RR – 84-12

NON-LINEAR CANONICAL CORRELATION WITH M SETS OF VARIABLES

Eeke van der Burg
Jan de Leeuw
Renée Verdegaal

Department of Data Theory
University of Leiden

An earlier version of this article has been presented at the Meeting of the Psychometric Society in Jouy-en-Josas, France, July 6, 7 and 8, 1984.

This research was partly supported by SWOV (Stichting Wetenschappelijk Onderzoek Verkeersveiligheid).

Requests for reprints should be addressed to Eeke van der Burg, Department of Data Theory, University of Leiden, Middelstegracht 4, 2312 TW Leiden, The Netherlands.

NON-LINEAR CANONICAL CORRELATION ANALYSIS WITH M SETS OF VARIABLES

ABSTRACT

Non-linear canonical correlation analysis, OVERALS, is introduced as a generalization of both linear canonical correlation analysis and homogeneity analysis. The basic notion of OVERALS is minimizing a loss function under certain conditions. In the OVERALS program the parameters are solved for by alternating least squares. OVERALS is invariant under certain transformations of the data. There are multiple and single transformations. The latter ones consist of three types: nominal, ordinal and numerical. An application of OVERALS is treated in the last section.

Keywords

Canonical correlation, loss function, homogeneity analysis, indicator matrix, multiple and single quantifications or transformations, optimal scaling, rank one restriction, nominal, ordinal and numerical transformation, non-linear, alternating least squares, eigenvector/ value decomposition, copies.

Introduction

OVERALS is a computer program for non-linear canonical correlation analysis (CCA) with m sets of variables. Gifi (1981) and de Leeuw (1983) give a global introduction to the method of OVERALS. We will present a more detailed treatment of the algorithm.

CCA for two sets of variables is a standard technique, first described by Hotelling (1936). A non-linear version was introduced by Young, de Leeuw & Takane (1976), Gifi (1981a) and van der Burg & de Leeuw (1983). CCA with m sets of variables is a generalization of two set CCA. Many generalizations are possible. Descriptions can be found in Horst (1961), Carroll (1968), Kettenring (1971), Gifi (1981a), de Leeuw (1984) and van de Geer (1984). As in the two set case non-linear CCA with m sets of variables is again a generalization of linear CCA. Which generalizations are used depends on subjective choices. There are interpretive reasons, algorithmic arguments, and historical grounds, van de Geer (1984) treats the impacts of various generalizations. We preferred a relationship with homogeneity analysis as described by Gifi (1981a,b). Therefore we have chosen a method which also generalizes homogeneity analysis to groups of variables (de Leeuw, 1984). In the following sections we explain how various types of techniques are related.

As there are many types of non-linear CCA we will call the method OVERALS. Then the computer program is one realization of the OVERALS method. Throughout the text we will make clear if the are dealing with the method or the program.

Definition of OVERALS as a generalization of linear CCA

Linear CCA (or in short CCA) is a technique applied to m sets of variables. It aims at defining an optimal relationship between the sets at the same time. To be more explicit about the technique we need some notation.

Given a data matrix H of objects (rows) by variables (columns) of dimensions n x s. Suppose the data matrix can be devided into m sets of standardized variables:

$$H=(h_1,......,h_s)=(H_1,......,H_m).$$

Furthermore suppose there are, for each variable j, weight vectors a_j (dimension p). Collect the weight vectors as columns of a matrix A' (dimensions p x m) and group the weights for each set also into submatrices $A_1',......,A_m'$. Thus:

$$A'=(a_1,......,a_s)=(A_1',......,A_m').$$

Linear combinations of the sets are denoted as:

$$H_1A_1,......,H_mA_m.$$

CCA defines for each set those linear combinations of the variables, the canonical variates, which are optimally related under certain conditions. In case of two sets of variables there exists only one criterion for relatedness: the canonical correlation coefficient. However is case of m sets of variables there are many kinds of conditions. Define R as the correlation matrix of $H_1A_1,.....,H_mA_m$. Horst (1961), Kettenring (1971) and van de Geer (1984) optimize properties of R over choice of weights A. Their optimality criteria vary from the largest eigenvalue of R to the sum of squared elements of R. The conditions vary from orthogonality constraints imposed on the canonical variates within sets to an overall orthogonality constraint. Carroll (1968), Gifi (1981a) and de Leeuw (1984) use a different formulation. They introduce an extra parameter X (dimensions n x p) to define their criterion and impose the orthogonality constraints to this parameter. Carroll maximizes the sum of the squared correlations, which gives the same solution. In the terminology of Gifi (1981a) CCA is:

$$\sigma(X,A)=\Sigma_{k=1}^{m}\text{trace}(X-H_kA_k)'(X-H_kA_k) \qquad (1)$$

under the conditions that u'X=0 and X'X=nI.

The vector u is a column vector of ones. The condition u'X=0 guarantees that X is in deviations from column means and X'X=nI ensures X to be orthonormal. Names used for X are object scores, principal components, canonical axes and canonical variates. We will use the first name.

One way to introduce non-linear CCA is by defining non-linear transformations of the variables:

$$t^j(h_j)=q_j. \qquad (2)$$

The q_j (dimension n) are called the optimally scaled variables because the transformations are chosen in such a way that the q_j optimize the CCA criterion. We use three types of transformations: single nominal, single ordinal and single numerical (single because we take only one quantification for each variable). Young, de Leeuw & Takane (1976) describe these types of transformations in relation with CCA of two sets of variables. Young (1981) describes them in the more general framework of ALSOS programs. We summarize their definitions briefly. Single nominal transformations maintain the ties within variables, single ordinal transformations maintain the ordening and ties of a variable, and single numerical transformations are linear transformations.

