Data Analysis and Informatics
E. Diday et al. (eds.)

89

HOMALS & PRINCALS

SOME GENERALIZATIONS OF PRINCIPAL COMPONENTS ANALYSIS

Jan De Leeuw
Jan van Rijckevorsel *
Department of Datatheory
University of Leiden
Leiden
The Netherlands

This paper deals with a system of loss functions, measurement levels and process types for some forms of principal components analysis (= PCA). Generalizations are made for incomplete data-matrices and non-metric options. PCA for strictly nominal data (Guttman-Hayashi-Benzécri) and " non-metric " PCA (Young,Takane & De Leeuw or Tenenhaus or Kruskal & Shepard) are integrated in a single loss function and a corresponding algorithm. The ALS algorithm and its convergence properties are discussed.

Il s'agit d'un système des fonctions de perte, d'une typologie de données décrites à l'aide de variables caractéristiques nominales, ordinales ou numériques, et des transformations disjonctives ou continues pour l'analyse des ensembles des données incomplèts et mélangés en composants principales (= PCA). Deux sortes de PCA sont combinées dans une fonction de perte et dans l'algorithme correspondant. Les deux catégories de PCA sont PCA pour des données exclusivement nominales (Guttman-Hayashi-Benzécri) et PCA pour des données mélangées, comme les methodes de Young, Takane & De Leeuw, ou de Kruskal & Shepard, ou de Tenenhaus. Aussi sont décrits les qualités de convergence de l'algorithme.

* This study was supported in part by grant nr. 56 - 97 from the Netherlands organization for the advancement of pure research (Z.W.O.).

INTRODUCTION

Principal components analysis is one of the most popular data-reduction techniques in the social sciences. But the existing computerprograms for PCA impose rather severe restrictions both with respect to completeness of datamatrices and interval measurement of variables. If these restictions are violated, one is forced to take ad hoc measures. The assumption of interval scales in the social sciences is usually not justified, and often data matrices are incomplete. So it seems desirable to generalize the existing programs in such a way that more general types of variables with arbitrary patterns of missing data can be analyzed.

Some of these generalizations have been discussed earlier in the psychometric literature. In the first place there are PCA techniques for strictly nominal variables. Guttman gave a rather complete description in 1941, which has been extended later by Mosteller, Lord, Burt, Guttman, Hayashi, Lingoes, Nishisato and others. For more complete references see De Leeuw (1973). In the sixties Benzécri a.o. developed an equivalent form of PCA for nominal data in France which was called " analyse des correspondances ". This method is also worked out in many ways and a recent survey in french of the many applications, specializations and interpretations is given by Benzécri a.o. (1973). A useful review of this method in english is given by Hill (1974). The most complete and most general description of the results of the french school is Dauxois and Pousse (1976).

Besides the Guttman-Hayashi-Benzécri branch of PCA for nominal data, there exist several "nonmetric" generalizations of PCA. Probably the best known technique of this kind is Shepard and Kruskal (1974), but there exist also nonmetric PCA techniques in the GL-SSA series (Lingoes, 1972), in PRINCIPALS (Young, Takane and De Leeuw, 1978), in PRINQUAL (Tenenhaus, 1976), in POLYCON (Young, 1972) and in a program series made by Roskam (1968). The aim of our project is twofold: in the first place we have developed a general theoretic framework in which the two different forms of generalized PCA can be described and elaborated. Secondly great attention is given to the development of computerprograms, whose first and foremost aim is the solution of very big problems in a reasonable small amount of computing time. This makes high demands upon both algorithms and computerprograms. The algorithms are constructed according the alternating least squares method and they are written following the principles of structured programming. Optimization of subprograms is based upon their timing profiles during the analysis of large datasets.

INTERVAL DATA

One can define PCA in several ways. Here we prefer geometrical starting points and loss functions, derived from the geometry of the problem; that is, we prefer to see PCA as a multidimensional scaling method. Suppose we have n observations y_{ij} on m numerical variables (i=1,...,n;j=1,...,m). We want to represent these observations as points in a p-dimensional space and the variables as directions in that space i.e. as lines through the origin. Observation i is represented as the point $x_i = (x_{i1},x_{i2},\ldots,x_{ip})$, and variable j is represented as the direction cosines $a_j = (a_{j1},a_{j2},\ldots,a_{jp})$. Thus $a_j^T a_j = 1$ for all p. We require that the orthogonal projections of the observation points $x_1,x_2,\ldots,x_n$ on the direction defined by variable j are proportional with the measurement y_{ij}.

