Component and Correspondence Analysis
Edited by J.L.A. van Rijckevorsel and J. de Leeuw

Chapter 1

Relations Between Variants of Non-linear Principal Component Analysis

Paul Bekker
University of Groningen, The Netherlands
and
Jan de Leeuw
UCLA, Los Angeles, USA

INTRODUCTION

Classical principal component analysis can be generalized in different directions to yield non-linear principal component analysis. This means that, in addition to variables measured on an interval scale level, other variables such as ordinal or nominal variables can be analysed in a similar way.

The first generalization is a technique that has many names, but we shall refer to it as 'multiple correspondence analysis'. For an extensive description of multiple correspondence analysis, we refer to Nishisato (1980), Gifi (1981a, 1988), or Greenacre (1984). Another generalization, described by Kruskal and Shepard (1974) and Gifi (1981a, 1988), is 'non-metric principal component analysis'.

In the first section of this chapter a brief account is given of both techniques, which describes them as direct generalizations of principal component analysis. This account is far from being complete and is meant as an introduction only. One can introduce principal component analysis in many different ways. Also, multiple correspondence analysis and non-metric principal component analysis can be looked upon from quite another angle, without regarding them as generalizations of principal component analysis. Both techniques can be presented as multidimensional scaling techniques, using the concept of distance as the central one rather than the concept of correlation.

The principal aim of this chapter is to discuss, in detail, the relationships between multiple correspondence analysis and non-metric principal component analysis. To this end the techniques will be reformulated in terms of optimal scaling in Section 2, and, in the third section, a condition will be derived under which both techniques find essentially the same solution. In Section 4, some important theoretical examples are given for which this condition is satisfied. To conclude, an alternative algorithm for non-linear principal component analysis is applied which combines features of both previous generalizations.

For previous attempts to integrate different generalizations of principal component analysis, we refer to de Leeuw and van Rijckevorsel (1980), Gifi (1981a, 1988), de Leeuw (1982), Tenenhaus and Young (1985). The general approach to PCA in this chapter is taken from Gifi (1981a), the results we discuss are mainly due to de Leeuw (1982).

1. VARIOUS GENERALIZATIONS OF PRINCIPAL COMPONENT ANALYSIS

In this section the concept of homogeneity will serve as a basis for the introduction of multiple correspondence analysis (MCA) and non-metric principal component analysis (NCA) as direct generalizations of ordinary metric principal component analysis (PCA). Homogeneity can be defined on the basis of various models (cf. Fischer, 1974). In this chapter the concept of homogeneity will be used in a data theoretical sense as being closely related to the concept of data reduction (cf. de Leeuw, 1973). In other words, homogeneity deals with the question to what extent different variables measure the same property or properties. In order to answer this question, we need a measure for the difference or resemblance of the variables. On the other hand, the measurement level of the data may allow us to transform the variables before comparing them with each other. By definition, the class of admissible transformations will be different for different types of data. The problem is then to find admissible transformations that maximize the resemblance, or homogeneity, of the variables. When the variables measure more than one property we may want to proceed in order to find another, orthogonal, solution. This is in accordance with the principle of data reduction which advocates that a small number of dimensions should be used to explain a maximum amount of information present in the data.

For the techniques discussed in this chapter, we might say that they share this approach of finding a transformation within a class of admissible transformations such as to minimize the difference between the variables. There are two kinds of distinctions amongst the techniques: either the way the differences between the variables are measured varies, or different classes of admissible transformations are used.

1.1. Homogeneity and linear transformations

Linear transformations of variables may both change their means and their variances. To begin we shall leave the mean out of consideration (we exclude variables which are not in deviations from the mean) and only attach weights to the variables. The difference between the variables will be expressed in a loss function. We want this loss function to attain a minimum when all variables are alike, so one possibility is to use the mean squared Euclidean distance between the transformed variables and one hypothetical common variable (or vector). The resemblance of the transformed variables and the hypothetical variable will be maximized if the loss function is minimized. In that case, the hypothetical variable can serve as a scale for the individuals or objects. The values of this variable are called object scores, component scores or—analogous to factor analysis—factor scores.

The loss function reads:

$$\sigma(x; \phi) = m^{-1} \sum_j SSQ(x - \phi_j(h_j)), \tag{1.1}$$

where x is the common variable and ϕ is the transformation of variable $h_j (j = 1, \ldots, m)$. In the present case of linear weighting, we can write $\phi_j(h_j) = h_j y_j$, where $y_j (j = 1, \ldots, m)$ are the weights for the variables.