Let us organize the optimally scaled variables q_j in the same way as the columns of H. This gives:

$$Q=(q_1,......,q_s)=(Q_1,......,Q_m).$$

We always assume q_j to be standardized. Gifi (1981a) defines nonlinear CCA with single transformations as:

$$\begin{aligned}&\text{minimize } \sigma(X,Q,A)=\Sigma_{k=1}^{m}\mathrm{tr}(X-Q_kA_k)'(X-Q_kA_k)\\&\text{under the conditions that } u'X=0,\ X'X=nI \text{ and } q_j\varepsilon C_j.\end{aligned} \qquad (3)$$

C_j defines the set of nominal, ordinal or numerical transformations corresponding to variable h_j. For characteristics of C_j compare de Leeuw (Note 1). With (3) we have introduced the generalization of linear CCA which we call OVERALS with single transformations.

Definition of OVERALS as a generalization of homogeneity analysis

A different way to introduce non-linear CCA is as a generalization of homogeneity analysis or multiple correspondence analysis (Guttman, 1941; de Leeuw, 1973; Benzécri et al., 1973; Gifi, 1981a and b). Gifi formulates homogeneity analysis with a loss function using indicator matrices for each variable. Assume variable j has k_j different categories, then indicator matrix G_j has dimensions n x k_j. The elements of G_j are defined as:

$$g^j_{1,r}=\begin{cases}1 \text{ if object 1 falls in category r}\\0 \text{ if object 1 does not fall in category r.}\end{cases}$$

If we use vector y_j (dimension k_j) for the quantifications of the categories of variable j, the expression G_jY_j is a single nominal transformation q_j of the original variable h_j. Suppose we have more than one quantification (say p). Collect them in a matrix Y_j with dimensions k_j x p. Also combine the Y_j into a super-matrix Y of dimensions Σk_j x p. The Gifi (1981a) formulation of homogeneity analysis for m variables is:

$$\text{minimize } \sigma(X,Y)=\Sigma^m_{k=1}\text{tr}(X-G_jY_j)'(X-G_jY_j) \tag{4}$$

under the conditions that u'X=0 and X'X=nI.

This technique is implemented in the computer program HOMALS (Gifi, 1981b). We can look upon homogeneity analysis as a form of non-linear CCA with one variable in each set. Generalization to more variables in the sets is possible for instance by summation of the quantified variables within sets. This gives:

$$\text{minimize } \sigma(X,Y)=\Sigma^m_{k=1}\text{tr}(X-\Sigma_{j\varepsilon I_k}G_jY_j)'(X-\Sigma_{j\varepsilon I_k}G_jY_j) \tag{5}$$

under the conditions that u'X=0 and X'X=nI.

I_k contains the variables indices of set k. We call this non-linear CCA, OVERALS, with multiple nominal quantifications. De Leeuw (1973) gives a description of this technique. Saporta (Note 3) discussed it at the Meeting of the Psychometric Society at Groningen. Gifi (1981a) shows how multiple and single transformations can be combined into one and the same loss function. In fact multiple OVERALS (i.e. OVERALS with multiple nominal transformations) is the most general one. It can be seen as a generalization of single OVERALS. We will show this in the next paragraph.

The definitions of OVERALS are given in (3) for single transformations and in (5) for multiple transformations. To see the relationship between the two definitions we restrict the multiple category quantification Y_j in (5) to be rank one matrices. In that case Y_j can be written as:

$$Y_j=z_ja_j' \tag{6}$$

z_j= single category quantifications (k_j x 1)

a_j= canonical weights (p x 1).

Expression (6) means that the multiple category quantifications are the same for all p solutions, viz. z_j, except for a constant given in a_j. A variable which is quantified p times and of which the quantifications are restricted to be a rank one matrix looks as follows:

$$G_jY_j=G_jz_ja_j'=q_ja_j'. \tag{7}$$

Let us collect the q_j and a_j into matrices Q_k and A_k' and also into supermatrices Q and A' in the same way as we did earlier. Then the sum over transformed variables of set k is:

$$\Sigma_{j\epsilon I_k}G_jY_j=\Sigma_{j\epsilon I_k}q_ja_j'=Q_kA_k. \tag{8}$$

Using (8) we can reformulate the loss of (5):

$$\sigma(X,Y)=\Sigma_{k=1}^{m}\text{tr}(X-\Sigma_{jeI_k}G_jY_j)'(X-\Sigma_{j\epsilon I_k}G_jY_j)$$

$$=\Sigma_{k=1}^{m}\text{tr}(X-Q_kA_k)'(X-Q_kA_k). \tag{9}$$

The loss in (9) is identical to the loss in (3), which shows that single quantifications are a special case of multiple quantifications.

Now we have seen that single and multiple OVERALS are related we can come to a definitive definition of OVERALS, namely with mixed single and multiple quantifications. It means that for some variables there are rank one restrictions as in (6) and not for others. The OVERALS problem is:

$$\text{minimize } \sigma(X,Y)=\Sigma_{k=1}^{m}\text{tr}(X-\Sigma_{j\varepsilon I_k}G_jY_j)'(X-\Sigma_{j\varepsilon I_k}G_jY_j)$$

$$\text{under the conditions that } u'X=0,\ X'X=nI \text{ and} \qquad (10)$$

$$\text{for some variables } Y_j=z_ja_j' \text{ and } G_jz_j\varepsilon C_j.$$

C_j defined as in (3). With (10) we have formulated OVERALS in the most general sense.

Relationship with other non-linear techniques

It is interesting to consider the relationship between OVERALS as defined in (10) and other non-linear multivariate techniques. Except for a link with homogeneity analysis there also exists a link with principal component analysis. OVERALS with one variable in each set defines a form of PCA. If variables can be multiple or single then PCA with OVERALS is the same as PRINCALS (Gifi, 1981a, 1983 and de Leeuw & van Rijckevorsel, 1981). PRINCIPALS described by Young, Takane & de Leeuw (1978) corresponds to OVERALS-PCA with single nominal and ordinal transformations. PRINQUAL (Tenehaus, 1977) corresponds to single nominal and numerical OVERALS-PCA.

