.).*

A		AB'		
.6071	.1794	2.026	0.910	1.194
.6240	-.0636	1.715	1.209	1.227
.5020	-.9984	-0.024	2.019	0.986
	B'	2.8991	1.8258	1.9667
		1.4813	-1.1042	0.0017

Goodness of fit = 99.26%

nomena that may occur in the case of general weights, namely: (*a*) possibility of "wrong convergence" of the criss-cross regression, and (*b*) loss of the orthogonality which made the stepwise dyadic fits to residuals (NIPALS) work for equal weights.

9. SOME APPLICATIONS

9.1 As a small example, consider Table 1*a* which gives ratios of amounts of precipitation under cloud seeding to corresponding amounts in the absence of seeding. The data are from the first Israeli rainfall stimulation experiment and the method of calculating the ratios and their standard errors has been described by Gabriel and Feder [8]. The rank two approximation AB′ is given in Table 1*b* and has, not surprisingly, excellent fit. (Goodness of fit is 99.26% by ratio of weighted Euclidean norms criterion).

Since the rank two approximation is so close, factorization AB′ can be used to represent the matrix in the plane. This is done by plotting the row vectors of A and B in the *biplot* (Gabriel [4]), where the i-th row vector of A provides a marker for row i of the matrix and the j-th row of B a marker for column j of the matrix. It was shown by Bradu and Gabriel [1] that collinearity of row markers and/or column markers is indicative of certain models for the matrix.

The row vectors of A and B are biplotted in Figure 1. It is immediately evident that the row markers $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ are almost perfectly colinear and the column markers $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ are also close to another straight line which is not perpendicular to the first. In terms of the diagnostic rules applicable to biplots (Bradu and Gabriel [1]) this indicates a concurrent model (i.e., $y_{ij} = \mu + \alpha_i + \beta_j + \lambda\alpha_i\beta_j$ subject to $\Sigma_i\alpha_i = \Sigma_j\beta_j = 0$). An iterative weighted least squares algorithm (Kester [14]) was therefore applied to provide the fitted model of Table 1*c*.

It should be noted that direct inspection of Table 1*a* would not have made it easy to diagnose the concurrent model, mostly because of the different precisions of the various entries in the table. And yet it fits sufficiently well so that the weighted sum of squared deviations of the model which has 6 fitted parameters does not exceed the 5% value of chi-square with 9 − 6 = 3 d.f.

9.2 For an example that is not quite so small, consider the data of Table 2*a* on science doctorates awarded in the United States. In a table of frequencies f_{ij} ($i = 1, \cdots, n; j = 1, \cdots, m$), contrasts $\Sigma_{ij}c_{ij}\log_e f_{ij}$ ($\Sigma_{ij}c_{ij} = 0$) have asymptotic variances $\Sigma_{ij}c_{ij}^2/f_{ij}$ (Plackett [18]) so that logarithms of frequencies are appropriately weighted by frequencies. In this application these logarithms are adjusted for their mean, i.e., $\log_e f_{ij} - (\Sigma_{ij}\log_e f_{ij})/nm$. This adjustment is convenient for graphical representation (see Bradu and Gabriel [1], Section 5).

Table 2*b* gives matrices A and B for the rank two

TABLE 1c—*Concurrent model fit to data of Table 1a—fitted by weighted least squares (Kester [14]).*

		Column Effects β_1 = .4566	β_2 = -.1325	β_3 = -.3241
Row Effects	α_1 = -.0310	2.040	0.926	1.315
	α_2 = -.2870	1.517	1.105	1.249
	α_3 = .3180	0.334	1.736	1.014

μ = 1.6475 λ = 2.2859

Goodness of fit = 98.77% Weighted SOS = 6.67

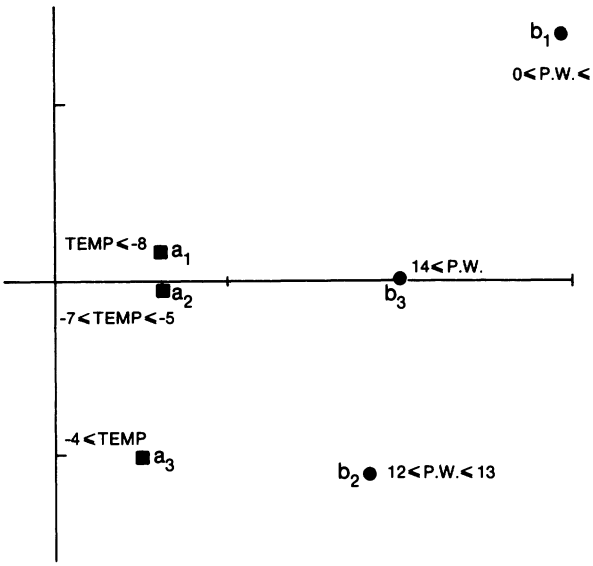


FIGURE 1. Biplot of ratios of seeded to unseeded precipitation.