A perfect representation is thus defined by the requirements

$$y_{ij} = \beta_j \Sigma_{s=1}^{p} x_{is} a_{js} \quad . \tag{1}$$

Clearly a perfect representation exists if and only if we can find a matrix X (n x p) and a matrix A (m x p) so that $Y = XA^T$ i.e. if and only if rank (Y) $\leq$ p. If a perfect solution does not exist for a chosen p we use the loss function:

$$\sigma = \mathrm{tr}\,(Y - XA^T)^T(Y - XA^T) \tag{2}$$

This loss function is to be minimized over X and A. The minimum can be found by means of a singular value decomposition of Y or by means of an eigenvalue-eigenvector decomposition of Y^TY or YY^T. These computational methods have two drawbacks. In the first place the computing time increases by the third power of the number of variables, which entails that really large datasets cannot be analyzed. Secondly these methods are not easily generalized to more general forms of PCA.

MISSING DATA

One of these more general forms of PCA is the case in which observations on several variables are missing. A suitable generalization of formula (2) is obtained by using diagonal matrices M_j, indicating which observations on variable j are missing. Diagonal element i of M_j is unity if y_{ij} is *not* missing, and equals zero if y_{ij} is missing. Hence we get:

$$\sigma_1 = \mathrm{tr}\ \Sigma_{j=1}^{m} (y_j - Xa_j)^T M_j (y_j - Xa_j) \tag{3}$$

This loss function has some serious disadvantages, of which the most important one is that minimization of (3) is not an eigenproblem anymore.

We consequently use an alternative:

$$\sigma_2 = \Sigma_{j=1}^{m} \text{tr } (X - y_j a_j^T)^T M_j (X - y_j a_j^T) \qquad (4)$$

One has to prove that formula (4) is a generalization of formula (2). Suppose there are no missing data, i.e. $M_j = I$ for all j. Under the normalization requirements $X^TX = I$ and $a_j^T a_j = 1$ for all j we find:

$$\sigma_2 = \sigma_1 + m(p - 1) \qquad (5)$$

In this special case the minimization of σ_2 is equivalent to the minimization of σ_1. When $M_j \neq I$ for some j we have two different problems.

OPTIMAL SCALING

Until now we have assumed that all variables had at least an interval measurement level i.e. the elements y_{ij} are constants within the iterative process. We will generalize the idea by making the weaker assumption that the vectors $M_j y_j$ have to be in known convex cones K_j. The loss function in formula (4) has to be minimized over X, over A and over Y. The minimization of Y for temporary fixed values of X and A during an iteration is sometimes called the " Optimal scaling " or the optimal quantification of variables (Young, De Leeuw and Takane, 1979). This so called " optimality " is defined in terms of a particular loss function. Every minimizing sub-operation within an iteration corresponds with a type of partitioning of the residual sum of squares. The particular optimal scaling partitioning is:

$$\sigma = \Sigma_{j=1}^{m} \text{tr } (a_j^T a_j)(y_j - \hat{y}_j)^T M_j (y_j - \hat{y}_j) + \Sigma_{j=1}^{m} \text{tr } M_j X \{I - \frac{a_j a_j^T}{a_j^T a_j}\} X^T M_j$$

where $\hat{y}_j = \frac{1}{a_j^T a_j} M_j X a_j$ . . .

After some substitution we find:

$$\sigma = \Sigma_{j=1}^{m} \text{tr } (a_j^T a_j)(y_j^T M_j y_j) - \Sigma_{j=1}^{m} \text{tr } 2 y_j^T M_j X a_j + \Sigma_{j=1}^{m} \text{tr } M_j X X^T M_j ,$$

which is evidently equal to formula (4) :

$$\sigma = \Sigma_{j=1}^{m} \text{tr } X^T M_* X - \Sigma_{j=1}^{m} \text{tr } 2 y_j^T M_j X a_j + \Sigma_{j=1}^{m} \text{tr } a_j y_j^T M_j y_j a_j^T . \qquad (8)$$

We minimize the first term of (6) over y_j under the restriction that y_j is in the cone K_j. This defines a cone regression problem, the general theory of which is dealt with by De Leeuw (1977).