If we restrict the number of sets to two we find a relationship between OVERALS and CANALS, which is non-linear CCA with two sets of variables (Gifi, 1981a; van der Burg & de Leeuw, 1983 and van der Burg, 1983). OVERALS is not exactly a generalization of CANALS, but there exist a close relationship. By eliminating parameter X the single OVERALS problem (3) for two sets of variables can be translated into:

$$\text{minimize } \sigma(Q,A)=\text{tr}(Q_1A_1-Q_2A_2)'(Q_1A_1-Q_2A_2) \tag{11}$$

under the conditions that

$$A_1'Q_1'Q_1A_1=nI,\ A_2'Q_2'Q_2A_2=nI \text{ and } q_j\varepsilon C_j.$$

Van der Burg & de Leeuw (1983) discuss this problem. The solution they suggest corresponds to a problem very similar, but not identical, to (11).

Another example of non-linear CCA with two sets of variables is given by Young, de Leeuw & Takane (1976). They treat the case of one variable in one of the sets (MORALS) and of more variables in both sets (CORALS). CANALS is the successor of CORALS. MORALS is a special case of CANALS, but also of OVERALS with single transformations.

Here we finish our introduction to OVERALS. We will concentrate in the next sections on the solutions of the OVERALS-parameters and the value of the loss. The last section will deal with an application of OVERALS.

The solution of the OVERALS parameters

The parameters of the multiple OVERALS problem (5) have the following solutions. The optimal object scores X, for given quantification Y_j, are:

$$X\Phi=J\frac{1}{m}\Sigma_{k=1}^{m}\Sigma_{j\varepsilon I_k}G_jY_j \tag{12}$$

with Φ symmetric Lagrange multiplier. The matrix J (n x n) is the operator which transforms the next matrix to deviations from column means. The optimal category quantifications Y_j, for given X, are:

$$Y_j=D_j^{-1}G_j'(X-V_j^k) \text{ with} \tag{13}$$

$$V_j^k=\Sigma_{\substack{l\varepsilon I_k\\ l\neq j}}G_lY_l \text{ and } D_j=G_j'G_j.$$

The matrix D_j is diagonal and contains the frequencies of the different categories of variable j. The matrix $D_j^{-1}G_j'$ averages over objects belonging to the same category. We average the object scores X minus a correction term V_j^k for the other variables in the set. Remark that in the 'one variable in each set' case the correction term is zero.

The OVERALS problem is solved in the program by an iterative procedure, alternating the least squares (ALS) steps (12) and (13) (cf. Young, 1981). The matrix $J\frac{1}{m}\Sigma\Sigma G_jY_j$ is orthogonalized in every iteration step by a Procrustus procedure, which gives a new update for X[1].

In the OVERALS program, for single variables the transformations t^j are added as an additional least squares step to (13). In the single nominal case the quantfications are defined as weighted multiple quantifications:

$$t^j_{nom}:\ z_j=Y_ja_j/a_j'a_j. \tag{14}$$

Single ordinal transformations are monotonically regressed (MR) single nominal quantifications. The regression is based on the original ordening of the categories contained in h_j.

$$t^j_{ord}:\ z_j=MR(Y_ja_j/a_j'a_j). \tag{15}$$

The single numerical quantifications are linearly regressed (LR) single nominal quantifications.

$$t^j_{num}:\ z_j=LR(Y_ja_j/a_j'a_j). \tag{16}$$

For computation of the weights another step is added to the second main iteration step:

$$a_j=Y_j'D_jz_j/z'_jD_jz_j. \tag{17}$$

[1] The final version of X (and Y_j) are rotated such that the object scores are the principal axes of JP, see (30) in the following section.

Step (12) to (17) define the OVERALS program.

To prove that it is correct to add separate steps for transformations and weights to the computation of the multiple category quantifications, we rewrite the loss for single variables. Consider the part concerning variable j of set k.

$$\sigma_{z_j,a_j}=\sigma_j(X,Y_k)=\mathrm{tr}(X-V_j^k-G_jz_ja_j')'(X-V_j^k-G_jz_ja_j'). \tag{18}$$

If we leave out all the subscripts and set $U=X-V_j^k$, then:

$$\sigma_{z,a}=\mathrm{tr}(U-Gza')'(U-Gza'). \tag{19}$$

Define:

$$\bar{Y}=D^{-1}G'U \text{ (i.e. multiple quantifications) and} \tag{20}$$

$$\bar{z} = \bar{Y}a/a'a \text{ (i.e. single nominal quantifications).} \tag{21}$$

The loss $\sigma_{z,a}$ can now be written as:

$$\sigma_{z,a}=\mathrm{tr}\{(U-G\bar{Y})'(U-G\bar{Y})+(\bar{Y}-\bar{z}a')'D(\bar{Y}-\bar{z}a')+a'a(\bar{z}-z)'D(\bar{z}-z)\}. \tag{22}$$

Expression (22) shows that if there is a rank one resstriction to the quantifications the least squares solution of z is $\bar{z}$. This gives t_{nom}. In case there is, in addition, an order or linear restriction to z, expression (22) shows that the least squares solution for z is either the monotonically or the linearly regressed $\bar{z}$. This gives t_{ord} and t_{num}. Thus we see that the least squares step to compute z_j is defined by (13) followed by (14), (15) or (16).

In a similar way we prove that the computation of the canonical weights as described above is correct. Define:

$$\bar{a}=\bar{Y}'Dz/z'Dz. \tag{23}$$

Using this definition the loss of (19) can be rewritten as:

$$\sigma_{z,a}=\mathrm{tr}\{(U-G\bar{Y})'(U-G\bar{Y})+(\bar{Y}-z\bar{a}')'D(\bar{Y}-z\bar{a}')+z'Dz(a-\bar{a})'(a-\bar{a})\}. \tag{24}$$

The third part of this sum of squares gives the solution for the weights as given in (17).

