Principal Components Analysis on a mixture of quantitative and qualitative data based on generalized correlation coefficients

Kiers, Henk A.L.

Published in:
The many faces of multivariate analysis (Vol.I)

Publication date:
1988

THE MANY FACES OF MULTIVARIATE ANALYSIS

Proceedings of the SMABS-88 conference,
held in Groningen, December 18-21

VOLUME I

Editors:

Margo G.H. Jansen, IDOK
Wijbrandt H. van Schuur, S & M

RION, Institute for
educational research
University of Groningen

FPPSW,
University of Groningen

PRINCIPAL COMPONENTS ANALYSIS ON A MIXTURE OF QUANTITATIVE AND QUALITATIVE DATA BASED ON GENERALIZED CORRELATION COEFFICIENTS[1]

Henk A.L. Kiers[2]

University of Groningen

In this paper a method for principal components analysis (PCA) is proposed that yields a compromise between PCA on generalized correlation coefficients for a mixture of qualitative and quantitative variables, as suggested by Janson and Vegelius, and a generalization of multiple correspondence analysis for the analysis of such variables. The method proposed here is based on INDSCAL on a set of similarity matrices, that are used as quantification matrices for each of the variables. This method is compared to the two original methods between which it is a compromise.

Principal Components Analysis (PCA) is a useful technique for the exploratory analysis of quantitative variables. The purpose of PCA is to yield optimal representations of the variables and the observation units (called "objects" here) simultaneously. That is, PCA yields component scores for the objects on a limited number of components such that these component scores are the best possible predictors for the scores of the objects on the variables. Optimal representation of the variables is given in an analogous way. The usefulness of PCA lies in the possibility of describing most of the information present in one's data by means of a number of components that is usually much smaller than the number of variables.

PCA can only provide such a solution when the variables are quantitative. For qualitative variables an essentially different approach is needed. On the one hand, there are methods that optimally represent

[1] Financial Support by the Netherlands Organization for the advancement of Pure Research (Z.W.O.) is gratefully acknowledged.
The author is obliged to Jos ten Berge for helpful comments.

[2] Dept. of Psychology, Grote Markt 31/32, 9712 HV Groningen, Tel.:050-636339
EARN: KELINCI at HGRRUG5

qualitative variables, but do not represent objects at all. On the other hand there are methods that optimally represent objects, but by no means represent the qualitative variables in any optimal way. In fact, it seems impossible to obtain an optimal representation of variables and objects simultaneously, when the variables are qualitative. Therefore, Kiers (in press) proposed a method that offers a compromise between methods that optimally represent the variables without representing the objects at all and methods that optimally represent objects only. Kiers' method is a compromise between those two types of methods in that it yields the best possible representation for the variables which simultaneously yields a representation for the objects. Kiers' approach only applies to data sets that consist exclusively of qualitative variables. For the exploratory analysis of data consisting of a mixture of qualitative and quantitative variables a different approach is needed.

The exploratory analysis of a mixture of qualitative and quantitative variables seems to have received far less attention than the exploratory analysis of qualitative variables. Two types of approaches have been considered, both of which are generalizations of methods for the exploratory analysis of qualitative variables. The first type of method generalizes Multiple Correspondence Analysis (MCA) to the effect that it can handle a mixture of qualitative and quantitative variables. The second type of method is based on PCA on generalized correlation coefficients that measure the association between a qualitative and a quantitative variable. As has been shown by Kiers (in press) for the case of qualitative variables, neither MCA nor PCA based on generalized correlation coefficients for qualitative variables are "complete". That is, MCA does not provide an optimal representation of the variables and PCA based on generalized correlation coefficients for qualitative variables does not provide a useful representation for the objects. In the present paper it will be shown that the two types of methods currently available for the analysis of a mixture of qualitative and quantitative variables are similarly incomplete. The purpose of the present paper is to propose a method that yields a compromise between the two existing types of methods for the analysis of a mixture of qualitative and quantitative variables. Before discussing this compromise a more detailed discussion of the two existing types of methods will be given.

