PSYCHOMETRIKA—VOL. 53, NO. 2, 177–197
JUNE 1988

HOMOGENEITY ANALYSIS WITH k SETS OF VARIABLES: AN ALTERNATING LEAST SQUARES METHOD WITH OPTIMAL SCALING FEATURES

EEKE VAN DER BURG

DEPARTMENT OF EDUCATION
UNIVERSITY OF TWENTE

JAN DE LEEUW
RENÉE VERDEGAAL

DEPARTMENT OF DATA THEORY
LEIDEN UNIVERSITY

Homogeneity analysis, or multiple correspondence analysis, is usually applied to k separate variables. In this paper we apply it to sets of variables by using sums within sets. The resulting technique is called OVERALS. It uses the notion of optimal scaling, with transformations that can be multiple or single. The single transformations consist of three types: nominal, ordinal, and numerical. The corresponding OVERALS computer program minimizes a least squares loss function by using an alternating least squares algorithm. Many existing linear and nonlinear multivariate analysis techniques are shown to be special cases of OVERALS. An application to data from an epidemiological survey is presented.

Key words: homogeneity analysis, correspondence analysis, optimal scaling, transformation, alternating least squares, canonical correlation analysis, principal component analysis.

Introduction

Approximately ten years ago Young, de Leeuw, and Takane started to apply the optimal scaling ideas, that had originated in multidimensional scaling, to multivariate analysis. This made it possible to link the developments in multidimensional scaling with older but related developments in multivariate analysis centering around the notion of coding categorical variables by using matrices with zeroes and ones. The resulting ALSOS (alternating least squares with optimal scaling) approach to multivariate data analysis was based on the idea of alternating the transformation or quantification of variables with the fitting of model parameters in an iterative way, using least squares loss functions. This resulted in a series of programs for *nonlinear multivariate analysis*, with special programs for additivity analysis, multiple regression, canonical correlation analysis, principal component analysis, and factor analysis. A review of the general ALSOS approach and of the results that have been obtained, is given by Young (1981).

This research was partly supported by SWOV (Institute for Road Safety Research) in Leidschendam, The Netherlands.

The authors wish to thank Wilfrid van Pelt, Department of Physiology and Physiological Physics, University Medical Centre, Leiden, for his help and cooperation with the chronic lung disease application. The data were obtained from the Department of Epidemiology, Institute for Social Medical Research, University of Groningen. The authors kindly thank R. van der Lende who made the data available to them.

We also thank the editor and referees of *Psychometrika* for their very helpful comments on previous versions of this paper.

Requests for reprints should be sent to Eeke van der Burg, University of Twente, Department of Education, PO Box 217, 7500 AE Enschede, THE NETHERLANDS.

0033-3123/88/0600-7002$00.75/0

The ALSOS approach to algorithm construction is quite general, but the framework is a bit too narrow for some applications in multivariate analysis, for example, *correspondence analysis* (Benzécri et al., 1973; Benzécri et al., 1980; Nishisato, 1980; Lebart, Morinaux, & Warwick, 1984; Greenacre, 1984). Although correspondence analysis does not fit directly into the ALSOS approach, it is still possible to relate it to the computational developments in ALSOS. This has been done in considerable detail by Gifi (1981), which is summarized briefly in de Leeuw (1984a). In this paper we discuss some of the more specific principles of algorithm construction used by Gifi, and we apply them to OVERALS, a very general nonlinear multivariate analysis technique, covering both ALSOS and correspondence analysis.

The major feature of the Gifi-system for nonlinear multivariate analysis is that it takes *homogeneity analysis* as its starting point. Homogeneity analysis, also known as *multiple correspondence analysis*, is discussed in great detail in the references on correspondence analysis mentioned above, and by Tenenhaus and Young (1985). Gifi introduces homogeneity analysis as the cornerstone of multivariate data analysis, and then specializes to other multivariate techniques by imposing various forms of restrictions on the parameters. Imposing restrictions is one way of dealing with prior information. As a consequence the number of parameters is reduced, which generally improves both the stability and the interpretability of the solution. The most important restrictions are the *additivity* restrictions. These are discussed in detail in this paper in the section on *sets of variables*. In order to fit the classical linear techniques smoothly into the system we also need the *rank-one* restrictions, which can be combined with additivity restrictions to produce a very general class of techniques. Finally *measurement* restrictions are build into the system, in much the same way as in ALSOS. We shall treat these notions in more detail in the section on rank-one restrictions and optimal scaling.

The technique that results if we minimize the general least squares loss function of homogeneity analysis under the types of restrictions mentioned above is called OVERALS. We have to be careful here, because terminological confusion is possible at this point. In the first place we discuss a restricted minimization problem, which we call the OVERALS problem. In the second place we propose an alternating least squares algorithm to solve this minimization problem. This is called the OVERALS algorithm. And thirdly we have written a FORTRAN computer program implementing this algorithm. This is the OVERALS program. It is quite important to keep these three meanings of the word OVERALS apart, although in this paper the context will always indicate which one of the three meanings we are using at any given moment.

Homogeneity Analysis

Homogeneity analysis or *multiple correspondence analysis* is a method to maximize the homogeneity of a number of variables (de Leeuw, 1984b, chap. 3; Greenacre, 1984, chap. 5; Guttman, 1941; Meulman, 1982; Lebart, Morineau, & Warwick, chap. 6; Nishisato, 1980, chap. 5). To define homogeneity analysis we need some notation. Suppose we have an $n \times m$ multivariate data matrix, with rows corresponding to objects and columns to variables. Assume that variable j takes k_j different values (has k_j *categories*) and define the matrix G_j as the $n \times k_j$ *indicator matrix* corresponding to this variable. An indicator matrix indicates which categories are scored by which objects. Rows correspond to objects, columns to categories. Its elements consist of zeroes (not scored) and ones (scored).

Homogeneity analysis determines *quantifications* or *transformations* of the categories of each of the variables such that homogeneity is maximized. A definition of homogeneity follows. Let us use the vector $\mathbf{y}_j$, with k_j elements, for the quantifications of the categories of variable j. Expression $G_j\mathbf{y}_j$ represents a single quantification or transformation of the n

objects, *induced* by variable j. Without further conditions on the $\mathbf{y}_j$ the quantification is restricted only by the ties in the data, that is, objects in the same category get the same quantification. In homogeneity analysis we work with p simultaneous quantifications for each variable (or, to put it differently, with *p-dimensional* quantifications). Let us collect them in $k_j \times p$ matrices Y_j, and let us call these the *multiple nominal quantifications* of variable j. Then the matrices $G_j Y_j$ induce m multiple quantifications of the objects. *Perfect homogeneity* is defined if all multiple quantifications of the objects are the same, say $X(n \times p)$, thus if $X = G_1 Y_1 = \cdots = G_m Y_m$, (de Leeuw, 1984b, chap. 2). Homogeneity analysis minimizes the loss of homogeneity, with loss defined in terms of squared deviations, over normalized object quantifications:

$$\min \sigma(X, Y) = \sum_{j=1}^{m} \mathrm{SSQ}(X - G_j Y_j),$$

$$\text{subject to the condition that } X'X = nI \quad \text{and} \quad \mathbf{u}'X = 0, \tag{1}$$

where $\mathbf{u}$ is a column with n elements equal to one. Symbol $\mathrm{SSQ}(\cdot)$ is used for the sum of squares of the elements of a vector or matrix. The condition $\mathbf{u}'X = 0$ guarantees that X is in deviations from the column means, while $X'X = nI$ makes the columns of X uncorrelated, with variances equal to one. Elements of X are called *object scores.* At this point we do not go further into the formal development of homogeneity analysis, or into computational implementations. We come back to this at a later stage of the paper.

Rank-One Restrictions and Optimal Scaling

In homogeneity analysis with the dimensionality $p \geq 2$ we work with multiple quantifications. Each dimension adds another quantification of the categories of each variable, and the different quantifications of the same variable have no simple relation to each other. This makes interpretation sometimes complicated, especially in the case of variables whose categories have a clear ordinal or even numerical interpretation. For this reason we introduce rank-one restrictions into homogeneity analysis, which make it possible to have multidimensional solutions for object scores with only a single quantification (or *optimal scaling*) for categories. As another benefit the use of rank-one restrictions makes it possible to relate homogeneity analysis to many of the classical multivariate techniques. Mathematically the rank-one restriction (for variable j) is

$$Y_j = \mathbf{z}_j \mathbf{a}_j', \tag{2}$$

with $\mathbf{z}_j$ the k_j-vector of *single category quantifications*, and $\mathbf{a}_j$ the p-vector of *weights.* Thus the quantification matrix Y_j is restricted to be a rank-one matrix. The columns of Y_j are all the same, apart from weight factors.