Metric PCA is the special case where the cones K_j are one-dimensional subspaces i.e. lines through the origin. In the case of ordinal variables the cones K_j are

sets of isotonic or monotonic vectors i.e. those vectors on which the observations are ordered in the same way as on the original raw data vector. Because of the peculiar character of nominal variables we will deal with those separately.

PROCESS TYPES

A description of the several types of scales which can be used within the ALS framework is given in De Leeuw, Young and Takane (1976). We suggest here a different classification based upon the same ideas. All variables are considered to be categorical and they can be interpreted as discrete or continuous. The categories of discrete categorical variables are represented as points on a scale and the categories of continuous categorical variables are represented as nonoverlapping intervals. The scale values of the individual observations have to fall within the quantifications of the corresponding categories in both cases. This means that for discrete variables the quantifications of observations and the quantification of the category to which they belong have to coincide.

We use indicator matrices to show the relation between quantifications of observations and categories (De Leeuw, 1973). An indicator matrix G for n observations on a k-category variable is a n x k binary matrix with $g_{ic} = 1$, if observation i scores in category c, and $g_{ic} = 0$, if i scores in another category. We distinguish discrete and continuous variables of interval, ordinal and nominal level. Continuous interval is a new type which is not mentioned in De Leeuw, Young, and Takane (1976), but which has recently been discussed in De Leeuw and Walter (1979). Continuous and discrete nominal variables have been discussed in De Leeuw, Young and Takane (1976). We do not discuss them in detail because in most cases it does not make much sense to map purely nominal data on a linear scale. Therefore we will use the alternative definition of continuous and discrete process types in respect to ordinal and interval data.

Assume an ordinal variable with k categories. In the continuous case we quantify a category c as an interval (z_c^-, z_c^+) with the restriction:

$$z_1^- \leq z_1^+ \leq z_2^- \leq z_2^+ \leq \ldots \leq z_k^- \leq z_k^+ \qquad (9)$$

and for all i :

$$\Sigma_{c=1}^{k} g_{ic} z_c^- \leq y_i \leq \Sigma_{c=1}^{k} g_{ic} z_c^+ . \qquad (10)$$

In the ordinal discrete case we have the extra restriction:

$$z_c^- = z_c^+ \ (= z_c) \text{ for all } c \qquad (11)$$

and thus:

$$y_i = \Sigma_{c=1}^{k} g_{ic} z_c \tag{12}$$

In the case of continuous interval variables z_c^- and z_c^+ are known numbers with $z_c^- < z_c^+$. There have to exist an $\alpha \geqslant 0$ and a β such that

$$\alpha \Sigma_{c=1}^{k} g_{ic} z_c^- + \beta \leqslant y_i \leqslant \alpha \Sigma_{c=1}^{k} g_{ic} z_c^+ + \beta . \tag{13}$$

In the discrete case again we have the restriction:

$$z_c^- = z_c^+ \quad (= z_c) \quad \text{for all } c, \tag{14}$$

thus:

$$y_i = \alpha \Sigma_{c=1}^{k} g_{ic} z_c + \beta \tag{15}$$

One can often expect the upperbound of a category to be equal to the lowerbound of the next category in case of continuous interval data.

One can see from formula (6) that the condition that Xa_j is in the cone K_j, is necessary for a perfect solution. This means that for ordinal discrete data the y_i have to be on k parallel hyperplanes (dimension p-1) orthogonal on the direction a_j. The projections on this direction must be in corresponding order. For ordinal continuous variables there have to be k-1 hyperplanes which separate the k clusters of points belonging to the observation scores in k categories. The y_i in category c are between the hyperplanes of category c-1 and category c+1. With discrete interval data y_i must be on corresponding hyperplanes but also these hyperplanes must be at certain distances to each other. Those distances are on an interval scale i.e. known up to a multiplicative constant, and for continuous interval data the y_i must be between an upper- and a lower bound with the necessary numerical properties. If the upper hyperplane of a category coincides with the lower hyperplane of the next category, there exist k-1 hyperplanes that separate the clusters of points belonging to the categories. See fig. 1 .

NOMINAL VARIABLES

In our earlier paragraphs we have discussed a generalization of non-metric PCA with the according loss functions, a system of measurement levels, algorithms and geometrical representations. In the sequel we will deal with the question how nominal data do fit in this framework and what is the relationship between PRINCALS and the Guttman-Hayashi-Benzécri approach (= HOMALS) to categorical data.