approximation AB' fitted to these adjusted logarithms by least squares weighted with the frequencies f_{ij} of Table 2*a*.

The rank two fit is biplotted in Figure 2. A clear time trend is evident from 1960 to 1970, with a less striking movement in a different direction after 1970. The disciplines are scattered from Anthropology with fewest Ph.D.'s to Biology and Engineering with most. The scatter is elongated in a direction perpendicular to the 1960–70 trend: this indicates (Bradu and Gabriel [1]) that an additive model is appropriate for those years. Since we are dealing with logarithms, that implies a multiplicative model for the frequencies themselves, i.e., independence of discipline and time.

The biplot markers for some disciplines are not quite on the line orthogonal to the 1960–70 trend. Thus, on the one side, Agriculture, Earth Sciences and Chemistry had smaller 1960–70 increases; whereas, on the other side, Psychology and other Social Sciences (which includes Statistics) had larger increases.

As to the 1970–75 changes, these appear less pronounced and go in a different direction of the biplot. Considering the roughly SWS direction of that change, the biplot indicates a strong increase in Anthropology Ph.D.'s and a less strong increase in Sociology. In the other direction, it indicates decreases in Biology, Chemistry and Engineering. These biplot patterns mostly conform to the frequencies of Table 2*a*, except that the number of Biology Ph.D.'s did not diminish after 1970, but remained pretty much constant.

On the whole, the biplot is seen to allow rapid appraisal of the main features of the data but some deviations occur. This is to be expected, since the rank two approximation is not perfect.

9.3 Missing values in a two-way table (or a higher order table collapsed into matrix form) can be fitted by lower rank approximation of the entire matrix, entering the available values with weight one and arbitrary values with weights zero for the missing cells.

TABLE2a—*Science doctorates conferred in the U.S. in 1976.*

Ph.D.s	1960	1965	1970	1971	1972	1973	1974	1975
Total	6263	10477	17731	18880	18940	18948	18316	18352
Engineering	794	2073	3432	3495	3475	3338	3144	2959
Mathematics	291	685	1222	1236	1281	1222	1196	1149
Physics & Astro.	530	1046	1655	1740	1635	1590	1334	1293
Chemistry	1078	1444	2234	2204	2011	1849	1792	1762
Earth Sciences	253	375	511	550	580	577	570	556
Biological Sci.	1245	1963	3360	3633	3580	3636	3473	3498
Agric. Sci.	414	576	803	900	855	853	830	904
Psychology	772	954	1888	2116	2262	2444	2587	2749
Sociology	162	239	504	583	638	599	645	680
Economics	341	538	826	791	863	907	833	867
Anthropology	69	82	217	240	260	324	381	385
Other Soc. Sci.	314	502	1079	1392	1500	1609	1531	1550

Source: Statistical Abstract of the US, 1976 -- Table 958

TABLE 2b—*Reduced rank fit coefficients.*

A		B	
.4482	-.0166	.2453	4.0512
.0663	-.2389	1.4937	2.3897
.1780	-.1158	2.6466	.3549
.2741	.0106	2.7480	-.0524
-.2260	-.2924	2.5615	-.0631
.4847	.0147	2.5565	-.2542
-.0564	-.1974	2.5911	-.2651
.3320	-.1386	2.4121	-.3454
-.2012	-.4361		
-.0496	-.2361	Goodness of fit to matrix of	
-.4874	-.6620		
.1374	-.3273	$\log_e f_{ij}$-(mean of $\log_e f$'s)	
		is 98.95%	

For an illustration see Bradu and Gabriel [1], Section 10. Christofferson [2] had earlier used this method with rank one fits.

This method of interpolation assumes that the missing values fit the general dyadic, rank two or rank three pattern which approximates the available values. Since one is unlikely to have a priori reasons to assume a pattern of a certain rank one would presumably do well to try fits of several ranks and choose the least rank that gave a close fit.