If no further conditions are imposed on the single quantifications $\mathbf{z}_j$ we call them *single nominal.* Incorporating prior ordinal information on the categories can be done by requiring that the elements of $\mathbf{z}_j$ are in the appropriate order. This defines the *single ordinal* treatment of a variable. *Single numerical* restrictions can also be quite useful. We may require that $\mathbf{z}_j$ is linear with known scores for the categories. All these restrictions are *discrete*, because variables have a restricted number of categories. There are consequently many tied observations, and ties in the data remain ties in the representation. In the *continuous* treatment of variables, as in the *primary approach* to ties of Kruskal (1964), ties can become untied. Because homogeneity analysis is firmly based on the indicator matrix, it does not allow untying of ties, and consequently our approach has no continuous treatment of variables.

The combination of homogeneity analysis with the rank-one restrictions defines a form of nonlinear principal component analysis. We shall discuss this as one of the various special cases below, but first we introduce the implementation of sets of variables into homogeneity analysis.

Sets of Variables

In many applications of multivariate analysis the variables are grouped in a natural way into *sets of variables*. Think of multiple regression for instance, where one has a number of independent variables, or of canonical correlation analysis. One way of dealing with sets of variables in homogeneity analysis is by using *interactive coding*, familiar from the analysis of variance. Variables which belong together are collected as *subvariables* of one interactive variable, and the analysis is applied to the interactive codings instead of to the original variables.

For a set of r subvariables the interactive variable has categories corresponding to all cells of the r-dimensional cross table. Thus using interactive coding can rapidly lead to a very large number of categories. For 5 subvariables with 5 categories, the interactive variable has 3125 categories, which is far too much for any data analysis technique. Almost all cells will be empty, especially if we cross this gigantic variable with others. Nevertheless we may still feel that the subvariables really belong together for the purposes of the analysis we are interested in, and that they form a set of variables in a natural way. We can try to avoid the empty cell problem by imposing *additivity restrictions* on the interactive variables. In analysis of variance terminology this means that we require that the category quantifications for the interactive variables consist of *main effects* only, without interactions between subvariables.

We now translate the above into mathematical notation. The index set $J = \{1, \ldots, m\}$ for variables is partitioned into subsets $J(1), \ldots, J(k)$, where k is the number of sets of variables. We use t for the index indicating sets, thus in the sequel always $t = 1, \ldots, k$. The homogeneity analysis problem with k sets of variables is now defined (de Leeuw, 1984b) as

$$\min \sigma(X, Y) = \sum_t \mathrm{SSQ}\left(X - \sum_{j \in J(t)} G_j Y_j\right),$$
$$\text{subject to the condition that } X'X = nI \quad \text{and} \quad \mathbf{u}'X = 0. \tag{3}$$

Subvariables within sets are treated by (3) as *additive*. Thus, conceptually, sets of variables are dealt with by first creating interactive variables, and then by imposing additivity restrictions. Therefore all within set interactions vanish if variables are coded as concatenated indicators. It is also possible to require that only some within set interactions vanish by leaving some of the interactive codings intact. For instance a set with 4 variables can be coded as 6 concatenated indicators corresponding with all pairs of variables, or as two concatenated indicators, the first one corresponding with three subvariables, and the second one with the remaining subvariable.

The Definition of OVERALS

In the introduction we defined OVERALS as the combination of homogeneity analysis with optimal scaling and additivity restrictions. Now we are ready for a more formal definition. This involves the combination of (2) and (3) into the problem

$$\min \sigma(X, Y) = \underline{\sum_t \mathrm{SSQ}}\left(X - \sum_{j \in J(t)} G_j Y_j\right),$$
$$\text{subject to the condition that } X'X = nI \quad \text{and} \quad \mathbf{u}'X = 0, \tag{4}$$
$$\text{and for some (sub)variables } Y_j = \mathbf{z}_j \mathbf{a}_j' \quad \text{and} \quad \mathbf{z}_j \in C_j,$$

which is the definition of the OVERALS problem. In (4) we have used the general notation $\mathbf{z}_j \in C_j$ to indicate that there may be *measurement restrictions* on the category quantifications (numerical, ordinal, and nominal). The measurement level in (4) is consequently *mixed*, not only because we can choose between single nominal, single ordinal, and single numerical, but also because we have multiple nominal as an option as well. We still consider (4) as a form of homogeneity analysis, with restrictions, and we have implemented a technique for solving the OVERALS problem in the OVERALS computer program. In the following section we discuss the algorithm used in this program.

The OVERALS Algorithm

In this section we explain how the OVERALS problem is solved by using an alternating least squares (ALS) algorithm. First we solve the multiple OVERALS problem, which is the OVERALS problem with all measurement levels multiple nominal. Then we solve the general OVERALS problem (with variables having multiple and/or single measurement levels, from now on briefly called *multiple* and *single variables*) by imposing rank-one restrictions on the multiple quantifications corresponding with single variables.

First we introduce some notation which is more convenient than the summation notation within sets used in (3) and (4). We write all G_j corresponding with variables in set t next to each other in the matrix $\underline{G}_t$, and the Y_j for set t above each other in $\underline{Y}_t$. Thus $\underline{G}_t \underline{Y}_t$ is the sum of all $G_j Y_j$ in set t.

The stationary equations for the OVERALS problem (4) are the following. The optimal object scores $\hat{X}$, for given Y_j, must satisfy the equation

$$\hat{X}\Phi = \underline{M \sum_t} \underline{G}_t \underline{Y}_t, \tag{5}$$

with Φ a symmetric matrix of Lagrange multipliers, and $M = [I - n^{-1}\mathbf{u}\mathbf{u}']$ the operator which transforms a vector into deviations from the mean. Equation (5) is obtained by differentiating the loss function with respect to X under the restrictions that $\mathbf{u}'X = 0$ and $X'X = nI$. If we write Z for the right-hand side of (5), then premultiplying both sides by their transposes gives $n\Phi^2 = Z'Z$. Thus $\Phi = (Z'Z/n)^{1/2}$, and $\hat{X} = n^{1/2}Z(Z'Z)^{-1/2}$. Computing the optimum X is actually a form of the Orthogonal Procrustes problem, for which the solution is classical (Cliff, 1966). The right hand side of (5) is the average of the multiple transformed sets of variables, where each transformed set is the sum of a number of transformed variables. The optimal matrix of object scores is an orthogonalized version of this average.

The optimal category quantification of variable j of set t is

$$\begin{gathered} \hat{Y}_j = D_j^{-1}G_j'(X - V_{tj}), \qquad \text{with} \\ V_{tj} = \underline{G}_t\underline{Y}_t - G_jY_j \quad \text{and} \quad D_j = G_j'G_j. \end{gathered} \tag{6}$$

In order to show that (6) does indeed give the optimal multiple quantifications we write

$$\begin{aligned} \mathbf{SSQ}(X - \overline{\mathrm{G}}_t \overline{\mathrm{Y}}_t) &= \mathbf{SSQ}((X - V_{tj}) - G_j Y_j) \\ &= \mathbf{SSQ}((X - V_{tj}) - G_j \hat{Y}_j) + \mathrm{tr}\,(\hat{Y}_j - Y_j)'D_j(\hat{Y}_j - Y_j). \end{aligned} \tag{7}$$

Clearly the minimum over Y_j is obtained by setting Y_j equal to $\hat{Y}_j$. The matrix D_j is diagonal, and contains the frequencies of the different categories of variable j. The operator $D_j^{-1}G_j'$ averages over objects belonging to the same category, that is, computes category means. We average the object scores X minus a correction term V_{tj} for the other variables in set t. Note that in the "one variable in each set" case, the correction term is zero. In that case the optimal category quantification is the average or centroid of the object scores of all objects in the category.

The two equations (5) and (6) illustrate the *centroid principle* which is one of the leading principles in correspondence analysis. Category quantifications are centroids of objects scores (with a correction for other variables, if necessary), and object scores are averages of quantified variables (with an orthogonalization, if necessary). The multiple OVERALS problem is solved by an ALS-procedure which alternates Step (5), combined with the Procrustes orthogonalization, and Step (6). The centroid principle in the stationary equations (5) and (6) is implemented by a *reciprocal averaging* algorithm.

The general OVERALS problem is the multiple OVERALS algorithm with an extra inner iteration step for single variables (i.e., variables with rank-one restrictions) added. The inner iteration step consists of estimation of weights and single category quantifications, again it alternates two steps of an inner ALS-procedure. We could continue the inner iterations until convergence before proceeding with outer iterations again, but computational experience has indicated that performing only one inner iteration is generally more efficient (Takane, Young, & de Leeuw, 1980).