We can reformulate the loss function as

$$\sigma(x; y) = x'x + m^{-1} y'Dy - 2m^{-1} x'Hy, \tag{1.2}$$

where H is the data matrix of order $(n \times m)$, D the diagonal of $H'H$ and y the vector of weights. When minimizing this function, we have to impose a restriction on the length of x or y in order to avoid the trivial minimum where both x and y contain zeros.

A possible restriction is $x'x = 1$, which normalizes the object scores. Another restriction where the normalization focuses on the transformed variables is given by $y'Dy = 1$. Both approaches given essentially the same results. Therefore we shall only work out the first normalization.

The minimum of $\sigma(x; y)$, subject to the restriction $x'x = 1$, can be found by minimizing the function $f(x, y, \lambda) = \sigma(x; y) - \lambda(x'x - 1)$, where λ is a Lagrange multiplier. The stationary values can be found in the usual way by equating the partial derivatives to zero.

We find

$$H'x = Dy,$$

$$Hy = m(1 - \lambda)x.$$

Hence $\sigma(x; y) = \lambda$, and the loss function will reach a minimum for the smallest possible value of λ. This λ can be found by combining the equations to

$$HD^{-1}H'x = m(1 - \lambda)x.$$

Clearly, the object scores form a latent vector of $HD^{-1}H'$ and $m(1-\lambda)$ is the latent root. The loss function will thus be minimized if x is the latent vector that corresponds to the largest latent root of $HD^{-1}H'$. If we define the singular value decomposition (SVD-solution, cf. van de Geer, 1986)

$$HD^{-1/2} = V\psi W',$$

where V is of order $(n \times m)$, W is of order $(m \times m)$ and $V'V = I$, $W'W = I$, and ψ is a diagonal matrix of order $(m \times m)$, then we can find the latent roots and vectors in the matrix

$$HD^{-1}H' = V\psi^2 V'.$$

If the latent roots are arranged in descending order of magnitude, the solution is given by $x = v_1$ and $m(1-\lambda) = \psi_1^2$:

$$HD^{-1}H'x = x\psi_1^2. \tag{1.3}$$

At the same time,

$$H'Hy = Dy\psi_1^2. \tag{1.4}$$

Hence,

$$y = D^{-1/2}w_1\psi_1,$$

and

$$y = D^{-1}H'x, \quad \text{or} \quad x\psi_1^2 = Hy. \tag{1.5}$$

Which determines completely the solution of the minimization of the loss function.

For variables measuring more than one property, we may want to find another scale x_2 for the individuals, orthogonal to the first scale x_1. This means minimizing $\sigma(x_2; y_2)$ subject to the restriction $x_1'x_2 = 0$. For a third solution we have to impose $x_1'x_3 = 0$ and $x_2'x_3 = 0$, etc.

The successive solutions can now easily be found by the same decomposition as we used for the first solution: $HD^{-1/2} = V\psi W'$, where V is orthogonal, $V'V = I$. We simply use successive singular vectors and values for successive minimizations. The total number of successive solutions equals m, the number of variables. If we collect the vectors x_i and $y_i (i = 1, \ldots, m)$ as columns in the matrices X and Y respectively, the simultaneous solution is given by:

$$HD^{-1/2} = XY'D^{1/2}. \tag{1.6}$$

Apparently, the original and transformed (i.e. weighted) variables are linear combinations of the column vectors of X. This means that the vectors $x_i (i = 1, \ldots, m)$ form an orthonormal basis of the vector space spanned by the original, or transformed, variables. The sum of the latent roots equals m, for:

$$SSQ(HD^{-1/2}) = \mathrm{tr}(HD^{-1}H') = \mathrm{tr}(XY'DYX')$$
$$= \mathrm{tr}(Y'DY) = \mathrm{tr}(\psi^2) = \sum_i \psi_i^2 .$$