The first type of method has been proposed independently by many authors (De Leeuw, 1973; Hill & Smith, 1976; Tenenhaus, 1977; ~~Young, Takane & De Leeuw, 1978~~; Escofier, 1979; Nishisato, 1980, pp.103–107). Most of these authors present their methods as generalizations of MCA to the effect that they can handle a mixture of qualitative and quantitative variables. Although

the methods slightly differ in the way in which quantitative variables are transformed, all methods essentially use the same approach to handle qualitative variables. That is, all methods can be described as a kind of PCA on a matrix containing as columns the indicator variables for the categories of the qualitative variables and (a transformation of) the quantitative variables. The particular method that performs PCA on the indicator matrices of the qualitative variables and on standardized versions of the quantitative variables will be denoted here as MCAMIX. This is the generalization of MCA as it has been proposed originally by Hill and Smith (1976), Tenenhaus (1977), ~~and Young, Takane and De Leeuw (1978).~~

As has been demonstrated by Kiers (in press), the very fact that MCA performs a PCA on the complete set of indicator variables for all qualitative variables causes it to yield a non–optimal representation of the qualitative variables. In fact, MCA yields an optimal representation of the *categories* of the qualitative variables, not of the variables themselves. Analogously, the generalizations of MCA for analyzing a mixture of qualitative and quantitative variables, optimally represent only the categories of the qualitative variables, rather than the qualitative variables themselves, because they use the same approach for the qualitative variables as MCA does.

The second type of method has been outlined by Saporta (1976). Saporta's method finds an optimal representation of a mixture of qualitative and quantitative variables by means of a PCA on a matrix of generalized correlations between all variables. For the generalized correlation between two qualitative variables, Saporta chooses Tschuprow's T^2 coefficient (Tschuprow, 1939), which is the χ^2 measure normalized to the effect that its maximum is 1. For the generalized correlation between a qualitative and a quantitative variable, he proposes to use the correlation ratio, also normalized such that its maximum is 1. Saporta does not explicitly mention which measure one should take for the correlation between two quantitative variables, but following his line of reasoning it is clear that this should be the squared product moment correlation. Janson and Vegelius (1978,1982) propose several generalized correlation coefficients for the purpose of performing a PCA on the generalized correlation matrix. Zegers and Ten Berge (1986) discuss the coefficients proposed by Janson and Vegelius, and show some advantages of some of these coefficients above others.

The generalized correlation coefficients proposed by Saporta (1976) and Janson and Vegelius (1982) share one basic property. That is, all these generalized correlation coefficients are normalized scalar products between "operators" as Saporta calls them, or "quantification matrices" as Zegers and Ten Berge (1986) choose to name them. A quantification matrix is a matrix of

similarities between objects. Each quantification matrix uniquely represents a variable. For a quantitative variable usually the matrix that is the outer product moment of the standardized column of scores on the variable is used. For a qualitative variable many different quantification matrices might be used (Janson & Vegelius, 1978). All of these are transformations of the indicator matrix for the variable concerned.

Because the generalized correlation coefficients between variables that result from the quantification matrices are scalar products, the variables can be analyzed by means of PCA on the matrix of generalized correlations. This PCA will yield an optimal representation of the variables. However, in general, it does not provide a representation for the objects. For the case with only qualitative variables, Cazes, Bonnefous, Baumerder and Pages (1976) do propose a method for representing objects after a PCA has been performed on generalized correlation coefficients. This method could also be used for representing objects after a PCA on generalized correlation coefficients for a mixture of qualitative and quantitative variables. However, their method for representing objects is based on the first principal component of the variables only. In this way, the representation of the objects is based only on part of the information present in the solution for the variables, which might imply an undesirable amount of loss of information.

Above, it has been shown that MCAMIX and its variants do not provide an optimal representation for the variables, and that Saporta's method does not provide an optimal representation for the objects. In the present paper a compromise is provided that yields representations of the variables and of the objects simultaneously. It will be shown that this method yields a better representation for the variables than MCAMIX does, but this representation is not as good as the one provided by Saporta's method. On the other hand, this method provides a representation for the objects which is better than the one that might be given by Saporta's method, because these are defined for all principal components, and not only for the first. As in Kiers (in press), the compromise is given by applying a variant of INDSCAL (Carroll & Chang, 1970) to the set of quantification matrices that are defined for each of the variables. Firstly, this variant of INDSCAL will be treated. Secondly, the INDSCAL analysis for a mixture of qualitative and quantitative variables will be described. Next, it will be shown that this method is a compromise between the two existing types of methods for the analysis of a mixture of qualitative and quantitative variables.