Now we have shown that the computation of both transformations and weights are least squares solutions, we have proved that we follow a proper ALS-method with our definitions of the OVERALS solutions by steps (12) to (17).

The minimum OVERALS loss

To find the minimum OVERALS loss we need a closer look at the solutions for the different parameters. Let us start with only multiple transformations. We need some new notation. We shift from simple matrices G_j and Y_j to super-matrices $\tilde{G}_k$ and $\tilde{Y}_k$. Define:

$$I_k = (k_1, \ldots\ldots, k_\ell)$$

$$\tilde{G}_k = (G_{k_1}, \ldots\ldots, G_{k_\ell})$$

$$\tilde{Y}_k = (Y_{k_1}, \ldots\ldots, Y_{k_\ell}).$$

Now the sum over transformed variables within sets is:

$$\sum_{j \varepsilon I_k} G_j Y_j = \tilde{G}_k \tilde{Y}_k. \tag{25}$$

This changes the loss of (5) into:

$$\sigma(X,Y) = \sum_{k=1}^{m} \operatorname{tr}(X-\tilde{G}_k\tilde{Y}_k)'(X-\tilde{G}_k\tilde{Y}_k). \tag{26}$$

Minimization of $\sigma(X,Y)$ of (26) over $\tilde{Y}_k$, for fixed X, gives:

$$\hat{\tilde{Y}}_k=(\tilde{G}_k'\tilde{G}_k')^{+}\tilde{G}_k'X. \tag{27}$$

A^{+} is used for the Moore-Penrose invers. Substitution of $\tilde{Y}_k$ from (27) in the loss of (26) gives:

$$\sigma(X,*)=\sum_{k=1}^{m}\mathrm{tr}(X-P_kX)'(X-P_kx) \text{ with} \tag{28}$$

$$P_k=\tilde{G}_k(\tilde{G}_k'\tilde{G}_k)^{+}\tilde{G}_k'. \tag{29}$$

Minimization of σ(X,*) over X also considering the conditions imposed on X gives:

$$mX\Phi=J\sum_{k=1}^{m}P_kX, \tag{30}$$

with Φ symmetric Lagrange multiplier. This shows that X is a rotation of the eigenvectors of the matrix:

$$P=\frac{1}{m}\sum_{k=1}^{m}P_k \tag{31}$$

in deviations from column means[1]. The minimum loss σ(*,*) is attained by substituting (30) into (28). This minimum is:

$$\sigma(*,*)=nmp(1-\frac{1}{p}\sum_{i=1}^{p}\Phi_{ii}), \tag{32}$$

which shows the minimum loss to be proportional to the sum of the p larger eigenvalues of matrix JP.

From (30) and (32) we see that a multiple OVERALS analysis is similar to solving an eigenvalue problem defined in terms of G_j. This implies that the iteration process leads to an absolute minimum.

For single OVERALS we follow a similar procedure to compute the minimum loss. Then we solve (3) for X, Q and A, which gives:

$$A_k=(Q_k'Q_k)^{+}Q_k'X \tag{33}$$

$$\tilde{P}=\frac{1}{m}\sum_{k=1}^{m}\tilde{P}_k=\frac{1}{m}\sum_{k=1}^{m}Q_k(Q_k'Q_k)^{+}Q_k' \tag{34}$$

[1] X is also a rotation of the p+1 larger eigenvectors of P minus the first one (which is proportional to u). In fact JP is the deflated P matrix with the first eigenvector removed (de Leeuw, 1973).

$$X\Phi=\tilde{J}PX \qquad (35)$$

$$\sigma(*,*,*)=nmp(1-\frac{1}{p}\Sigma_{i=1}^{p}\Phi_{ii}(Q)). \qquad (36)$$

Thus in the single OVERALS case the minimum loss is a function of the quantifications. This means that, in every iteration step, another eigenvalue problem is approximated (except for the single numerical case, which gives Q as a constant). This implies that the iterative process may be bothered by local minima, which is not the case with multiple (and single numerical) OVERALS.

In the mixed situation the loss is also a function of Q. To find out what function, we use the notion that a multiple variable can be considered as a repeated or copied single variable (this does not change the idea that single quantifications are restricted multiple quantifications). Let us be more explicit. The matrix Y_j can be decomposed in many ways. We define the following decomposition:

$$Y_j=Z_jA_j \text{ with} \qquad (37)$$

$$A_j=\Delta^{\frac{1}{2}}=\{diag(Y_j'D_jY_j)\}^{\frac{1}{2}} \text{ and} \qquad (38)$$

$$Z_j=Y_j\Delta^{-\frac{1}{2}}. \qquad (39)$$

We assume that $p \leqq k_j-1$. This is a real restriction, but does not change the general idea. If we write z_{jr} for the columns of matrix Z_j, we get

$$Y_j=Z_jA_j=\Sigma_{r=1}^{p}z_{jr}a_{jr}'. \qquad (40)$$

Thus variable j has a contribution to the loss of

$$G_jY_j=\Sigma_{r=1}^{p}G_jz_{jr}a_{jr}'. \qquad (41)$$

This means that we are dealing with p single variables corresponding to the same indicator matrix. Therefore multiple variables can be viewed as copied single variables. For a more detailed treatment of copies we refer to de Leeuw (Note 2) and de Leeuw (1983).

If all variables have to be transformed in a multiple way and $p \leq k_j - 1$ for every variable, then the maximum number of variables including copies in the single OVERALS problem is sxp. With some multiple variables the dimensions of the Q matrix is between s and sxp.

Now we have fitted the situation of multiple transformations into the single case, we do not need to compute the minimum loss again. We simply generalize (33), (34), (35) and (36) to larger Q, say $\bar{Q}$, with dimensions n x $\bar{s}$. The minimum loss of mixed OVERALS is:

$$\sigma(*,*,*) = nmp(1-\frac{1}{p}\Sigma_{i=1}^{p}\phi_{ii}(\bar{Q})). \tag{42}$$

As in the single case mixed OVERALS may be bothered by local minima.