The multiple category quantifications (6) are computed for all variables, both multiple and single. Weights and single category quantifications are solved for each single variable separately. In order to show how this must be done optimally, we use the partitioning of the sum of squares in (7), assuming now that Y_j is the currently optimal multiple quantification, and $\mathbf{z}_j$ the current single quantification. Thus

$$\begin{aligned}\mathrm{SSQ}(X - \overline{\mathrm{G}}_t \overline{\mathrm{Y}}_t) &= \mathrm{SSQ}(X - V_{tj}) - G_j Y_j) \\ &= \mathrm{SSQ}((X - V_{tj}) - G_j Y_j) + \mathrm{tr}\,(Y_j - z_j' a_j)' D_j (Y_j - \mathbf{z}_j' \mathbf{a}_j). \qquad (8)\end{aligned}$$

Define

$$\hat{\mathbf{a}}_j = (\mathbf{z}_j{}' D_j \mathbf{z}_j)^{-1} Y_j' D_j \mathbf{z}_j. \qquad (9)$$

The last term of (8) can now be written as

$$\mathrm{tr}\,(Y_j - \mathbf{z}_j' \mathbf{a}_j)' D_j (Y_j - \mathbf{z}_j' \mathbf{a}_j) = \mathrm{tr}\,(Y_j - \mathbf{z}_j' \hat{\mathbf{a}}_j)' D_j (Y_j - \mathbf{z}_j' \hat{\mathbf{a}}_j) + \mathbf{z}_j{}' D_j \mathbf{z}_j (\hat{\mathbf{a}}_j - \mathbf{a}_j)' (\hat{\mathbf{a}}_j - \mathbf{a}_j), \qquad (10)$$

which shows that $\hat{\mathbf{a}}_j$ is optimal. In the same way we can define

$$\hat{\mathbf{z}}_j = (\mathbf{a}_j' \mathbf{a}_j)^{-1} Y_j \mathbf{a}_j, \qquad (11)$$

and write

$$\mathrm{tr}\,(Y_j - \mathbf{z}_j' \mathbf{a}_j)' D_j (Y_j - \mathbf{z}_j' \mathbf{a}_j) = \mathrm{tr}\,(Y_j - \hat{\mathbf{z}}_j' \mathbf{a}_j)' D_j (Y_j - \hat{\mathbf{z}}_j' \mathbf{a}_j) + \mathbf{a}_j' \mathbf{a}_j (\hat{\mathbf{z}}_j - \mathbf{z}_j)' D_j (\hat{\mathbf{z}}_j - \mathbf{z}_j). \qquad (12)$$

Now Y_j and $\mathbf{a}_j$ are the current values of the multiple category quantifications and the weights, respectively. We see from (12) that (11) is optimal for single nominal variables. For single ordinal variables the transformations are obtained by using monotone regression (MR), with weights D_j, on the single nominal solution. (See also Young, 1981.) The regression is based on the original ordering of the categories in the data matrix. Thus for single ordinal the optimum is

$$\hat{\mathbf{z}}_j = \mathrm{MR}\{(\mathbf{a}_j' \mathbf{a}_j)^{-1} Y_j \mathbf{a}_j\}, \qquad (13)$$

and for single numerical transformations we use linear regression (LR) instead. Thus

$$\hat{\mathbf{z}}_j = \mathrm{LR}\{(\mathbf{a}_j' \mathbf{a}_j)^{-1} Y_j \mathbf{a}_j\}. \qquad (14)$$

Summarizing the OVERALS algorithm we have: an alternating least squares procedure estimating the objects scores plus orthogonalization (5), and for each variable the multiple category quantifications (6). If there are single variables the single category quantifications and the weights are also estimated in a separate ALS-procedure of which one step is carried out in each major iteration. Then (6) is followed by (9), (11), (13) and (14).

Relationship Between OVERALS and Eigenvalue Problems

In this section we discuss the OVERALS loss function for the multiple case, and the general mixed case a bit more in detail. We do this to relate the technique to various more familiar concepts from linear multivariate analysis. More specifically we want to investigate if and in how far OVERALS solves eigenvector-eigenvalue problems.

Let us start with multiple OVERALS. Remember that G_j was the indicator matrix of variable j, and $\underline{G}_t$ was the supermatrix containing all G_j in set t, obtained by writing the G_j next to each other. It follows directly from (4) that the minimum of the loss over the $\underline{Y}_t$, for fixed X, is attained at $\underline{Y}_t = [\underline{G}_t]^+X$, with $[\cdot]^+$ denoting the Moore-Penrose inverse. Substituting in (4) gives

$$\sigma(X,*) = \sum_t \operatorname{tr} X'\{I - \underline{P}_t\}X, \tag{15}$$

with $\underline{P}_t = \underline{G}_t[\underline{G}_t]^+$, the orthogonal projector on the subspace spanned by the columns of $\underline{G}_t$. Minimization of (15) over X, subject to the normalization conditions specified in (4), gives the stationary equation

$$\sum_t \{M\underline{P}_t M\}X = kX\Theta, \tag{16}$$

with Θ a symmetric matrix of Lagrange multipliers. This shows that the optimal X is a basis for the eigenspace spanned by the p principal eigenvectors of the matrix $M\underline{P}*M$, with $\underline{P}*$ the average of the projectors $\underline{P}_t$. The minimum loss is given by

$$\sigma(*,*) = nkp\left\{1 - p^{-1}\sum_s \lambda_s\right\}, \tag{17}$$

with λ_s the p largest eigenvalues of $M\underline{P}*M$ (and also of Θ). This shows that solving the multiple OVERALS problem corresponds to solving the eigenvalue problem for $M\underline{P}*M$, and that the minimum loss is a linear function of the average of the p largest eigenvalues. In fact it suffices to consider the eigenvalue problem for $\underline{P}_*$, as $M\underline{P}_*M$ is the deflated $\underline{P}_*$ matrix with the first trivial eigenvector, which has all elements the same, removed. The eigenvalue problem could also be solved directly, by using a Jacobi or Householder-Givens algorithm, but this is quite impractical in many situations, because the number of objects can be very large indeed.

It is of considerable interest to observe that instead of solving the eigenvalue problem for $\underline{P}_*$ in order to find the optimal X, we can also solve the generalized eigenvalue problem for the pair $(\underline{C}, k\underline{D})$ in order to find the optimal Y. Here $\underline{C}$ is the *Burt-matrix* of the problem, defined by $\underline{C} = \underline{G}'\underline{G}$, with $\underline{G}$ having all $\underline{G}_t$ next to each other (or, what amounts to the same thing, all G_j next to each other). Matrix $\underline{C}$ contains the bivariate cross tables of all pairs of variables. Compare Gifi (1981, p. 62), or Greenacre (1984, p. 140). Matrix $\underline{D}$ is block-diagonal, it is the direct sum of the $\underline{G}_t'\underline{G}_t$. Thus the optimal Y satisfies

$$\underline{C}Y = k\underline{D}Y\Theta. \tag{18}$$

The proof is short. Because $\underline{P}_*X = k^{-1}\underline{G}\underline{D}^{-1}\underline{G}'X = X\Theta$ and $\underline{D}^{-1}\underline{G}'X = Y$ we have $\underline{G}Y = kX\Theta$. Premultiplying both sides with $\underline{D}^{-1}\underline{G}'$ gives $\underline{D}^{-1}\underline{C}Y = kY\Theta$, which is (18).

Using (18) may be, at least in some situations, an attractive way to compute the optimal solutions of the homogeneity analysis problem with sets of variables. In other cases, however, this generalized eigenvalue problem may be simply too large. Above that, the whole development only applies if all variables are treated as multiple.