INDSCAL is a method for the analysis of a set of objects by objects similarity matrices. Each similarity matrix may represent similarities according to one judge, or more generally based on one "aspect". The method is designed to represent the objects in one (low-)dimensional space such that the similarities between objects with respect to each aspect are approximated. Because it is assumed that for different aspects the dimensions of the space have different saliencies, the similarities are approximated by means of weighted scalar products with different sets of weights for the dimensions, for each of the different aspects.

Let the matrix of similarities between n objects pertaining to the k^{th} aspect be given by S_k ($n \times n$), let the matrix of coordinates for the objects in a p-dimensional space be denoted by X ($n \times p$, $p \leq n$), and let W_k denote the $p \times p$ diagonal matrix providing the set of weights for the k^{th} aspect, $k = 1,..,m$. Then INDSCAL is the method that finds matrices X and diagonal matrices W_k such that the loss function

$$\sigma(X,W_1,..,W_m) = \sum_{k=1}^{m} \| S_k - XW_kX' \|^2 \qquad (1)$$

is minimized over X and $W_1,..,W_m$. Although no constraints need be imposed on the matrix X, it is often useful to do so. We constrain matrix X to be a column-wise orthonormal matrix, that is $X'X = I_p$, and call the resulting method "orthogonally constrained INDSCAL". Apart from being useful in a technical way, this constraint simplifies the interpretation of the results when the method is applied to quantification matrices for qualitative variables.

As has been shown by Kiers (in press), the problem of minimizing (1) over X and diagonal matrices $W_1,..,W_m$, subject to $X'X = I_p$, is equivalent to choosing $W_k = \text{Diag}(X'S_kX)$ and maximizing

$$f(X) = \sum_{k=1}^{m} \text{tr}[\text{Diag}(X'S_kX)]^2 \qquad (2)$$

over X, subject to $X'X = I_p$. It is useful to note that $f(X)$ can equivalently be expressed as

$$f(X) = \sum_{k=1}^{m} \sum_{t=1}^{p} (x_t'S_kx_t)^2, \qquad (3)$$

where x_t denotes the t^{th} column of X.

As in Kiers (in press), the maximum of (3), subject to $X'X = I_p$, can be found by means of the algorithm proposed by Ten Berge, Knol and Kiers (1988). This completes the description of orthogonally constrained INDSCAL. In the

next section it will be discussed how orthogonally constrained INDSCAL can be applied to a set of quantification matrices for a mixture of qualitative and quantitative variables.

INDSCAL for the analysis of a mixture of qualitative and quantitative variables

Above, it has been mentioned that various quantifications can be chosen for quantifying qualitative data. In the present paper the quantification matrices that have been used by Saporta (1976) are chosen. In case the k^{th} variable is a qualitative variable then let G_k denote the $n \times m_k$ indicator matrix for the k^{th} variable, where m_k is the number of categories of variable k, let D_k be defined as the diagonal matrix of frequencies of the categories of this variable, and let the $n \times n$ matrix J be defined by $J = (I - 11'/n)$, where $\mathbf{1}$ is the vector of order n with unit elements and J is the centering operator. Then the quantification matrix chosen here is defined as

$$P_k = \frac{1}{(m_k - 1)^{1/2}} JG_kD_k^{-1}G_k'J. \tag{4}$$

For a quantitative variable the quantification matrix can be described as follows. When the k^{th} variable is quantitative, then let the column vector $\mathbf{z}_k$ contain the scores of the n objects on the quantitative variable k. The quantification matrix for variable k is then given by

$$Q_k = \frac{1}{\mathbf{z}_k'J\mathbf{z}_k} J\mathbf{z}_k\mathbf{z}_k'J. \tag{5}$$

From (4) and (5) it follows that we have $\text{tr } P_k^2 = 1$ and $\text{tr } Q_k^2 = 1$. That is, just as in ordinary PCA, the variables are normalized to unit sums of squares.