Nominal discrete data : Suppose a variable has k categories. We define k binary diagonal matrices M_c with the diagonal element M_{ci} equal to unity if observation i is in category c and equal to zero if i is not in category c. If there are no missing data the sum of all M_c is identity. The PRINCALS loss function for this variable is:

$$\sigma = \Sigma_{c=1}^{k} \text{ tr } (X - y_c a_c^T)^T M_c (X - y_c a_c^T) \qquad (16)$$

The optimal scaling restrictions for y_c are that all observations in category c will get the same score. So we treat our nominal variable with k categories as k binary variables with missing data. This is of course quite different from dividing a k-category variable into k complete binary variables. The relation with classical categorical PCA is rather straightforward (if no missing data exist). We can rewrite formula (4), using the indicator matrix G and the diagonal matrix of category frequencies $D = G^T G$, as

$$\sigma = \text{tr } X^T X - 2 \text{ tr } X^T G D^{-\frac{1}{2}} A + \text{tr } A^T A \qquad (17)$$

Define $E = D^{-\frac{1}{2}} A$. This implies

$$\sigma = \text{tr } (X - GE)^T (X - GE). \qquad (18)$$

Now we can say that in case of no missing data and only nominal discrete variables PRINCALS is equivalent to an eigen decomposition of the supermatrix C with submatrices

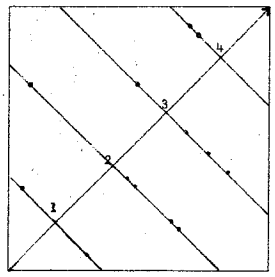
$$C_{j\ell} = D_j^{-\frac{1}{2}} G_j^T G_\ell D_\ell^{-\frac{1}{2}}, \qquad (19)$$

which is the Guttman-Hayashi-Benzécri PCA for categorical data, a generalization of which is the computerprogram HOMALS of Van Rijckevorsel and De Leeuw (1978) and De Leeuw (1976). An extra advantage, which is not accidental, is that we can apply the geometrical properties of this form of PCA to our PRINCALS and HOMALS solutions. Formula (18) shows that perfect fit is defined as the coincidence of all observation points with their corresponding category points, and the minimization of (25) over all variables equals the minimization of the within-category variance of the representation, for further details see De Leeuw (1973, 1976).

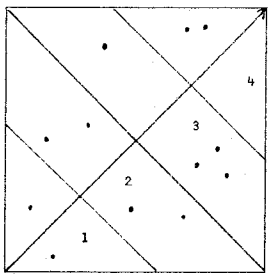
Nominal continuous variables are analysed by dividing a k category nominal variable in $\binom{k}{2}$ binary variables. Every pair of categories (c,c') defines a binary diagonal matrix $M_{cc'}$, with the element i equal to zero if i nor in c nor in c'. The loss function is

$$\sigma = \Sigma_{c=1}^{k} \Sigma_{c'=1}^{k} \text{tr}(X - y_{cc'} a_{cc'}^T)^T M_{cc'} (X - y_{cc'} a_{cc'}^T) \qquad (20)$$

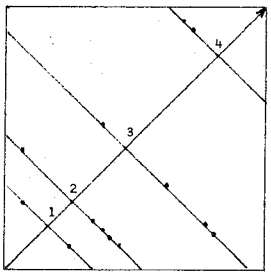
We impose the ordinal continuous restrictions on $y_{cc'}$, i.e. y_i corresponding with



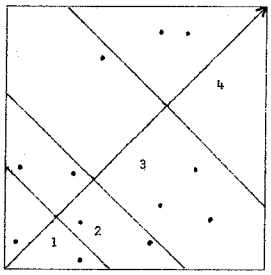
DISCRETE NUMERICAL



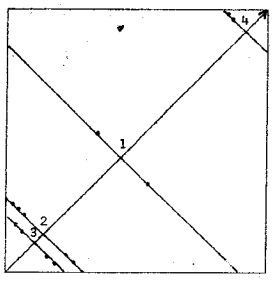
CONTINUOUS NUMERICAL



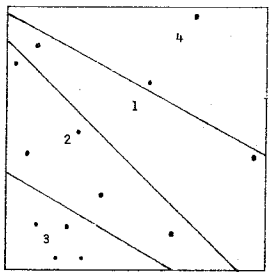
DISCRETE ORDINAL



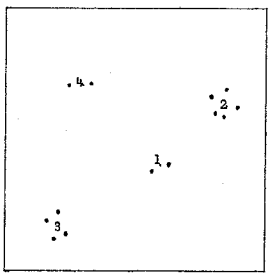
CONTINUOUS ORDINAL



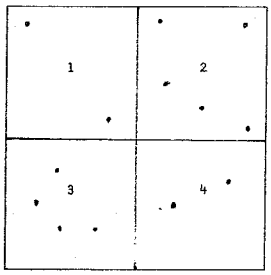
DISCRETE NOMINAL



CONTINUOUS NOMINAL (1)