In the OVERALS program the loss is never computed as a function of the eigenvalues, but it is directly computed from X and Y_j or $z_j a_j'$. The loss difference corresponding to two iteration steps is taken as criterion to stop the iteration proces.

Here we finish the theoretical part about the OVERALS method and program. The following section deals with an application from the medical field.

Application of Overals[1]

The data of this study are based on field surveys on chronic lung disease, carried out at three year intervals between 1972 and 1983 in the Netherlands[2] (van der Lende et al., 1981; van Pelt et al., 1984). The locations were a rural area, Vlagtwedde, and an industrial town, Vlaardingen, with a comparatively high grade of air pollution. The residents of both towns have been questioned, amongst other things, about their smoking behaviour, their respiratory symptoms and their personal background. The smoking behaviour 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 took SEX and AGE. The variables and the meaning of the categories are given in Table 1. There are 2870 individuals in the sample; we made the distribution of sex combined with age identical for both residences.

Insert Table 1 about here

Our main goal of the OVERALS analysis was to show the relative importance of living in an area of air pollution (Vlaardingen) and of smoking, both with respect to respiratory symptoms. We added sex and age to the analysis as we expected them to correlate with the other variables. We used the set partition as given in Table 1 and applied

[1] We kindly thank Wilfrid van Pelt, Department of Physiology and Physiological Physics, University Medical Centre, Leiden, for his help and cooperation in performing this application.

[2] The data were obtained from the Department of Epidemiology, Institute for Social Medical research, University of Groningen, The Netherlands. We kindly thank Dr. R. van der Lende, who made data available for us.

single nominal restrictions to all variables. For most variables the nominal restriction results in a nearly ordinal transformation. The exceptions concern the variables RATE, TIME and PER. The first category of each of those variables (never smokers) gets a much higher quantification than the categories for light smokers or for those who stopped smoking long ago. Never smokers vary independently from AGE and, as we will see below, (ex-)smokers vary systematically with AGE via PER and TIME, and with SEX via RATE. SEX and AGE are independent in the solution so that the first category of RATE, TIME and PER must be quantified in the mean. The other categories are ordinally transformed.

To get an overall impression of the analysis we give the component loadings, which are the correlations between the object scores and the (transformed) variables. We do not show the canonical weights as they are difficult to interpret due to the fact that they 'incorporate' the correlations with the other variables in the set (Thorndike, 1977). The component loadings are plotted in Figure 1. We computed a two-dimensional analysis, therefore we have two object scores for each object or individual and two component loadings for each variable. In Figure 1 the component loadings are the coordinates of the vectors. These vectors point towards a high quantification which means, in our case, a high category number. Thus a high age corresponds to a long period of smoking and severe dyspnoea. The respiratory symptoms COU, PHLE and WHE hardly correlate with AGE, the occurrence of these symptoms corresponds to a high rate of smoking and males. There is a large difference between men and women. Men in this sample smoke more and have more respiratory symptoms than women, with the result that the SEX-vector (pointing towards women) is nearly in the opposite direction of the SMO-vector (pointing towards (ex-)smokers).

Insert Figure 1 about here

TIME and AST are not very important in this two-dimensional solution, as indicated by their relatively short vectors. The variable VLA lies in the same direction as the respiratory symptoms and opposite to SEX, however it does not contribute much to the solution.

Insert Figure 2 about here

We also present the object scores labeled by the categories of VLA in Figure 2. We mentioned already that the effect of VLA is not very strong, nevertheless even the small effect can be seen. The twos (Vlaardingen) are located lower in the plot than the ones (Vlagtwedde). To obtain more insight in the plot of object scores we could use every variable for labeling the plot, which would result in eleven plots. To present this information more efficiently, the category quantifications have been projected onto the space of object scores (Figure 3). The categories are indicated by the first (two) letter(s) of their variable name and their number. The categories lie on lines with the same direction as the vectors of Figure 1. Note that Figures 2 and 3 have been plotted with different scales. With the same scales they can be combined into one figure simply by placing them on top of each other. Figure 3 shows how the categories are quantified

Insert Figure 3 about here

and tells how to interpret the object scores. For instance at the right (above the middle), we see categories for older people (A9, A10) who, most likely, smoked already a long time (P8 to P13) or who stoped smoking long ago (T2, T3), and probably with a severe dyspnoea (D3). This means that we find object scores for people characterized in this way, at the right side of Figure 2. In the slightly oblique vertical direction Figure 3 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 VLA. In the lower part of Figure 3 we find categories for people with respiratory symptoms (C2, PH2, W2, W3), most probably men (SE1) living in Vlaardingen (V2) who smoke(d) a lot (S2, S3, R7, R8, R9). In the upper part we find categories for females (SE2) and for never smokers (S1) or very light smokers (R2, R3, R4). Most likely they have no respiratory symptoms (C1, PH1, W1). Thus in the plot of object scores we find the healthier people, apart from having dyspnoea, more at the top. They are more often women than men, do not smoke or slightly so, live more often 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 4 and 5. Note that the two

Insert Figures 4 and 5 about here

plots are on the same scale. In both cases the respiratory symptoms (except DYS) are independent from AGE, and are strongly related to RATE. Compared to Figure 1, the variable DYS has moved away from age, apparently because we controlled for SEX. In fact shortage of breath (DYS) happens to the same extent to women as men and correlates 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 4 and 5 show that the time period smoked, PER, correlates more with AGE for men than women. Also we see that SMO has a different direction and length. This is a reflection of the fact that between 1972 and 1983 the larger part of the older women does not smoke, whereas most never smokers in males are found among the younger ones.

Another difference between the solutions for men and women is in the influence of residence. For men this variable is totally unexplained, for women it is very pronounced in the solution. We present the plot of object scores for women labeled by residence in Figure 6.