In the sequel the orthogonally constrained version of INDSCAL described above, with matrix S_k chosen as P_k when variable k is qualitative, and as Q_k when variables k is quantitative, $k = 1,..,m$, will be denoted as "INDSCAL for mixed variables".

INDSCAL for mixed variables as a compromise between Saporta's method and MCAMIX

INDSCAL for mixed variables can be interpreted in a number of different ways. It will be shown that the analysis can be interpreted as a method that optimally represents relations between variables (as a PCA technique does)

while retaining a clear link with the representation of the objects. In order to do this it is necessary to explain the idea behind Saporta's method in mathematical terms, because it supplies a good basis for interpreting INDSCAL for mixed variables.

Saporta's method consists of performing PCA on a mixture of qualitative and quantitative variables. This PCA is based on quantification matrices for the qualitative and quantitative variables. Such a quantification matrix can be seen as a vector in $\mathbf{R}^{n \times n}$. Such vectors that are in fact matrices strung out row-wise as vectors will be denoted here as tensors, because they are vectors of order n^2 representing $n \times n$ product moment matrices. The rank of such a tensor is defined as the rank of the corresponding product moment matrix. As has been mentioned above, Saporta chooses P_k as quantification matrix when the k^{th} variable is qualitative and Q_k when the k^{th} variable is quantitative.

Mathematically, Saporta's method can be described as follows. Let S_k denote the quantification matrix for the k^{th} variable, then Saporta's method consists of maximizing the function

$$g(F_1,..,F_p) = \sum_{k=1}^{m} \sum_{l=1}^{p} (\text{tr } F_l'S_k)^2, \tag{6}$$

over the $n \times n$ matrices F_l, $l = 1,..,p$, representing "factors" or "components" of the variables, subject to the constraint $\text{tr } F_l'F_{l'} = \delta_{ll'}$, where δ denotes the Kronecker symbol. It is well-known that maximizing the sum of squared loadings defines a PCA. Because $\text{tr } F_l'S_k$ can be considered as the loading of the vector representing variable k in $\mathbf{R}^{n \times n}$ on the vector representing component l in $\mathbf{R}^{n \times n}$, maximizing (6) can be seen as PCA on the variables represented by the quantification matrices.

Above, INDSCAL for mixed variables has been shown to be the method maximizing (3), with S_k chosen as P_k or Q_k, when the k^{th} variable is qualitative or quantitative, respectively. Function (3) can be rewritten as

$$f(X) = \sum_{k=1}^{m} \sum_{l=1}^{p} (\mathbf{x}_l'S_k\mathbf{x}_l)^2 = \sum_{k=1}^{m} \sum_{l=1}^{p} (\text{tr } \mathbf{x}_l\mathbf{x}_l'S_k)^2, \tag{7}$$

which has to be maximized over matrix X, subject to $X'X = I_p$. Obviously, maximizing $g(F_1,..,F_p)$ over $F_1,..,F_p$, subject to the constraint $\text{tr } F_l'F_{l'} = \delta_{ll'}$ and subject to the additional constraint $F_l = \mathbf{x}_l\mathbf{x}_l'$ is equivalent to maximizing $f(X)$ over X, subject to the constraint $\text{tr}(\mathbf{x}_l\mathbf{x}_l'\mathbf{x}_{l'}\mathbf{x}_{l'}') = \delta_{ll'}$. The latter constraint can be reformulated as $\text{tr}(\mathbf{x}_l\mathbf{x}_l'\mathbf{x}_{l'}\mathbf{x}_{l'}') = (\mathbf{x}_l'\mathbf{x}_{l'})^2 = \delta_{ll'}$. This in turn is equivalent to $X'X = I_p$, which shows that, when $F_l = \mathbf{x}_l\mathbf{x}_l'$, the constraints $\text{tr } F_l'F_{l'} = \delta_{ll'}$ and

$X'X = I_p$ are equivalent. As a consequence, maximizing $g(F_1,..,F_p)$ over $F_1,..,F_p$, subject to the constraint $\mathrm{tr}\, F_l'F_{l'} = \delta_{ll'}$, for all pairs l and l', and to the additional constraint that F_l has rank one for all l, is equivalent to maximizing (7) over X, subject to $X'X = I_p$. Hence, INDSCAL for mixed variables can be interpreted as a method of PCA on a mixture of qualitative and quantitative variables subject to the additional constraint that F_l be a rank-1-tensor.