For OVERALS with single quantifications only we follow a similar procedure to study the optimal solutions. We introduce some new notation to do this efficiently. Define, for each variable, the vector $\mathbf{q}_j = G_j \mathbf{z}_j$. The $\mathbf{q}_j$ are normalized induced scores for objects, or transformed variables. They are organized as columns of matrices $\underline{Q}_t$, one for each set. In a similar way the weight vectors $\mathbf{a}_j$ are organized as rows of matrices $\underline{A}_t$. We may rewrite the OVERALS problem (4), supposing that all variables are single, as

$$\min \sigma(X, \underline{Q}, \underline{A}) = \sum_t \mathrm{SSQ}(X - \underline{Q}_t \underline{A}_t),$$

$$\text{subject to the condition that } X'X = nI \quad \text{and} \quad \mathbf{u}'X = 0, \tag{19}$$

$$\mathbf{z}_j \in C_j.$$

Now problem (19) is very closely related to our previous OVERALS problem (4). We merely have to replace $\underline{G}_t$ in our previous formulas by $\underline{Q}_t$ and $\underline{Y}_t$ by $\underline{A}_t$. But this means that (16) also applies with $\underline{P}_t = \underline{Q}_t[\underline{Q}_t]^+$. Also $\sigma(*,*)$ is defined as in (17) from the eigenvalues of the average projector $\underline{P}_*$. If we write all $\underline{Q}_t$ next to each other in $\underline{Q}$, then we can also compute $\sigma(*,*)$ as in (17) from the generalized eigenvalues of $\underline{C} = \underline{Q}'\underline{Q}$ with respect to k times the direct sum of the $\underline{Q}_t'\underline{Q}_t$. There is one considerable difference between (19) and its predecessors, however. The vectors $\mathbf{q}_j$ are functions of the $\mathbf{z}_j$, which means that the average projector $\underline{P}_*$ and the Burt matrix $\underline{C}$ are a function of the single category quantifications as well. Thus we can write

$$\sigma(*, \underline{Q}, *) = nkp\left\{1 - p^{-1} \sum_s \lambda_s(\underline{Q})\right\}. \tag{20}$$

Result (17) shows that multiple OVERALS amounts to computing eigenvalues of a given matrix, result (20) shows that single OVERALS means choosing single quantifications of the variables in such a way that the sum of the p largest eigenvalues is maximized. Of course $\underline{Q}$ is constant if all variables happen to be single numerical.

We can now combine our results so far to obtain the interpretation of the minimum loss for the mixed case, in which some variables are single and some are multiple. But we shall introduce a somewhat different terminology, which makes the comparison more interesting. We use the notion that a multiple variable can be considered as a number of *copies* of a single variable. Or, somewhat differently, a multiple variable is really a set of single variables. This idea is due to de Leeuw (1983, 1984a).

Suppose Y_j is a given multiple quantification. We can decompose Y_j, a matrix with k_j rows and p columns, in many different ways in the form $Y_j = Z_j A_j$. One solution simply takes the columns of Z_j as the normalized version of the columns of Y_j, and takes A_j equal to the diagonal matrix of standard deviations of these columns. But Z_j could also be an orthogonalized version of Y_j, with A_j symmetric or upper-triangular, and so on. In any case the decomposition can be written as

$$Y_j = \sum_r \mathbf{z}_{jr} \mathbf{a}_{jr}', \tag{21}$$

and thus

$$G_j Y_j = \sum_r \mathbf{g}_{jr} \mathbf{a}_{jr}'. \tag{22}$$

Here index r is used for the columns of Z_j and the rows of A_j in the decomposition of Y_j.

If there are ρ_j such rows, then (21) and (22) show that having a multiple variable is equivalent to having ρ_j single variables *with the same indicator matrix* G_j, that is, ρ_j *copies*. Note that in general we can take $\rho_j \leq \min(p, k_j)$.

By using the idea of copies we reduce the mixed problem, with both single and multiple variables, to the single OVERALS problem, and we can use the interpretation of this problem in terms of eigenvalues of the Burt-tables and average projectors defined by means of the $\underline{Q}_t$ given above. An additional benefit of use of copies is that it becomes easy to define multiple ordinal and multiple numerical variables. We can fix the measurement level of each of the factors in the decomposition separately. Thus we can, for instance, use one variable three times in its set, once ordinal and two times nominal. If all copies of a variable are ordinal, then it is multiple ordinal. This opens many new possibilities, but we merely outline them here, because the use of copies is not yet implemented in the program OVERALS. If one wants to use the notion of copies in the program, one actually has to create the copies in the data set.

We have shown in this section that OVERALS can be interpreted in terms of eigenvalue problems. In the mixed multiple and single numerical case these eigenvalue problems could be defined completely in terms of the data. OVERALS then becomes the simultaneous iteration method for computing a few of the dominant eigenvalues of a matrix, and it consequently converges to the global minimum of its loss function (Rutishauser, 1969). In the other cases the eigenvalue problem varied with the single quantifications, and we had to choose the quantifications in such a way that the dominant eigenvalues were maximized. This is a nonlinear problem, which may have many local minima. We do not know how serious the local minimum problem is. All nonlinear multivariate analysis problems, except the eigenvalue problems, have to take the existence of local minima into account. The little research that has been done, by Segijn (1984) and Kuhfeld (1985) in the PRINCALS/PRINQUAL framework, shows that local minima do not appear to be a serious problem. But it is not known how general this finding is.

The Computer Program OVERALS

The OVERALS algorithm as described above has been implemented in a computer program which is also called OVERALS (Verdegaal, 1986). It has been developed at the Department of Data Theory by the authors of the article, and it has been written in FORTRAN.

In the OVERALS program three initializations are performed. The object scores X are initialized by using random values (the user determines p). For single variables the quantifications are set equal to the standardized versions of the original data. The multiple category quantifications are initialized as zero. The program starts by computing a solution which has all multiple variables multiple nominal and all single variables single numerical. After convergence of these initial iterations the measurement levels of the single variables are adjusted to the types specified by the user, and the iterations are restarted. This strategy seems to prevent the occurrence of local minima rather effectively, at least in the case in which the measurement level of the variables is single ordinal. A random initialization for the category quantifications is also possible. In case of single nominal variables we advise the use of one or several random starts.

In the program the iteration process is stopped when the loss difference between consecutive main steps is small enough. The user may define "small enough."

Another feature of the OVERALS program is the final rotation. After convergence the object scores X and the category quantifications Y_j are rotated in such a way that the X are the eigenvectors of the matrix $M\underline{P} * M$, and not merely a rotation of these eigenvectors. The eigenvalues of this matrix, which are called the *generalized canonical correlations* by de Leeuw (1984a), are a measure of the goodness-of-fit of OVERALS. To find some indication for the significance of these statistics. De Leeuw and van der Burg (1986) have studied their permutation distribution. They found that the significance testing methods they developed seemed to work rather well, but their study has a somewhat limited scope.

Geometry of OVERALS

In the preceding sections we have discussed object scores and multiple and single category quantifications. How do we interpret the values of these parameters geometrically? Let us make pictures in p-dimensional space (in practice, of course, we can only plot two- or three-dimensional projections of these pictures). The object scores X define a cloud of n points in this space, with unit variance in all directions. The projections on the different dimensions are uncorrelated.

We can compute the centroids of the objects which correspond to the same category of each variable (see Figure 7). We call these values the *category centroids*, in formula rows of $D_j^{-1}G_j'X$. In general these centroids are different from the multiple category quantifications given in (6), except if there is only one variable in the set. If we put category centroids and multiple category quantifications together in one plot, we can "see" the influence of the other variables in the set.

The single category quantifications $\mathbf{z}_j$, together with the weights $\mathbf{a}_j$, can be used to construct the *rank-one quantifications*. By plotting the multiple category quantifications and the rank-one quantifications $\mathbf{z}_j\mathbf{a}_j'$ in a single plot, we see the effect of the rank-one restrictions. The rank-one quantifications are on a line through the origin, with direction cosines proportional to $\mathbf{a}_j$. The transformed variables $\mathbf{q}_j = G_j\mathbf{z}_j$ can be correlated with the object scores X to produce the *component loadings* $\mathbf{c}_j$. The name is chosen in analogy with principal component analysis. They can be depicted as vectors representing transformed variables in the space of the object scores (see Figure 2). We can also plot, in the same space, the *average rank-one quantifications* $\mathbf{z}_j\mathbf{c}_j'$, which are the projections of each category into the space of object scores (see Figure 4). These are different from the $\mathbf{z}_j\mathbf{a}_j'$, because the $\mathbf{c}_j$ are the correlations of $\mathbf{q}_j$ with X, while the $\mathbf{a}_j$ are the correlations of $\mathbf{q}_j$ with $X - V_{tj}$. Thus again the difference is the contribution of the other variables.

In two-sets canonical correlation analysis it is more usual to show plots of the canonical variables for both sets, which are the $\underset{\sim}{G}_t\underset{\sim}{Y}_t$, than of the object scores. If there are only two sets, $\underset{\sim}{G}_1\underset{\sim}{Y}_1$ and $\underset{\sim}{G}_2\underset{\sim}{Y}_2$ are orthogonal, and related by a diagonal transformation. If the number of sets is larger the canonical variables are no longer orthogonal, and they may differ more fundamentally. Therefore we prefer object score plots, but one can, of course, plot canonical variables for each of the sets if this seems desirable.