Insert Figure 6 about here

The effect of residence is obvious. The respiratory symptoms correlate with rate of smoking in both cases, but only for women they also correlate with residence. The latter indicating that fewer women in Vlagtwedde smoke and/or smoke less than the women from Vlaardingen. It seems therefore that the difference in smoking behaviour between males and females, and between the two residences among females, is a more important determinant than place of living per se. Now we can also understand why the effect of residence in the combined solution was not very strong. Because only the women in our study have this difference, and not the men.

Concluding 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 habit between women from Vlaardingen and Vlagtwedde. Sex is correlated with both symptoms and smoking behaviour. Age is mostly related to smoking variables which have a time effect in it (TIME and PER). The symptoms are not related to age with the exception of shortage of breath.

Summary

A new method, OVERALS, for canonical correlation analysis with m sets of variables is introduced. OVERALS has theoretically attractive features. The algorithm is of an alternating least squares types. OVERALS belongs to a group of methods most thoroughly described by Gifi (1981a). In fact OVERALS is the most general method of this group to which homogeneity analysis (HOMALS), non-linear PCA (PRINCALS) and non-linear CCA (CANALS) belongs. OVERALS is a generalization of all those methods: it generalizes HOMALS and PRINCALS to more variables per set and it generalizes CANALS to m sets. Also OVERALS is a generalization of linear CCA, simply because it combines optimal scaling of the data with linear CCA.

The algorithm of OVERALS is based on minimization of a loss function. With OVERALS we search for directions (principal components or object scores) in the space of variables and for transformations or quantifications of the variables (optimal scaling). The object scores and transformations are such that linear combinations of the optimally scaled variables within sets correlate maximally with the object scores. Theoretically solving the loss for object scores means solving an eigenvector/value problem and solving for transformations means

solving a regression problem. As we prefer to approximate and iterate for the solutions, we alternate in the OVERALS program between object scores and transformations plus weights (necessary for the linear combinations), fitting least squares approximations in every iteration step.

The transformations can be of different types. There are multiple and single quantifications. Multiple means for every direction (dimension) a new transformation. Single means one transformation for all directions. The single transformations can be divided into nominal, ordinal and numerical. We can make a hierarchy of transformations. The multiple quantifications are the most general ones, followed by single nominal, single ordinal and finally single numerical quantifications. Multiple and single nominal impose only tie-restrictions per variable on the transformations, single ordinal imposes monotone restrictions and single numerical linear restrictions.

In case of multiple nominal and single numerical transformations the eigenvector/value problem only depends on the data, which means that we have a neatly defined problem with an absolute minimum without local minima. With single nominal and ordinal transformations the eigenvector/value problem depends on the optimally scaled data so that local minima can no longer be excluded.

An application of OVERALS is presented. It concerns a medical research on chronic lung disease.

Reference notes

(1) Leeuw. J. de (1977). Normalized cone regression. Mimeo. Department of Data Theory, University of Leiden.

(2) Leeuw, J. de (1983). Nonlinear joint bivariate analysis. Hand-out meeting Psychometric Society at Jouy-en-Josas, France.

(3) Saporta, G. (1980). About some remarkable properties of generalized canonical correlation analysis. Paper presented at the meeting of the Psychometric Society at Groningen, The Netherlands.

References

Benzécri, J.P. et al. (1973). **L'analyse des Données** (2 vols.). Dunod, Paris.

Burg, E. van der, (1983). CANALS user's guide. Department of Data Theory, University of Leiden.

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

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

Geer, J.P. van de (1984). Linear relations between k sets of variables. **Psychometrika**, 49, 79-94.

Gifi, A. (1981a). **Non-linear Multivariate Analysis**. Department of Data Theory, University of Leiden. In press: DSWO Press, Leiden, 1984.

Gifi, A. (1981b). HOMALS user's guide. Department of Data Theory, University of Leiden.

Gifi, A. (1983). PRINCALS user's guide. Department of Data Theory, University of Leiden.

Guttman, L. (1941). The quantifications of a class of attributes: a theory and method of scale construction. In: **The Prediction**

of Personel Adjustment. P. Horst (ed.). Social Science Research Council, New York, 251-364.

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

Hotelling, H. (1936). Relations between two sets of variates. **Biometrika**, 28, 321-377.

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

Leeuw, J. de (1973). **Canonical Analysis of Categorical Data.** Dissertation, University of Leiden. Reissued: DSWO Press, Leiden, 1984.

Leeuw, J. de (1984). The Gifi system of nonlinear multivariate analyis. In: **Data Analysis and Informatics.** E. Diday et al. (eds.), North-Holland Publishing Company, Amsterdam.

Leeuw, J. de & Rijckevorsel, J. van (1980). HOMALS and PRINCALS, some generalizations of principal components analysis. In: **Data Analysis and Informatics.** E. Diday et al. (eds.), North Holland Publishing Company, Amsterdam, 231-243.

Lende, R. van der, 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 Psychiopathologie Respiratoire**, 17, 775-792.

Pelt, W. van, Quanjer, Ph.H., Wise, M.E., Burg, E. van der & Lende, R. van der (1985). Analysis of maximum expiratory flow-volume curves using canonical correlation analysis. To be published in: Methods of Information in Medicine.

Tenenhaus, M. (1977). Analyse des composantes principales d'un ensemble de variables nominales et numériques. **Revue de Statistique Appliquée**, 25, 39-56.

Thorndike, R.M. (1977). Canonical analysis and predictor selection. The Journal of Multivariate Behavioral Research, 12, 75-87.

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

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

TABLE 1

Variables from the study of chronic lung disease

set 1	VLA:	Residence, (1) Vlagtwedde, (2) Vlaardingen.
set 2	SMO:	Smoking, (1) never smoker, (2) ex-smoker, (3) current smoker.
	RATE:	Rate of smoking (amount of tabacco), (1) never smoker, (2) low rate, ..., (9) high rate.
	PER:	Time period smoked, (1) never smoker, (2) short period, ..., (13) long period.
	TIME:	Time since last cigaret, (1) never smoker, (2) long ago, ..., (5) recently, (6) current smoker.
set 3	AGE:	Age discretisized 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.