In the foregoing we did not concern ourselves with the mean of the variables. In order for the transformations to be proper linear transformations, we have to take into account the means. This can be done by decomposing the variables into mean vectors and vectors in deviation from the mean. If we collect these two vectors, as columns, in a matrix F_j for every variable separately, so that $F_j u = h_j$, a linear transformation is given by $F_j z_j$, where z_j is a vector of two weights. The first weight of z_j transforms the mean and the second weight transforms the variable in deviation from the mean. The complete matrix $F = (F_1, \ldots, F_m)$ is of order $(n \times 2m)$ and the vector z is of order $(1 \times 2m)$. Clearly the column vectors of F_j are othogonal, i.e. $F_j'F_j$ is a diagonal matrix. If we now define D to be the diagonal matrix of $F'F$, we can rewrite the loss function (1.1) as:

$$\sigma(x; z) = x'x + m^{-1} \sum_j z_j' F_j' F_j z_j - 2m^{-1} \sum_j x' F_j z_j ,$$

or

$$\sigma(x; z) = x'x + m^{-1} z'Dz - 2m^{-1} x'Fz .$$

Because of the obvious resemblance of this expression to the one in (1.2), the minimization, subject to $x'x = 1$, is analogous to the one we already derived for (1.2).

The singular value decomposition $FD^{-1/2} = V\Psi W'$, and the analogous formulations of (1.3), (1.4) and (1.5), will now render $2m$ solutions, since F is of order $(n \times 2m)$. However, if we look more closely at the matrix $FD^{-1/2}$, we see that all columns related to mean vectors are identical to $n^{-1/2}u$ (u being a vector with unit elements only). This totals up to m columns. Consequently we shall find m trivial solutions resulting from transformations of the mean vectors only. It is easy to see that one trivial latent root equals m and $m-1$ trivial latent roots equal 0. For the non-trivial solutions only the variables in deviations from the mean are weighted. Obviously the object scores x_i will be in deviations from the mean too.

Thus, the minimization of the loss function for linear transformations renders m meaningful solutions in deviations from the mean. These solutions correspond to the ones we would have found if we had started with variables in deviations from the mean, and simple weighting would have generated them.

1.2. Principal component analysis: PCA

If, in the foregoing section, we had started with variables in deviations from the mean and normalized to unit length, so that $D = I$ (the identity matrix) and $H'H = R$ (the correlation matrix), then, according to (1.4), (1.5) and (1.6)

$$RY = Y\Psi^2, \quad H = XY', \quad R = YY', \quad X'X = I \quad \text{and} \quad Y'Y = \Psi^2 .$$

These formulae relate the latent root and latent vector solutions of the correlation matrix to the principal components X of the data matrix H. In fact, these formulae are well known expressions in principal component analysis (cf. van de Geer, 1986).

1.3. Homogeneity and non-linear transformations: MCA

As we have seen, PCA can be presented as a technique for minimizing differences among variables by transforming these variables linearly. A quite straightforward generalization of PCA in this context is given by extending the class of admissible transformations to include non-linear transformations as well. If we confine ourselves to categorical data, a non-linear transformation is simply found by weighting the categories of a variable. The differences amongst transformed variables can be measured by the same loss function, (1.1), as was used in the foregoing sections to introduce PCA. Minimization of this loss function, for non-linear transformations of discrete variables, adds up to a technique we refer to as multiple correspondence analysis.

If we use so-called indicator matrices, the transformed variables can easily be expressed in matrix notation. An indicator matrix is a binary matrix which indicates the category that an observation is in. Thus, if h_j has k_j categories, the indicator matrix G_j is $n \times k_j$. For the transformation of variable h_j we have,

$$\phi_j(h_j) = G_j y_j,$$

where G_j is the indicator matrix of variable h_j, and y_j is a vector comprising k_j weights for the k_j categories of variable h_j. MCA therefore can be presented as a technique which minimizes the following loss function

$$\sigma(x; y)) = m^{-1} \sum_j SSQ(x - G_j y_j), \tag{1.7}$$

or

$$\sigma(x; y) = x'x + m^{-1} \sum_j y_j' G_j' G_j y_j - 2m^{-1} \sum_j x' G_j y_j.$$

As the column vectors of G_j are orthogonal, the matrix $D_j = G_j' G_j$ is diagonal and hence we can write

$$\sigma(x; y) = x'x + m^{-1} y' D y - 2m^{-1} x' \mathrm{G} y, \tag{1.8}$$

where D is the diagonal supermatrix of univariate marginals and G is the indicator supermatrix. The resemblance with (1.2) is obvious. The minimization is completely analogous: instead of the data matrix H we simply use the indicator supermatrix G.