Relationship With Other Multivariate Techniques

It is interesting to consider the relationship between the OVERALS technique and other linear and nonlinear multivariate techniques. We can be brief about the relationship with homogeneity analysis. If each set contains only one variable, and all variables are multiple nominal, then OVERALS is identical to homogeneity analysis. This special case has been implemented in the program HOMALS (van de Geer, 1985). If there are only two variables, and both these variables are multiple nominal, then OVERALS is equivalent to correspondence analysis.

If each set contains only one variable, but the measurement levels are mixed, then OVERALS defines a form of nonlinear principal component analysis. This technique has been implemented in a separate program PRINCALS (Gifi, 1985). The related PRIN-

CIPALS program of Young, Takane, and de Leeuw (1978) does not have multiple options, but can handle continuous variables. PRINCIPALS is now implemented in PRINQUAL (Kuhfeld, Sarle & Young, 1985). If all variables are single numerical, and each set contains only one variable, OVERALS becomes ordinary principal component analysis.

If there are two sets of variables we move into the realm of canonical correlation analysis. In fact if all variables are considered single numerical OVERALS becomes equivalent to ordinary canonical correlation analysis. If only one interactive variable is reduced to a set of variables by using additivity restrictions, while the other interactive variable is left intact (coding treatment effects), we can use OVERALS to perform multivariate analysis of variance. If one set of single variables is combined with a set containing one multiple nominal variable (coding a partition of the objects), we can perform canonical discriminant analysis. An OVERALS of two sets of single variables is very close, but not exactly identical, to the nonlinear canonical correlation technique CANALS proposed by van der Burg and de Leeuw (1983), and van der Burg (1983). CANALS is an improvement of MORALS/CORALS proposed by Young, de Leeuw, and Takane (1976).

Canonical analysis techniques with k sets of variables were proposed in the single numerical case by many authors. Two early contributors are Horst (1961) and Carroll (1968). Kettenring (1971), Gifi (1981, chap. 6), and van de Geer (1986, pt. IV) provide reviews. It is possible to think of OVERALS, with all variables single, as a nonlinear generalization of one of these generalized forms of canonical correlation analysis. In fact it is a k-set canonical correlation analysis with optimal scaling. The difficulty with this interpretation (from the didactical point of view) is the step from single OVERALS to OVERALS with both single and multiple quantifications. This step is not very natural, and we need the notion of copies to bridge the gap between multiple and single (section on the relationship of OVERALS with eigenvalue problems). Therefore we have chosen the alternative route of starting with homogeneity analysis, and introducing OVERALS by discussing the use of additivity and rank-one restrictions. For the other route, via generalized canonical correlation analysis, we refer to van der Burg, de Leeuw, and Verdegaal (1984).

Application of OVERALS

The data of this study are based on field surveys on chronic lung disease, carried out at three year intervals between 1972 and 1982 in the Netherlands (van der Lende et al., 1981; van Pelt, Quanjer, Wise, van der Burg, & van der Lende, 1985). The locations were a rural area, Vlagtwedde, and an industrial town, Vlaardingen, the latter having a much higher grade of air pollution. The residents of both towns have been questioned, amongst other things, about their smoking behavior, their respiratory symptoms and their personal background. The smoking behavior has been operationalized by four variables: SMO, RATE, PER, and TIME; respiratory symptoms by five variables: COU, PHLE, DYS, WHE, and AST. As background variables we used SEX and AGE. The residence is denoted by RES. The variables and the meaning of the categories are given in Table 1.

There are 2870 individuals sampled from a data base of 3959 individuals under 56 years of age. Starting from the distribution of AGE for the total data base, we sampled four groups (denoted MR = men from rural Vlagtwedde, MI = men from industrial Vlaardingen, and WR, WI for the women) with identical AGE-distributions, so that there exists no correlation between AGE and SEX × RES. This was done to avoid trivial relationships, mainly between AGE and RES (on the average people in rural areas are older).

The goal of the OVERALS analysis was to find a common space in the four sets

TABLE 1

Variables from the study of chronic lung disease.

set 1	RES:	Residence, (1) Vlagtwedde, (2) Vlaardingen.
set 2	SMO:	Smoking, (1) never smoker, (2) ex-smoker, (3) current smoker.
	RATE:	Rate of smoking (amount of tobacco), (1) never smoker, (2) low rate, ... , (9) high rate.
	PER:	Smoking period, (1) never smoker, (2) short period, ..., (13) long period.
	TIME:	Time since last cigarette, (1) never smoker, (2) long ago, ... , (5) recently, (6) current smoker.
set 3	AGE:	Age discreticized into periods of 3.5 years, (1) age 19 - 22.5, ... , (10) age 52.5 - 56.
	SEX:	Sex, (1) male, (2) female.
set 4:	COU:	Coughing, (1) no, (2) persistent.
	PHLE:	Phlegm, (1) no, (2) persistent.
	DYS:	Dyspnoea or shortage of breath, (1) no, (2) slight/ moderate, (3) severe.
	WHE:	Wheezing, (1) never, (2) ever, (3) severe.
	AST:	Asthma, (1) ever, (2) never.

determined by the respiratory symptoms, smoking behavior, personal background, and residency.

We did four analyses, starting with 2870 individuals and all variables single nominal, except AGE which was taken as single ordinal. The same analysis was repeated for men and women separately. Finally another analysis on all 2870 individuals was performed, but now the variables AGE and SEX were combined to one interactive variable AGE × SEX, taken as multiple nominal, and the other variables were taken as single nominal. We considered only two-dimensional solutions. We discuss the results of the analyses with the help of plots. We show transformations of several variables (Figure 1), component loadings (Figures 2, 5, and 6), object scores (Figures 3 and 7), and average rank-one quantifications (Figure 4). In addition we have two tables which give correlations (Table 2) and eigenvalues (Table 3). We do not show the weights as they are difficult to interpret due to the fact that they "incorporate" the correlations with the other variables in the set (Geometry of OVERALS, or Thorndike, 1977).

An overall impression of the first analysis (men & women I) is obtained from the component loadings (Figure 2). However, before we are able to interpret this figure we have to study the transformations of the variables. We find that the single nominal restriction for most variables results in almost ordinal transformations. The exceptions are the smoking behavior variables RATE, PER and TIME. Transformation plots of all smoking variables and of AGE are given in Figure 1. The violations of ordinality occur

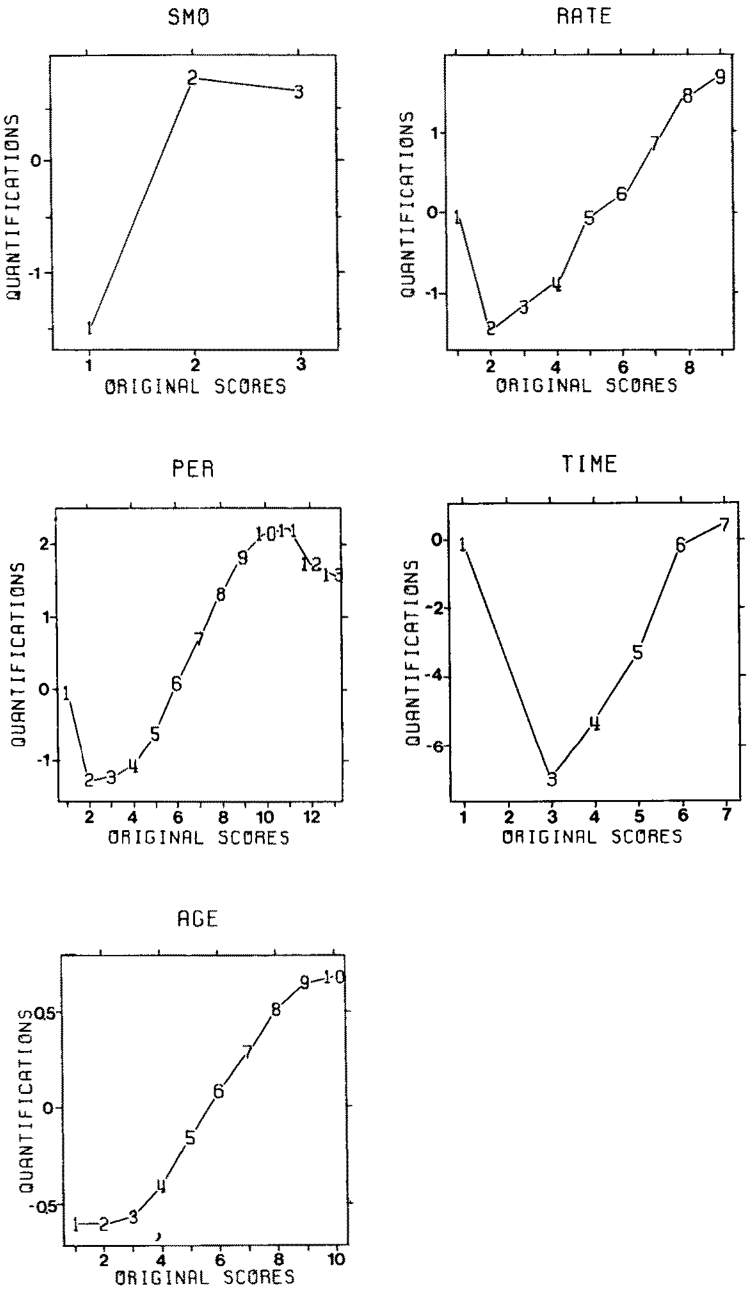

FIGURE 1
Transformations of smoking behavior variables and AGE, men & women I.