As has been mentioned above, Saporta's method is problematic in that it does not provide coordinates for the objects. An advantage of INDSCAL for mixed variables over Saporta's method is that it does yield coordinates for the objects. The constraint that the components are rank-1-tensors, $F_l = \mathbf{x}_l\mathbf{x}_l'$, implies that for every component of the variables there is a set of coordinates on a dimension in a low-dimensional space $\mathbb{R}^p$ for the *objects*. In this way, each dimension for the variables is directly and uniquely linked to a dimension for the objects. Moreover, because the rank-1-tensors are required to be orthogonal, the object coordinate dimensions are orthogonal as well.

It is well-known that the MCAMIX solution yields object coordinates for which $X'X = I_p$ (cf. Tenenhaus, 1977). That is, the MCAMIX solution satisfies the constraints imposed on the components in INDSCAL for mixed variables. Subject to these constraints, INDSCAL for mixed variables yields the best possible representation of the variables. Therefore, INDSCAL for mixed variables yields a representation of the variables that is *better* than the one given by MCAMIX.

In conclusion, we see INDSCAL for mixed variables as a method that optimally represents relations among a mixture of qualitative and quantitative variables, and simultaneously yields a representation of objects that is linked to the representation of the variables. Clearly, in this way, INDSCAL for mixed variables is a compromise between Saporta's method and MCAMIX, in that it consists of a (constrained) PCA on the variables (like Saporta's method) and simultaneously yields coordinates for the objects (like MCAMIX does).

Interpretation of results of an INDSCAL analysis for mixed variables

Because INDSCAL for mixed variables is a compromise between Saporta's method and MCAMIX, its results partly parallel those of Saporta's method and partly parallel those of MCAMIX. That is, like in MCAMIX, INDSCAL for mixed variables provides object coordinates, collected in a matrix X. These can be interpreted in the same way as in MCAMIX, but while doing this, it should be

noted that MCAMIX and INDSCAL for mixed variables stress different aspects. That is, whereas MCAMIX stresses optimal representation of objects and categories, INDSCAL for mixed variables stresses optimal representation of variables. As a consequence, MCAMIX does not provide an optimal representation for the variables, and INDSCAL for mixed variables does not provide an optimal representation for the categories. Yet, it is possible to provide category coordinates for the INDSCAL solution, by simply computing for every category of a qualitative variable the centroid of the object coordinates of the objects that fall in the category concerned.

The results of INDSCAL for mixed variables share with the solution of Saporta's method that a representation of variables is given. This representation is provided by means of the solution for the diagonal matrices W_k. The elements of these matrices can be interpreted as the loadings of the variables on the axes that represent the variables in a tensor space. In INDSCAL for mixed variables these loadings are always non-negative. This might seem to restrict the quality of these loadings, but, considering that relations between a mixture of qualitative and quantitative variables cannot sensibly be expressed in terms of negative correlations, (Janson & Vegelius, 1982), the non-negativity of these loadings merely reflects the inappropriateness of negative correlations for such pairs of variables .

Finally, we have an overall value for evaluating the quality of the solution. To this end, we use the proportion of explained inertia of the quantification matrices S_k. This proportion is given by the maximal value of $f(X)$ (cf.(3)), divided by the total inertia. The total inertia of matrix S_k is equal to 1, hence the overall total inertia is equal to m. Therefore, the proportion of explained inertia is given by the "quality measure" for INDSCAL (QM_I) as

$$QM_I = \frac{1}{m} \sum_{k=1}^{m} \sum_{l=1}^{p} (\mathbf{x}_l' S_k \mathbf{x}_l)^2. \qquad (8)$$

In order to provide an indication of the quality of the INDSCAL solution, it is useful to compare this measure to the inertia of the quantification matrices that is explained by means of Saporta's method and MCAMIX, respectively. For Saporta's method, as in ordinary PCA, the proportion of explained inertia of the quantification matrices is given by the sum of the first p eigenvalues of the "correlation" matrix, divided by m.