Figure 1
Component loadings

Figure 2
Object scores, 1=Vlagtwedde, 2=Vlaardingen.

Figure 3
Category quantifications.

Figure 4
Component loadings, men.

Figure 5
Component loadings, women.

Figure 6
Object scores for women, 1=Vlagtwedde, 2=Vlaardingen.

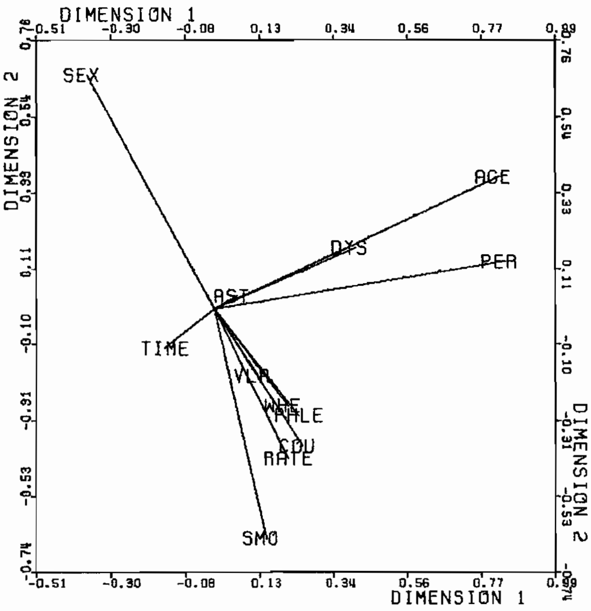
DIMENSION 1
-0.51
-0.30
-0.08
0.13
0.34
0.56
0.77
0.99
DIMENSION 2
0.76
0.54
0.33
0.11
-0.10
-0.31
-0.53
-0.74
SEX
AGE
DYS
PER
AST
TIME
VLA
WHE
PHLE
COU
RATE
SMO

Figure 1

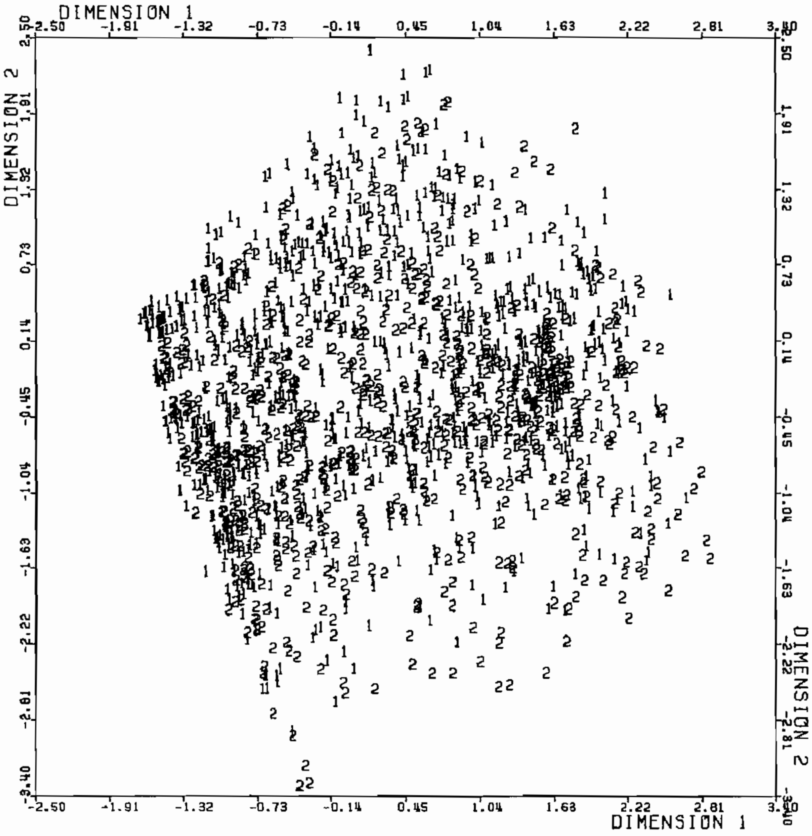
DIMENSION 1
-2.50
-1.91
-1.32
-0.73
-0.14
0.45
1.04
1.63
2.22
2.81
3.40
DIMENSION 2
2.50
1.91
1.32
0.73
0.14
-0.45
-1.04
-1.63
-2.22
-2.81
-3.40

Figure 2

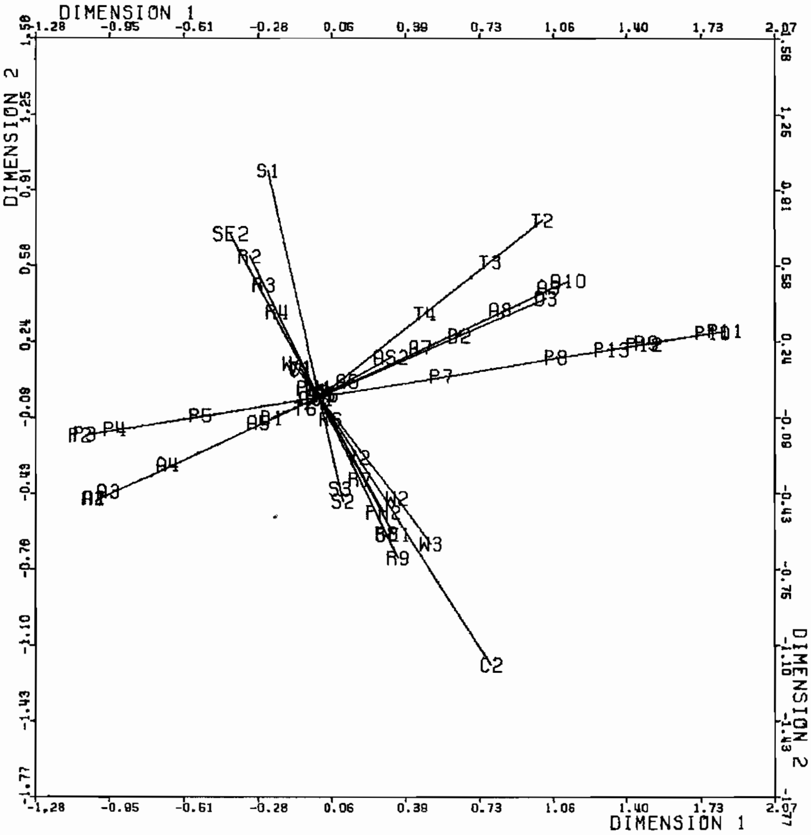
DIMENSION 1
-1.28
-0.95
-0.61
-0.28
0.06
0.39
0.73
1.06
1.40
1.73
2.07
DIMENSION 2
1.58
1.25
0.91
0.58
0.24
-0.09
-0.43
-0.76
-1.10
-1.43
-1.77
S1
SE2
A2
A3
A4
T2
T3
T4
A10
A9
A8
D3
D2
A7
AS2
P7
P8
P13
A12
P11
W1
A6
P5
P4
P2
A1
A5
T6
A4
A3
A2
S3
S2
A7
W2
A12
W3
A9
C2