mainly in the first categories of RATE, PER and TIME, which correspond to people who have never smoked. Due to the nonlinear transformations of the variables we expect differences between the correlations before and after transformation (respectively upper and lower triangle of Table 2). However the overall structure of the correlation matrix does not change a great deal, except for the submatrix of smoking habits. They form a tight cluster before transformation (mainly related to sex). After transformation they split

TABLE 2

Correlations before and after transformation, respectively upper and lower triangle, men and women I.

	RES	SMO	RATE	PER	TIME	AGE	SEX	COU	PHLE	DYS	WHE	AST
RES		.00	.04	.03	.02	.00	-.06	.09	.11	.05	.04	.04
SMO	.04		.75	.71	.97	-.07	-.32	.18	.12	.02	.17	-.02
RATE	.02	.03		.64	.74	-.03	-.39	.25	.18	.10	.20	-.01
PER	.02	.01	.26		.73	.41	-.43	.19	.15	.14	.18	.01
TIME	-.07	.03	.38	.17		-.08	-.34	.16	.11	.02	.16	-.01
AGE	.00	-.06	.01	.67	-.15		.00	.06	.07	.22	.08	.04
SEX	-.06	-.35	-.23	-.26	.03	.00		-.10	-.09	.06	-.06	.01
COU	.09	.15	.20	.12	.05	.06	-.10		.53	.24	.31	.17
PHLE	.11	.10	.16	.11	.04	.07	-.09	.53		.25	.31	.13
DYS	.05	.02	.12	.19	.01	.23	.06	.23	.24		.33	.20
WHE	.04	.15	.13	.09	.04	.07	-.06	.28	.27	.29		.31
AST	.04	.00	-.02	.02	-.04	.04	.01	.17	.13	.19	.32	

up into age-related smoking habits (PER and TIME) and sex-related smoking habits (SMO and RATE). This is mainly because of the quantification for the nonsmokers category.

The component loadings, which are the correlations between object scores and transformed variables, are plotted as vectors in Figure 2. They point towards a high quantification. As we have seen, this means that they point to individuals having high category numbers for all variables. We only have to keep in mind that the categories for nonsmokers are quantified around zero, and that ex-smokers and current smokers have the same quantification in this solution. The component loadings are interpreted in the usual way. Thus a high age corresponds to a long period of smoking and to severe dyspnoea. The respiratory symptoms, except DYS, are much more related to SEX than to AGE. As the vectors for symptoms and SEX point into opposite directions their relationship is negative. Thus in this sample men more often have symptoms than women. The SEX-vector and the SMO-vector are opposite too, thus also men in this sample are more often ex-smokers than women.

In addition to plotting variables we plotted individuals by their object scores (Figure 3). Together with the object scores we present the 90-percentile contours (equiprobability ellipses) of each of the four SEX × RES groups MR, MI, WR, and WI. The figure shows that men differ from women. Also that the difference between Vlagtwedde and Vlaardingen is larger for women than for men. To obtain more insight in the plot of object scores with respect to the other variables we projected single category quantifications of all variables onto the space of object scores (Figure 4). Above we referred to these projections as average rank-one quantifications. The categories of the variables lie on

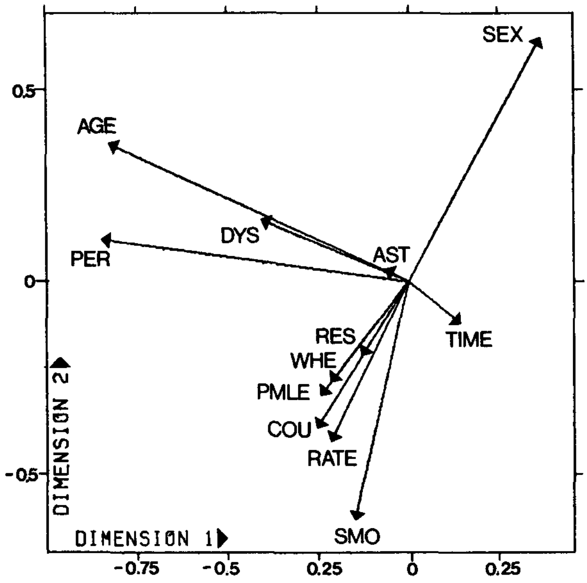

FIGURE 2
Component loadings, men & women I.

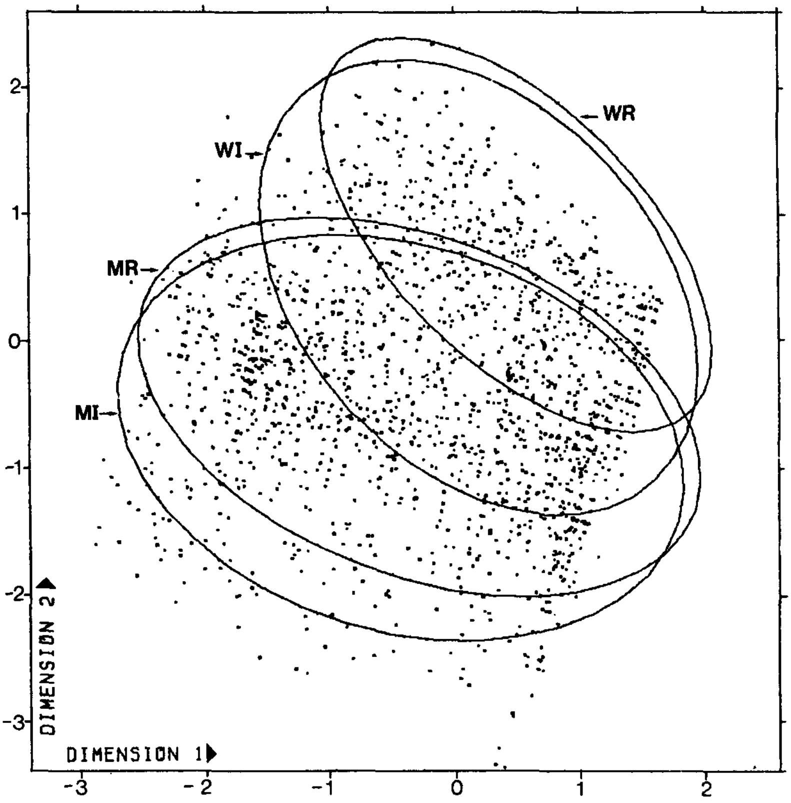

FIGURE 3
Object scores and 90-percent contours for SEX × RES, men & women I. (M = men, W = women, R = Vlagtwedde, I = Vlaardinger).

lines with the same direction as the vectors of Figure 2. To keep Figures 3 and 4 legible, they have been plotted with different scales. In Figure 4 the categories are indicated by the first (or first two) letters of their variable name and their category number (RE = RES, S = SMO, R = RATE, P = PER, T = TIME, A = AGE, SE = SEX, C = COU, PH = PHLE, D = DYS, W = WHE, AS = AST). Only the categories in the middle are left out of the plot. Thus categories which are missing in the plot have quantifications near zero.