The relation to the common method of interpolating missing values by an additive fit is simple. For rank one, the multiplicative fit is an alternative to the additive fit—a measure of goodness of fit should determine which method is more appropriate for a particular matrix. Additivity is a special case of the rank two model and, a fortiori, of the rank three model—hence the latter two will never fit worse than additivity. Whether calculation of their additional parameters is worthwhile will depend upon how much closer their fit is. Wishart [25] pointed out how some methods of fitting missing values introduce bias in clustering. It would seem that the higher the rank one uses in fitting, the less the bias—this is worth investigating.

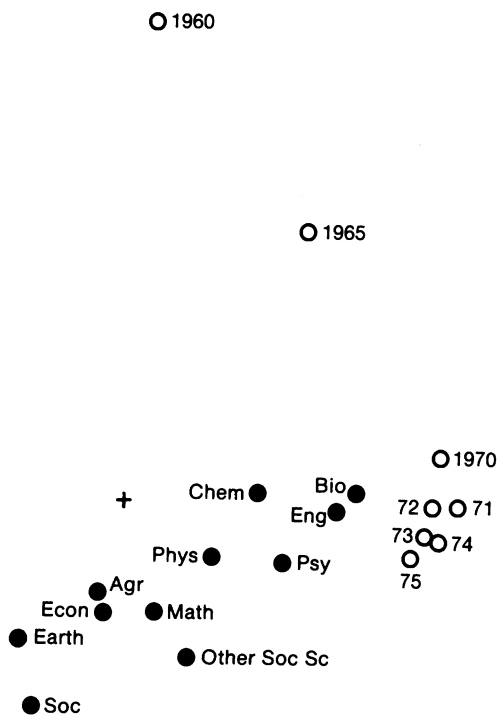


FIGURE 2. Biplot of science doctorates in the U.S.—logarithms of frequencies.

Another application (Shwertman and Allen [23]) is to the "smoothing" of covariance matrices whose elements are not all based on the entire sample because some observations were missing on some of the variables.

9.4 An obvious application of reduced rank approximation is to the MINRES method of factor analysis. This method aims at a reduced rank approximation of the off-diagonal elements of the correlation matrix and therefore fits into the present framework simply by setting all diagonal weights to zero and all off-diagonal weights to one. (Examples are discussed elsewhere (Gabriel [7]).)

9.5 A method for checking for outliers in a matrix would be to divide the elements into a number of subsets scattered over each row and column. Each element $y_{i,j}$ then belongs to a set S_k and may be compared with the reduced rank interpolation value $\hat{y}_{i,j}$ fitted by putting zero weights on itself and all other elements of S_k. Jackknifing or cross-validation techniques could provide tests of significance.

10. SUMMARY

The Householder-Young method is well known for approximation of matrices by matrices of lower rank. It provides the best approximation under unweighted least squares. However, there are many situations in which weighted least squares are more appropriate, be it because of differing precision of entries in the matrix or because of missing values (treated by assigning zero weights). No exact mathematical solution is available for weighted least squares and indeed many of the properties of the unweighted methods do not hold. Thus, one must not proceed by successive rank one fits to residuals from previous fits.

This paper presents some considerations in deriving an algorithm for weighted least squares and describes programs that will carry out such a fit iteratively. One of these programs is chosen for being faster than the others. A number of applications are shown: in modelling, biplotting, contingency table analysis, fitting of missing values and in checking outliers.

11. ACKNOWLEDGEMENT

The authors express their appreciation of Israel Einot's (Jerusalem) and Janet Gough's (Rochester) thoughtful and patient programming of successive

versions of these procedures. Much of the success of this work is due to their efforts. Computer programs in FORTRAN are available on request from the Division of Biostatistics, University of Rochester Medical Center, Rochester, NY 14642.

The referee's comments about relevant work in numerical analysis are greatly appreciated.

REFERENCES

[1] BRADU, D. and GABRIEL, K. R. (1978). The biplot as a diagnostic tool for models of two-way tables. *Technometrics, 20,* 47–68.

[2] CHRISTOFFERSON, A. (1969). The one-component model with incomplete data. Ph.D. Thesis, Uppsala University, Institute of Statistics.

[3] FISHER, R. A. and MACKENZIE, W. A. (1923). Studies in crop variation. *J. Agric. Sc., 13,* 311–320.