Figure 3

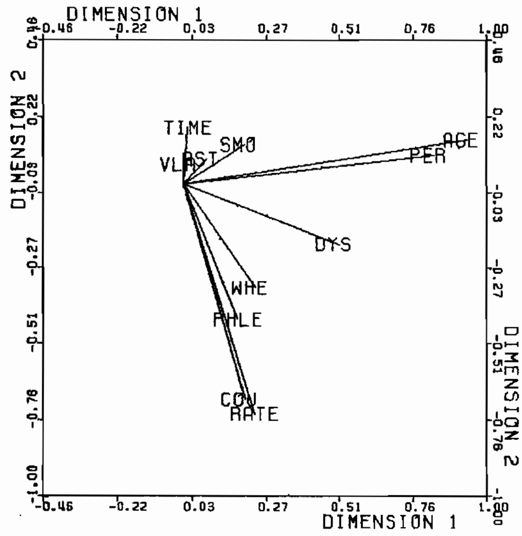

Figure 4

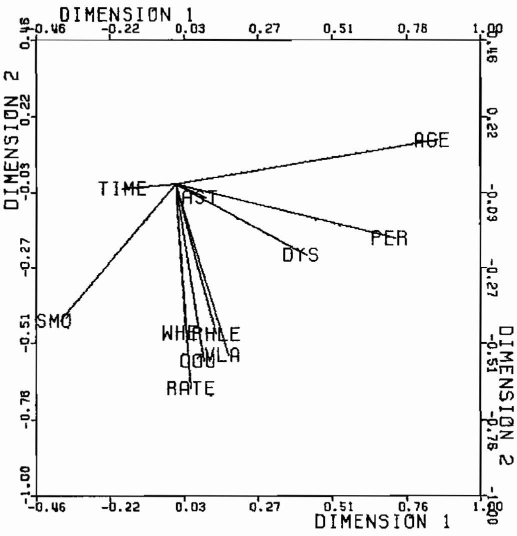

Figure 5

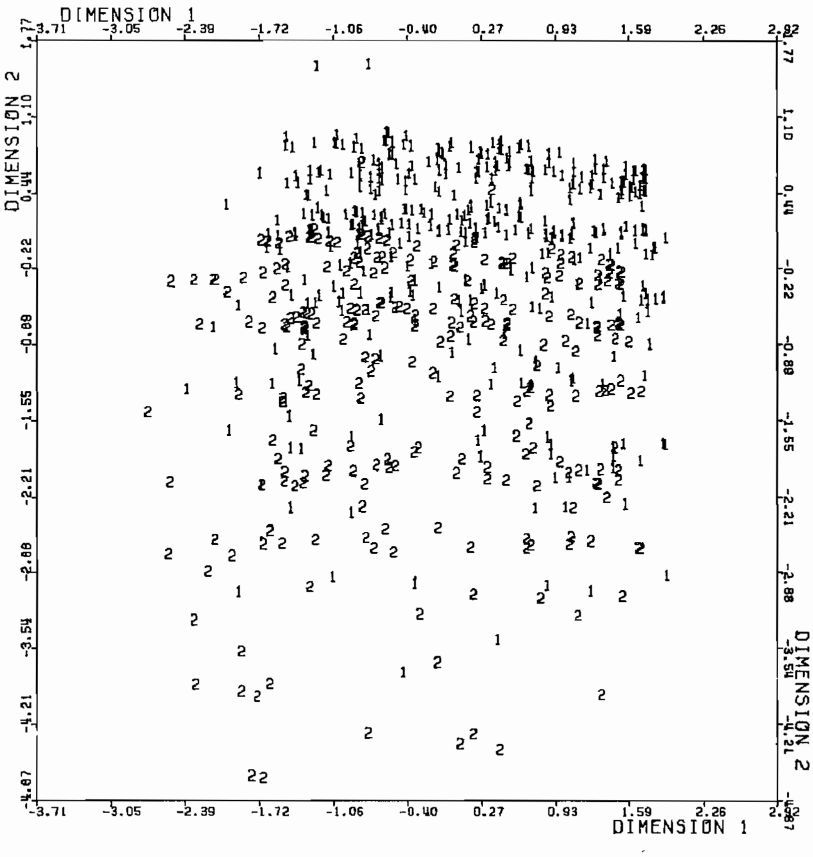
DIMENSION 1
-3.71
-3.05
-2.39
-1.72
-1.06
-0.40
0.27
0.93
1.59
2.26
2.92
DIMENSION 2
1.77
1.10
0.44
-0.22
-0.88
-1.55
-2.21
-2.88
-3.54
-4.21
-4.87

Figure 6