Figure 4 shows how the categories are quantified, and tells how to interpret the object scores. For instance at the left, above the middle, we see categories for older people (AGE-categories A9 and A10) who most likely smoked already a long time (PER-categories P8 to P13), or who stopped smoking long ago (T3 and T4, category T2 does not occur), and probably with a severe dispnoea (D3). This means that we find object scores for people characterized in this way at the left side of Figure 3. (In the slightly oblique vertical direction Figure 4 shows no variation in AGE but much variation in the respiratory symptoms COU, PHLE and WHE, in the smoking variables RATE and SMO, in SEX and in RES. In the lower part of Figure 4 we find categories for people with respiratory symptoms (C2, PH2, W2, W3), most probably men (SE1) living in Vlaardingen (RE2) who smoke(d) a lot (S2, S3, R7, R8, R9). In the upper part we find categories

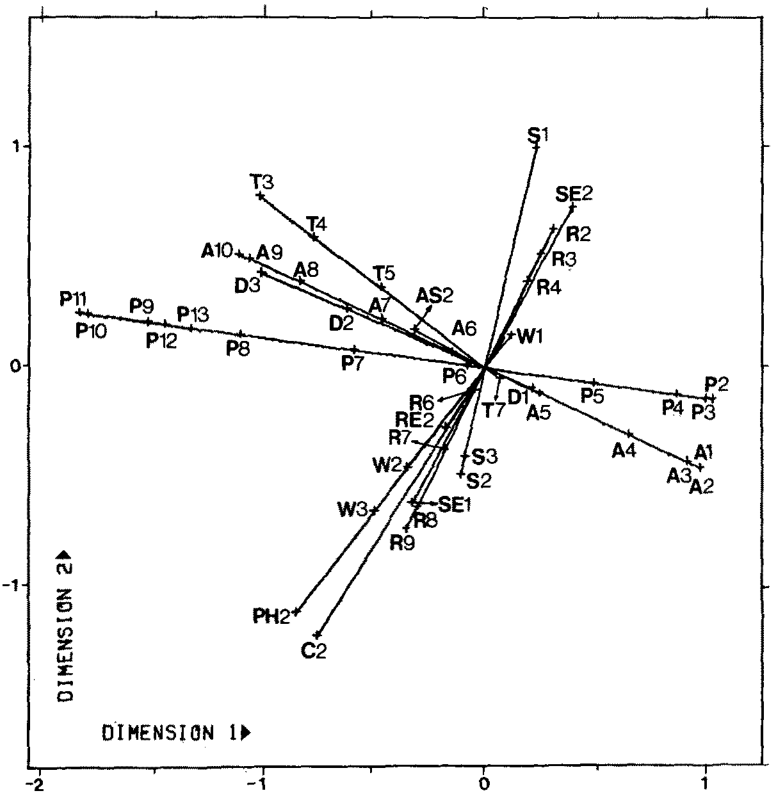

FIGURE 4

Average rank one quantifications, men & women I. (RE = RES, S = SMO, R = RATE, P = PER, T = TIME, A = AGE, SE = SEX, C = COU, PH = PHLE, D = DYS, W = WHE, AS = AST, 1, ..., 10 = category numbers).

for females (SE2) and for never smokers (S1) or very light smokers (R2, R3, R4). Most likely they have no respiratory symptoms (W1, and C1, PH1 in the center). Thus in the plot of the object scores we find healthier people, apart from heaving dyspnoea, more at the top. They are more often women than men, do not smoke or lightly so, live more in Vlagtwedde than Vlaardingen, and are found in all AGE categories.

Differences between men and women with respect to smoking habits and respiratory symptoms are a dominant feature in this solution. We therefore reanalyzed the data separately for men and women. We present the plots of component loadings in Figures 5 and 6. Note that the two plots are on the same scale. In both cases the respiratory symptoms (except DYS) are independent from AGE, and strongly related to RATE. Compared to Figure 2, the variable DYS has moved away from AGE, apparently because we have controlled for SEX. In fact shortage of breath (DYS) occurs equally often in women as in men and correlates mainly with age. It also correlates with the other symptoms, but in the two-dimensional solution of males and females together there was no "place" to show that.

Figures 5 and 6 show that the smoking period, PER, correlates more with AGE for men than women. Also we see that SMO has a different direction and length for the two solutions. This is a reflection of the fact that between 1972 and 1982 most older women do not smoke, whereas the neversmokers in males are usually the younger ones.

Another difference between the solutions for men and women is in the role of residence. For men this variable is totally unexplained, for women it is very pronounced in the solution. The respiratory symptoms correlate with the rate of smoking for both men and women, but they only correlate with residence for women (Figures 5 and 6). This indicates that fewer women in Vlagtwedde smoke than in Vlaardingen, or they smoke less. It seems therefore that the difference in smoking behaviour between males and females, and between the two residences among females, is a more important predictor than place of living as such.

Up till now we found a strong effect of AGE (independent from symptoms, except DYS) both in the total analysis and in the separate analyses for men and women. We also found a large difference between males and females. Therefore we reanalyzed the data, but in this case with the interactive variable AGE × SEX taken as multiple nominal (men & women II). The results confirm the conclusions of the first analysis. We show the catego-

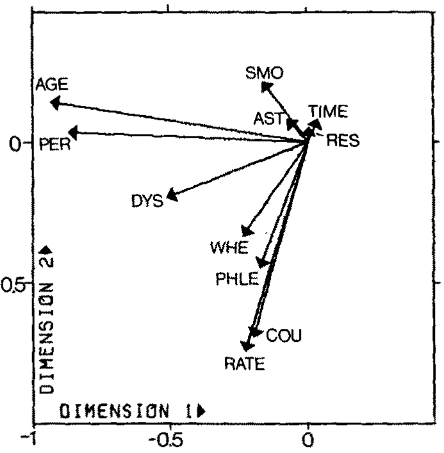

FIGURE 5
Component loadings, men.

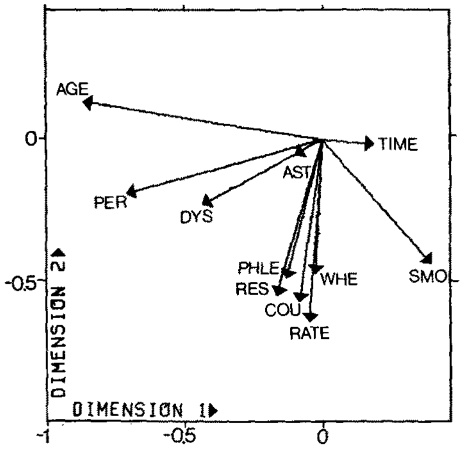

FIGURE 6
Component loadings, women.

ries of AGE × SEX (M1, ..., M10, W1, ..., W10) in the space of object scores (Figure 7). Each category point is in the centroid of (the object scores of) all individuals scored in that particular category. The quantifications form a letter *V* bent leftwards. In fact north-west is still the direction of increasing age, and north-east still the direction of SEX-difference. Categories M1 and W1 overlap, W2 and W3 have changed order, as have W9 and W10, and M9 and M10. But the interchanges are, on the whole, minor. The category quantifications of the other variables are very similar to those of Figure 4, we do not show them. Although there is an interaction effect between SEX and AGE (the younger females and males differ less from each other than older ones do) we can easily describe the effect by two separate variables as the results of the two analyses do not differ substantially.

Summarizing the four analyses we can say that we found a relationship between smoking behaviour and respiratory symptoms for both males and females. Only for women we also found an effect of residence with respect to respiratory symptoms. This effect can be reduced to a difference in smoking habits between women from Vlaardingen and Vlagtwedde. Sex is correlated with both symptoms and smoking behaviour. Age is mostly related to smoking variables with a time effect, such as TIME and PER. The symptoms are not related to age (in the age range we have considered), except shortage of breath. We found an interaction effect between AGE and SEX. Younger people differ less

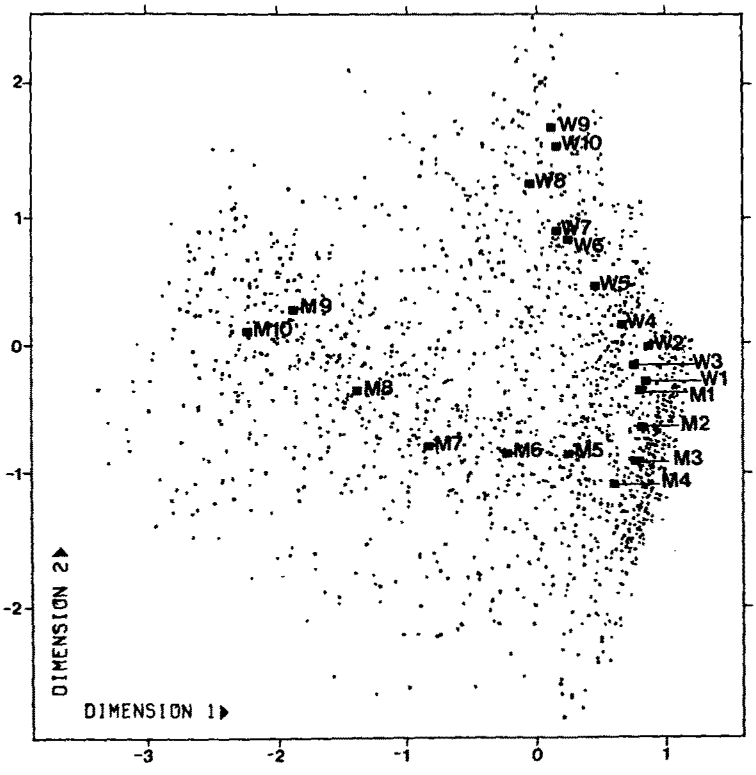

FIGURE 7
Object scores and category centroids for AGE × SEX, men & women II. (M = men, W = women, 1, ..., 10 = age categories).

in symptoms and smoking habits than older people do. The nonlinear transformation of the variables (first analysis) has effected mostly the smoking habit variables. Mainly due to the quantification for the category nonsmokers the cluster of smoking habits falls apart after transformation. For completeness we finish this application with an overview of the generalized canonical correlations (Table 3). Perfect homogeneity corresponds with a correlation of 1, and no relation at all with a canonical correlation of $1/k$. From Table 3 it can be seen that for men the first dimension is much more important than the second one. For the other analyses the two dimensions are more of equal importance.