In case the object coordinates are computed by means of MCAMIX, one might compute the proportion of explained inertia by means of (8) with the MCAMIX object coordinates substituted for the INDSCAL object coordinates. It

should be noted, however, that the thus computed "proportion of explained inertia" is the explained inertia of the MCAMIX object coordinates when the quantification matrices are represented by the INDSCAL model. Another interesting measure for the quality of the MCAMIX solution would be a measure that is based on the model for quantification matrices that is actually fit by MCAMIX. It can be shown that MCAMIX fits the quantification matrices to the model

$$\hat{S}_k = XWX', \tag{9}$$

for $k = 1,..,m$, where W is a diagonal matrix, and X is the orthonormal matrix of object coordinates (cf. Kiers, in press). Clearly, this model is a special case of the INDSCAL model. That is, MCAMIX fits the INDSCAL model, subject to the additional constraint that $W_k = W$, for all k, or, equivalently, subject to the additional constraint that the matrices $W_1,..,W_m$ be equal. This model is an interesting model in itself, because, when it adequately represents the quantification matrices, it implies that all variables can be represented by the same coordinates in the variable space. It can be shown that the explained inertia of this model is expressed by

$$QM_M = \sum_{k=1}^{m} \text{tr } \hat{S}_k^2 / \sum_{k=1}^{m} \text{tr } S_k^2 = \frac{1}{m^2} \sum_{l=1}^{p} \lambda_l^2, \tag{10}$$

where λ_l is the l^{th} eigenvalue of $\sum_{k=1}^{m} S_k$. Comparing QM_M to QM_I provides the user with a tool to choose between representing the variables by means of INDSCAL and representing the variables by means of the simpler model with poorer fit, MCAMIX.

Examplary analysis

In order to give an idea of what results of INDSCAL for mixed variables may look like, a simple data set has been analyzed by all three methods. That is, the data is analyzed by means of Saporta's method, a variant of MCAMIX, and INDSCAL for mixed variables. For the variant of MCAMIX the method is chosen that computes the object coordinates as the first p eigenvectors of $\sum_{k=1}^{m} S_k$, with S_k chosen as P_k or Q_k, depending on the measurement level of variable k. This differs slightly from MCAMIX because in MCAMIX S_k is chosen equal to $JG_kD_k^{-1}G_k'J$, instead of P_k, when variable k is qualitative. The choice for this variant of MCAMIX, is made for comparative purposes only.

The (artificial) data set that is analyzed here, is given by Hartigan (1975, p.228). The data consist of 24 objects like screws and nails,

that are classified according to 5 categorical variables (Whether or not they have a THread, what type of HEad they have, what INdentation they have in the heads, what kind of BOttom they have and whether or not they are made of BRass). In addition their LEngth (in half inches) is measured, which is considered as a numerical variable in the present analysis. In all analyses, the one–, two–, three–, and four–dimensional solutions were studied. For the INDSCAL solutions the percentages of explained inertia were 31.9 % for $r = 1$, 50.0 % for $r = 2$, 62.0 % for $r = 3$ and 68.5 % for $r = 4$. Because after the third dimension (with 62.0 % of explained inertia) the rate of increase in explained inertia was considerably diminished, it has been decided that the three–dimensional solution was the most useful. For $r = 3$, Saporta's method explained 72.1 % of the inertia, and the model fitted by MCAMIX explained 32.5 % of the inertia of the quantification matrices. The MCAMIX object coordinates used in the INDSCAL model explained 52.4 % of the inertia.

For reasons of space only the coordinates for the variables and the objects on the two most important dimensions resulting from the INDSCAL analysis with $r = 3$ are provided, in figure 1 and figure 2, respectively. The third dimension was dominated by the (numerical) variable LEngth. The object coordinates on this dimension mainly represent the objects along this axis according to increasing length.