DISCRETE NOMINAL CENTROID



CONTINUOUS NOMINAL (2)

FIGURE 1

c is not allowed to be smaller than y_i corresponding with c'. Geometrically we make the demand that all pairs of categories can be separated by hyperplanes. In other words we want the convex extensions of k clusters of points not to overlap. This is of course quite different from dividing a k category variable into k binary continuous variables, because this only requires that every category can be separated from the remaining k-1 categories together. See fig. 1.

CONVERGENCE

The loss function in its basic form, i.e. $M_j = I$, for all j, is

$$\sigma = \Sigma_{j=1}^{m} \text{tr } (X - y_j a_j^T)^T (X - y_j a_j^T) \qquad (21)$$

The loss σ must be minimized over three sets of parameters y,a,X under the restrictions $X^TX = I$, $a^Ta = 1$, and $y \in K$. We omit the variable subscripts because this is more convenient and it does not influence the results in this paragraph. The idea is that each substep of the algorithm minimizes the loss σ over one set for fixed values of the other two sets of parameters. Each iteration cycle consists of three substeps. In the beginning of an iteration we start with y^0, a^0, X^0. Each iteration gives us three updates y^+, a^+, X^+ each of which conditionally minimizes the loss σ.

Subproblems

To find the minimizing update y^+ with a^0 and X^0 is a cone regression problem, which has a unique solution (De Leeuw, 1977).

Finding an a^+ with y^+ and X^0 is an ordinary least squares problem with a unique solution. The component of σ depending on a is

$$\sigma(a) = \text{tr } (X - ya^T)^T (X - ya^T) \qquad (22)$$

$$= \text{tr } X^TX + \text{tr } ay^Tya^T - 2\text{tr } X^Tya^T \qquad (23)$$

$$= p + a^Ta - 2\, a^TX^Ty \qquad (24)$$

We have to minimize this over a satisfying $a^Ta = 1$. The solution is

$$a^+ = X^{0T}y^+ \,/\, (y^{+T}X^0X^{0T}y^+)^{\frac{1}{2}} \quad . \qquad (25)$$

Finding a X^+ which minimizes σ with a^+ and y^+ under the requirement $X^{+T}X^+ = I$ is an orthogonal prucrustus problem, which has an unique solution if and only if

the p-th singular value of the matrix $\frac{1}{m}\sum_{j=1}^{m} y_j^{\dagger} a_j^{+\tau}$ is greater than the (p+1)-th singular value (Cliff, 1966).

Properties of the algorithm

The loss function decreases except at a stationary point

Because all the suboperations are continuous operations the transformations

$$\begin{bmatrix} y^k \\ a^k \\ X^k \end{bmatrix} \rightarrow \begin{bmatrix} y^{k+1} \\ a^{k+1} \\ X^{k+1} \end{bmatrix}$$

are continuous.

Because $a^{\tau}a = 1$, $X^{\tau}X = I$ and $y^{\tau}y = a^{\tau}X^{\tau}Xa \leqslant 1$ all our updates are in a compact set.

Convergence

Define u^k as the parameterset of the k-th iteration

Φ as the operation of the three step algorithm

σ as the loss, bounded from below

Suppose the algorithm generates an infinite sequence u^k, none of the u^k is a stationary point, then

$$\sigma(u^k) > \sigma(u^{k+1}) > 0 \tag{26}$$

$$\Phi(u^k) = u^{k+1} \tag{27}$$

$$u^k \text{ belongs to the compact set } u \tag{28}$$

From (26) it follows that $\sigma(u^k)$ converges to σ^*

Bolzano-Weierstrass shows that u^k has an accumulation point u^*, i.e. there is a subsequence (u^l) such that $u^l \rightarrow u^*$. Define (u^{l+1}). This sequence again has a subsequence (u^{v+1}) converging to, say, u^{**}. And finally we construct (u^v) which is a subsequence of (u^l) and which converges consequently to u^*. For example

if (u^l) is $u^{(1)}, u^{(3)}, u^{(5)}, u^{(7)}, u^{(9)}, \ldots$

then (u^{l+1}) is $u^{(2)}, u^{(4)}, u^{(6)}, u^{(8)}, u^{(10)}, \ldots$

if (u^{v+1}) is $u^{(2)}, u^{(6)}, u^{(10)}, \ldots$

then (u^v) is $u^{(1)}, u^{(5)}, u^{(9)}, \ldots$

Because Φ is continuous and $u^{v+1} = \Phi(u^v)$ it follows that $u^{**} = \Phi(u^*)$