We emphasize that this example is only a tiny demonstration of the capabilities of OVERALS. There are so many choices and options in the program, that we can never cover the complete range of possibilities. We refer to Gifi (1981) for other examples. Many applications of special cases of OVERALS can be found throughout that book.

Discussion and Extensions

The OVERALS algorithm opens many possibilities in data analysis. It covers most of the usual linear and nonlinear multivariate analysis techniques. But this generality comes at a price. In the first place there is the possibility of local minima in some of the more complicated special cases. It is necessary to study the seriousness of this problem in more detail in the future. In the second place we do not have information on the stability of the results. For several special cases of OVERALS (two variables, or k sets each with one variable) research has been done, however for the more general cases of OVERALS not very much is known. De Leeuw and van der Burg (1986) make a start by means of randomization methods. They compare several methods and obtain promising results. They investigate the stability of generalized canonical correlation in a small study. More work in this direction has been planned. Van der Burg and de Leeuw have investigated ways of computing confidence regions for the OVERALS results. For this they use the Delta method combined with the Jackknife. Their results are encouraging, but still very preliminary.

Another apparent disadvantage of the OVERALS method is the fact that it can only handle complete data matrices. We did not discuss missing values in this article. The computer program OVERALS does handle missing data, however, on the basis of equations given by Gifi (1981, chap. 6). Verdegaal (1985, 1986) gives an extensive discussion of the OVERALS program with missing data.

The nonlinear transformations in OVERALS are a real extension of the usual linear transformations in multivariate analysis. However the transformations we use are necessarily step functions. This can be a disadvantage in some cases. To make transformations more smooth we can, for instance, use splines. De Leeuw, van Rijckevorsel, and van der

TABLE 3

Generalized Canonical Correlations.

	1	2
men & women I	.469	.390
men	.510	.317
women	.426	.362
men & women II	.486	.398

Wouden (1981) have implemented splines in the principal component algorithm. We plan to integrate these transformations into OVERALS as well.

With these extensions OVERALS can effectively be applied in even more data analysis situations.

References

Benzécri, J. P. et al. (1973). *L'Analyse des données* [Data analysis] (2 vols.). Paris: Dunod.

Benzécri, J. P. et al. (1980). *Pratique de l'Analyse des données* [Practice of data analysis] (3 vols). Paris: Dunod.

Carroll, J. D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th Annual Convention of the American Psychological Association, 5,* 227–228.

Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika, 31,* 33–42.

de Leeuw, J. (1983, July). *Nonlinear joint bivariate analysis.* Paper presented at the meeting of the Psychometric Society, Jouy-en-Josas, France.

de Leeuw, J. (1984a). The Gifi-system of nonlinear multivariate analysis. In E. Diday, M. Jambu, L. Lebart, J. Pagès, & R. Thomassone (Eds.), *Data Analysis and Informatics III* (pp. 415–424). Amsterdam: North Holland.

de Leeuw, J. (1984b). *Canonical analysis of categorical data.* (Doctoral dissertation, University of Leiden, 1973). Leiden: DSWO-Press.

de Leeuw, J. & van der Burg, E. (1986). The Permutational limit distribution of generalized canonical correlations. In E. Diday, Y. Escoufier, L. Lebart, J. P. Pagès, Y. Schektman, & R. Thomassone (Eds.), *Data analysis and informatics IV* (pp. 509–521). Amsterdam: North Holland.

de Leeuw, J., van Rijckevorsel, J., & van der Wouden, H. (1981). Nonlinear principal component analysis using B-splines. *Methods of operations research, 23,* 211–234.

Gifi, A. (1981). *Nonlinear multivariate analysis.* Department of data theory, University of Leiden. (In Press, Leiden: DSWO-Press).

Gifi, A. (1985). *PRINCALS.* (User's Guide UG-85-03). Department of Data Theory, University of Leiden.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.* New York: Academic Press.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), *The prediction of personal adjustment.* New York: Social Science Research Council.

Horst, P. (1961). Relations among *m* sets of measures. *Psychometrika, 26,* 129–149.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika, 56,* 433–451.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika, 29,* 1–28.

Kuhfeld, W. F. (1985). *Principal components of ordered categorical data.* Unpublished doctoral dissertation, University of North Carolina.

Kuhfeld, W. F., Sarle, W. S., & Young, F. W. (1985). Methods in generating model estimates in the PRINQUAL macro. *SUGI-Proceedings* (pp. 962–971), Cary, NC: SAS-Institute.

Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate descriptive analysis.* New York: Wiley.

Meulman, J. (1982). *Homogeneity analysis of incomplete data.* Leiden: DSWO-Press.

Nishisato S. (1980). *Analysis of categorical data: Dual scaling and its applications.* Toronto: University of Toronto Press.

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

Segijn, R. (1984). Lokale minima in PRINCALS [Local minima in PRINCALS]. Unpublished master's Thesis, Department of Data Theory, University of Leiden.

Takane, Y., Young. F. W., & de Leeuw, J. (1980). An individual differences additive model: An alternating least squares method with optimal scaling features. *Psychometrika, 45,* 183–209.

Tenenhaus, M. & Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika, 50,* 91–120.

Thorndike, R. M. (1977). Canonical analysis and predictor selection. *Multivariate Behavioral Research, 12,* 75–87.

van de Geer, J. P. (1985). *HOMALS.* (User's guide UG-85-02). Department of data theory, University of Leiden.

van de Geer, J. P. (1986). *Introduction to multivariate data analysis* (2 vols.). Leiden: DSWO-Press.

van der Burg, E. (1983). *CANALS* (User's guide UG-85-05). Department of Data Theory, University of Leiden.

van der Burg, E., & de Leeuw, J. (1983). Non-linear canonical correlation. *British Journal of Mathematical and Statistical Psychology, 36,* 54–80.

van der Burg, E., & de Leeuw, J. (1985). Use of the multinomial jackknife in generalized canonical correlation analysis. Paper presented at the Multidimensional Data Analysis Workhop, Cambridge, G.B.

van der Burg, E., de Leeuw, J., & Verdegaal, R. (1984). *Nonlinear canonical correlation with m sets of variables* (Research Report RR-84-12) Leiden: University of Leiden, Department of Data Theory.

van der Lende, R., Kok, T. J., Peset Reig, R., Quanjer, Ph. H., Schouten, J. P., & Orie, N. G. M. (1981). Decreases in VC and FEV with time: Indicators for effects of smoking and air pollution. *Bulletin Européen de Psychopathologie Respiratoire, 17,* 775–792.

van Pelt, W. J., Quanjer, Ph. W., Wise, M. E., van der Burg, E., & van der Lende, R. (1985). Analysis of maximum expiratory flow-volume curves using canonical correlation analysis. *Methods of Information in Medicine, 24,* 91–100.

Verdegaal, R. (1985). *Meer-sets analyse voor kwalitatieve gegevens* [Multi-set analysis of qualitative data] (Research Report RR-85-14). Department of Data Theory, University of Leiden.

Verdegaal, R. (1986). *OVERALS* (User's guide UG-86-01). Department of Data Theory, University of Leiden.

Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika, 46,* 347–388.

Young, F. W., de Leeuw, J., & Takane, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika, 41,* 505–529.

Young, F. W., Takane, Y., & de Leeuw, J. (1978). The principal components of mixed measurement multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika, 43,* 279–281.

Manuscript received 7/2/84
Final version received 10/24/86