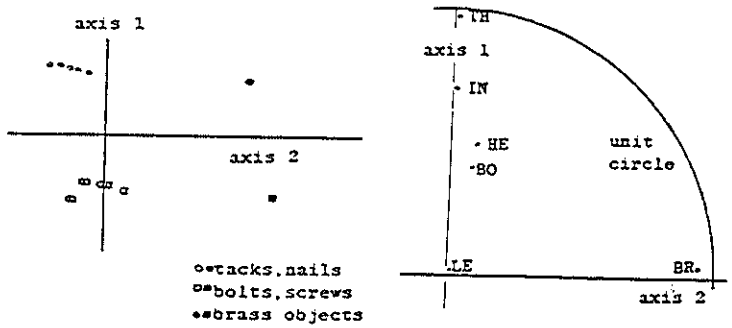

Figure 1. Object coordinates from INDSCAL for mixed variables.

Figure 2. Variable loadings from INDSCAL for mixed variables.

Clearly, Saporta's method represents the variables best, and the model for MCAMIX yields the poorest representation for the variables. The quality of the representation provided by INDSCAL lies in between these two extremes. In conclusion, when one is interested in a PCA on variables that also provides object coordinates than one should certainly prefer to represent the present examplary data set by means of the INDSCAL model to representing

it by means of MCAMIX.

The interpretation of the results is facilitated by the fact that the the axes for the objects and those for the variables are linked. From the plot for the variables it is clear that the first (vertical) axes represents quite well the variables THread, HEad, INdentation and BOttom. These variables are precisely the variables that distinguish screws and the like from nails and the like. This interpretation of the first axis corresponds well to the coordinates of the objects on this axis. That is, this axis contrasts two clusters of objects: the screws and the nails. It should be noted that this contrast had not been used as a variable explicitly in the analysis. The second axis is dominated by the variable BRass. In the objects plot it contrasts the (few) brass objects to those that are made of other material. The fact that the variables BRass and LEngth can only be represented by means of extra dimensions is in agreement with the general notion that the distinction between screws and nails has nothing to do with the material they are made of or with their length. It can be conluded that the INDSCAL analysis of these data yields well interpretable results.

Beyond INDSCAL

In the present paper, INDSCAL for a mixture of qualitative and quantitative variables has been described as a method that can be seen as a compromise between MCAMIX and Saporta's method. It is a compromise in that it yields a good representation for the variables (like Saporta's method) and at the same time it yields object coordinates (like MCAMIX). However, INDSCAL for mixed variables is not the only method that yields such a compromise. In fact, at least two other methods provide a compromise between MCAMIX and Saporta's method. These are TUCKALS-3 (Kroonenberg & De Leeuw, 1980; cf. Marchetti, 1988) and unconstrained INDSCAL, both applied to quantification matrices. Kiers (1988) has shown that TUCKALS-3, unconstrained INDSCAL and orthogonally constrained INDSCAL, together with two other three-way methods form a hierarchy, such that the matrices to which these are applied are increasingly well fitted. In fact, it readily follows from Kiers (1988) that for methods for the analysis of a mixture of qualitative and quantitative variables a similar hierarchy is possible, as follows. MCAMIX yields the poorest representation of the quantification matrices. INDSCAL for mixed variables yields a better representation for the quantification matrices, unconstrained INDSCAL yields a representation that is still better than the one given by (orthogonally constrained) INDSCAL, TUCKALS-3 yields a representation which is even better than the one given by unconstrained

INDSCAL, and Saporta's method yields the best possible representation for the quantification matrices. As explained by Kiers (1988), increase of fit is obtained at the cost of growing complexity of the model. Therefore, a priori it is never clear which method might be the most useful for representing one's data. The data analyst must decide on the basis of the fit values and interpretability of the results of the various methods, which method yields the most useful representation for the data set at hand.

Discussion

In the present paper, a particular choice has been made for the quantification matrices for the qualitative and quantitative variables. This choice served two purposes. On the one hand, it provided a basis from which a series of methods could be worked out in detail. On the other hand, the specific choice made here resulted in a series of methods of which the two extremes are methods that have been proposed before (Saporta's method and MCAMIX). However, as has been said, it is an open question whether other choices of quantification matrices might be more useful. It is conceivable that the choice of quantification matrices can only be made sensibly on the basis of the one data set at hand and the research question, which is to be answered. Further research is needed to develop a strategy for choosing quantification matrices. An important consequence of the possibility of different choices for quantification matrices is that, apart form INDSCAL on other quantification matrices, also alternatives for MCAMIX and Saporta's method can be developed. That is, alternatives of MCAMIX can be developed as methods that find object coordinates as the first p eigenvectors of the sum of the (alternative) quantification matrices. Likewise, alternatives for Saporta's methods are conceivable, and have in fact been proposed by Janson and Vegelius (1978, 1982), by representing the associations in a set of qualitative and quantitative variables by other generalized correlation coefficients than the ones chosen by Saporta (1976).