[4] GABRIEL, K. R. (1971). The biplot-graphic display of matrices with application to principal component analysis. *Biometrika, 58,* 453–467.

[5] GABRIEL, K. R. (1972). Analysis of meteorological data by means of canonical decomposition and biplots. *J. App. Meteor., 11,* 1071–1077.

[6] GABRIEL, K. R. (1978). Least squares approximation of matrices by additive and multiplicative models. *J. Roy. Statist. Soc., B, 40,* 186–196.

[7] GABRIEL, K. R. (1978). The complex correlational biplot. *Theory Construction and Data Analysis in the Behavioral Sciences* (S. Shye, ed.), pp. 350–370. San Francisco: Jossey-Bass.

[7a] GABRIEL, K. R. and BARAS, M. (1970). The Israeli rainmaking experiment. Jerusalem: Hebrew University (mimeographed).

[8] GABRIEL, K. R. and FEDER, P. (1969). On the distribution of statistics suitable for evaluating rainfall stimulation experiments. *Technometrics, 11,* 149–160.

[9] GOOD, I. J. (1969). Some applications of the singular value decomposition of a matrix. *Technometrics, 11,* 823–831.

[10] GOLUB, G. H. and PEREYRA, V. (1973). The differentiation of pseudo-inverses and non-linear least squares problems whose variables separate. *SIAM J. Numer. Anal, 10,* 413–432.

[11] GOLUB, G. H. and REINSCH, C. (1970). Singular value decomposition and least squares solution. *Numer. Math., 14,* 403–420.

[12] HABER, M. and GABRIEL, K. R. (1977). Weighted least onical correlation and biplot display (mimeographed). Department of Statistics, University of Rochester.

[13] HOUSEHOLDER, A. S. and YOUNG, G. (1938). Matrix approximation and latent roots. *Am. Math. Monthly, 45,* 165–171.

[14] KESTER, N. K. (1979). Diagnosing and fitting concurrent and related models for two-way and higher-way layouts. Unpublished Ph.D. thesis, University of Rochester, Rochester, NY.

[15] MANDEL, J. (1969). A method of fitting empirical surfaces to physical and chemical data. *Technometrics, 11,* 411–430.

[16] MANDEL, J. (1971). A new analysis of variance model for non-additive data. *Technometrics, 13,* 1–18.

[17] McNEIL, D. R. and TUKEY, J. W. (1975). Higher order diagnoses of two-way tables illustrated on two sets of demographic empirical distributions. *Biometrics, 31,* 487–510.

[18] PLACKETT, R. L. (1962). A note on interactions in contingency tables. *J. Roy. Statist. Soc., B, 24,* 162–166.

[19] ROY, S. N. (1957). *Some Aspects of Multivariate Analysis.* New York: Wiley.

[20] ROY, S. N., GNANADESIKAN, R. and SRIVASTAVA, J. N. (1971). *Analysis and Design of Certain Quantitative Multiresponse Experiments.* Oxford: Pergamon.

[21] RUHE, A. (1974). Numerical computation of principal components when several observations are missing. University of Umea, Institute of Mathematics and Statistics Report (mimeographed).

[22] RUHE, A. and WEDIN, P. A. (1974). Algorithms for separable non-linear least squares problems. University of Umea, Institute of Mathematics and Statistics Report (mimeographed).

[23] SCHWERTMAN, N. C. and ALLEN, D. M. (1973). The smoothing of an indefinite matrix with applications to growth curve analysis with missing observations. University of Kentucky, Department of Statistics, Technical Report No. 56 (mimeographed).

[24] WHITTLE, P. (1952). On principal components and least square methods of factor analysis. *Skand. Aktuar., 25,* 232–239.

[25] WISHART, D. (1978). Treatment of missing values in cluster analysis. *Compstat 1978—Proceedings in Computational Statistics* (L. C. A. Corsten and J. Hermans, eds.), pp. 281–287. Vienna: Physica-Verlag.

[26] WOLD, H. and LYTTKENS, E. (1969). Nonlinear iterative partial least squares (NIPALS) estimation procedures. *Bull. Inter. Statist. Inst., 43,* 29–51.

[27] WOLD, H. (1966). Nonlinear estimation by iterative least squares procedures. *Research Papers in Statistics* (F. N. David, ed.), pp. 411–444, New York: Wiley.