Because σ is continuous and $\sigma(u^k) \to \sigma^*$ it follows that $\sigma(u^{**}) = \sigma(u^*)$
But, if u^* is not a stationary point then $\sigma(u^{**}) > \sigma(u^*)$, because of $u^{**} = \Phi(u^*)$.
Thus u^* is stationary. We have proved that all accumulation points of u^k are stationary points with the same function value σ^*.

COMPUTER PROGRAMS

The several forms of PCA discussed in this paper are incorporated in two computer programs: HOMALS and PRINCALS. HOMALS entails the Guttman-Hayashi-Benzécri branch of PCA for strictly nominal data. PRINCALS combines the " non-metric " generalizations of the vector-model with the HOMALS approach but the application is restricted to discrete or discretized data. More information about algorithms, size, details of the programs and several examples and applications are to be found in Van Rijckevorsel & De Leeuw (1978) for HOMALS and in Van Rijckevorsel & De Leeuw (1979) for PRINCALS. Both programs are written according the ANSI Fortran conventions. They are available from:

Department of Datatheory
Faculty of the Social Sciences
University of Leiden
Leiden
The Netherlands

REFERENCES

Benzécri, J.P., 1974. Analyse des Données, (Volume II). Paris, Dunod.

Cliff, N., 1966. Orthogonal rotation to congruence, PM, 31, 33 - 42

Dauxois, J. & Pousse, A., 1976. Les analyses factorielles en calcul des probabilités et en statistique: essai d'étude synthetique. Thèse, Université Paul Sabatier, Toulouse.

De Leeuw, J., 1973. Canonical analysis of categorical data. Psychologisch Instituut, Leiden.

De Leeuw, J., 1976. HOMALS, Paper presented at the Psychometric society meeting, april 1976, Murray Hill, New York.

De Leeuw, J., 1977. A normalized cone regression approach to alternating least squares algorithms. an unpublished note, Dep. of Datatheory, Leiden.

De Leeuw, J., 1979. Optimal scaling of continuous numerical data. Submitted.

De Leeuw, J., Young, F.W. & Takane, Y., 1976. Additive structure in qualitative data: An alternating least squares approach with optimal scaling features. PM, 41, 471 - 503.

Hill, M.O., 1974. Correspondence analysis: A neglected multivariate method. Appl. Statist., 23, 3, 340 - 354.

Lingoes, J.C., 1972. A general survey of the Guttman - Lingoes nonmetric program series. In: R.N.Shepard, A.K.Romney, & S.Nerlove (Eds.), Multidimensional scaling: Theory and application in the behavioral sciences, (Volume I, Theory). New York: Seminar Press. Pp 49 - 68.

Roskam, E., 1968. Metric analysis of ordinal data in psychology. Voorschoten: VAM.

Shepard, R.N., & Kruskal, J.B., 1974. A nonmetric variety of linear factor analysis. PM, 39, 123 - 157.

Tenenhaus, M., 1976. Analyse en composantes principales d'un ensemble de variables nominales et numériques. Cahiers de recherche du CESA, Jouy-en-Josas.

Van Rijckevorsel, J., & De Leeuw, J., 1978. An outline of HOMALS-1. Dept. of Datatheory, Leiden.

Van Rijckevorsel, J., & De Leeuw, J., 1979. An outline of PRINCALS. Dept. of Datatheory, Leiden.

Young, F.W., 1972. A model for polynomial conjoint analysis algorithms. In: R.N. Shepard, A.K.Romney, & S.Nerlove (Eds.), Multidimensional scaling: Theory and application in the behavioral sciences,(Volume I, Theory). New York: Seminar Press. Pp 69 - 102.

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

Young, F.W., Takane, Y., & De Leeuw, J., 1978. The principal components of mixed measurement level data: An alternating least squares method. PM, 43, 279-282