The analysis of a set of only qualitative variables can be performed equally well by the method "INDSCAL for categorical data", described by Kiers (in press), as by the methods provided in the present paper. In fact, if all variables are qualitative, choosing $S_k = P_k$, for all k, then INDSCAL for mixed variables and INDSCAL for categorical data yield exactly the same results. Hence the method developed in the present paper generalizes the method developed by Kiers (in press). On the other hand, INDSCAL for mixed variables can also be applied to a set of only quantitative variables. In that case, PCA is performed on the matrix of squared product moment

correlations between quantitative variables. Typically, this PCA will not yield the same solution as ordinary PCA on quantitative variables. The resulting object coordinates will equal those of ordinary PCA only, in case MCAMIX is used for analyzing the quantification matrices for the quantitative variables.

References

Carroll, J.D. & Chang, J.J.(1970) Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. Psychometrika, 35, 283-319.

Cazes, P., Bonnefous, S., Baumerder, A. & Pages, J.P. (1976) Description cohérente des variables qualitatives prises globalement et de leurs modalités. Statistique et Analyse des Données, 1, 48-62.

De Leeuw, J. (1973) Canonical Analysis of Categorical Data. Doctoral Dissertation, University of Leyden.

Escofier, B. (1979) Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. Cahiers de l'Analyse des Données, 4, 137-146.

Hartigan, J.A. (1975) Clustering algorithms. New York: Wiley.

Hill, M.O. & Smith, A.J.E. (1976) Principal Component Analysis of Taxonomic data with multi-state discrete characters. Taxon, 25, 249-255.

Janson, S & Vegelius, J. (1978) Correlation coefficients for more than one scale type. (Res. Report,78-2) University of Uppsala, Dept. of Statistics.

Janson, S & Vegelius, J. (1982) Correlation coefficients for more than one scale type. Multivariate Behavioral Research, 17, 271-284.

Kiers, H.A.L. (1988) Hierarchical relations between three-way methods. Paper presented at the 20th meeting of the ASU, Grenoble, May 30 - June 2.

Kiers, H.A.L. (in press) INDSCAL for the analysis of categorical data. Proceedings of the meeting "MULTIWAY '88", Rome, March 28-30.

Kroonenberg, P.M. & De Leeuw, J. (1980) Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika, 45, 69-97.

Marchetti, G.M.(1988) Three-way analysis of two-mode matrices of qualitative data. Manuscript.

Nishisato, S. (1980) Analysis of Categorical Data: Dual Scaling and its Applications. Toronto: University Press.

Saporta, G. (1976) Quelques Applications des Opérateurs d'Escoufier au traitement des variables qualitatives. Statistique et Analyse des Données, 1, 38-46.

Ten Berge, J.M.F., Knol, D.L. & Kiers, H.A.L. (1988) A treatment of the Orthomax rotation family in terms of diagonalization, and a re-examination of a singular value approach to Varimax rotation. Manuscript submitted for publication.

Tenenhaus, M. (1977) Analyse en Composantes Principales d'un ensemble de variables nominales ou numériques. Revue de Statistique Appliquée, 25, 39-56.

Tschuprow, A.A. (1939) Principles of the mathematical theory of correlation. New York: William Hodge.

Young, F.W., Takane, Y. & De Leeuw, J. (1978) The Principal Components of Mixed Measurement Level Multivariate Data: an Alternating Least Squares Method with Optimal Scaling Features. Psychometrika, 43, 279-281.

Zegers, F.E. & Ten Berge, J.M.F. (1986) Correlation coefficients for more than one scale type: an alternative to the Janson and Vegelius approach. Psychometrika, 51, 549-557.