All solutions will now be rendered by the singular value decomposition of

$$GD^{-1/2} = V \Psi W'.$$

If we write $G'G=C$ for the matrix of bimarginals we have, analogous to (1.3), (1.4) and (1.5)

$$X=V, \quad \text{and} \quad Y=D^{-1/2}W\psi, \tag{1.9}$$

$$GD^{-1}G'X=X\Psi^2, \quad \text{and} \quad CY=DY\Psi^2, \tag{1.10}$$

$$X\Psi^2=GY, \quad \text{and} \quad Y=D^{-1}G'X. \tag{1.11}$$

Since G is of order $(n\times\sum k_j)$, we find $\sum k_j$ solutions. Although some of these, the so-called trivial solutions, are completely meaningless, their existence is, as we saw in Section 1.1, a pleasant circumstance. If we look at the indicator matrices more closely, we see that they comprise a certain amount of redundant information. For, when all vectors of G_j but one, are known, then this one vector is also fixed (we assume that there are no missing data). As a result of these m redundant vectors in G, we shall find in our analysis m trivial solutions. The most prominent trivial solution is given by the vector y_0, where all $\sum k_j$ weights equal $n^{-1/2}$, and the associated object scores in x_0, which also equal $n^{-1/2}$. It is easy to see that this pair x_0, y_0 is a solution; and in fact the loss function reaches its absolute minimum so that there is no loss at all. The corresponding latent root equals m. The $m-1$ remaining trivial solutions are found for weights which are the same for the categories within a variable, but which vary across variables, so that, for all remaining trivial solutions, we have $y'Cy=0$. The corresponding latent roots all equal zero. This situation resembles the situation where we are maximizing homogeneity by means of linear weighting, while the data matrix H comprises units only. As a result of the existence of these trivial solutions, all non-trivial, meaningful solutions are in deviations from the mean. On the one hand we have $X'X=I$, consequently for dimensions we have, $x_o'x_s=0$, or $u'x_s=0$, and so the object scores are in deviations from the mean. On the other hand, we have transformed variables G_jy_j in deviations from the mean.

In order to prove this, we define u_j $(j=1,\ldots,m)$ as vectors comprising k_j units, and $U=u_1\,\#\ldots\#\,u_m$, the direct sum of these vectors. As a result, the m rows of the matrix $U'C$ are all identical to the row vector $u'D$. Consequently we have $U'Cy_s=0$, because $y_0'Dy_s=0$. According to (1.10) we may also write $U'Cy_s=U'Dy_s\psi_s^2$. Then, if $\psi_s^2\neq 0$, this means that $U'Dy_s=0$. So we have for every variable separately $u'D_jy_{js}=0$, or $u'G_j'G_jy_{js}=0$, whence $u'G_jy_{js}=0$, which is a reflection of the fact that the transformed variables for non-trivial solutions are in deviations from the mean. Ultimately we have $\sum k_j-m$ meaningful solutions in deviations from the mean; the inner products of the solutions x_s and G_jy_{js} can now be interpreted in terms of variances, covariances or correlations. For the sum total of the non-trivial latent roots, we find: $\sum\psi_j^2=\sum k_j-m$.

1.3.1. *A relation with* χ^2

We may rewrite the latent vector solutions $D^{-1/2}CD^{-1/2}=W\Psi^2W'$ as a summation of matrices of rank one:

$$D^{-1/2}CD^{-1/2}=\sum_s w_s\psi_s^2w_s'.$$

From this summation $(s=1,\ldots,\sum k_j)$ we may remove the trivial matrices for which $\psi_s^2=0$ For the trivial solution for which $\psi_0^2=m$, we have

$$w_0=D^{1/2}y_0\psi_0^{-1}=D^{1/2}u(n\,.\,m)^{-1/2},$$

so that

$$w_0\psi^2w_0'=D^{1/2}uu'D^{1/2}n^{-1}.$$

If we now remove all trivial solutions from W and Ψ, we could write

$$D^{-1/2}(C-Duu'Dn^{-1})D^{-1/2}=W\Psi^2W'. \tag{1.12}$$

For a non-diagonal submatrix of (1.12) we have

$$D_i^{-1/2}(C_{ik}-D_iuu'D_kn^{-1})D_k^{-1/2}. \tag{1.13}$$

C_{ik} being the contingency table of the variables i and k. The matrix $D_iuu'D_kn^{-1}$ contains the expected frequencies on the hypothesis of independence, based on the univariate marginals of C_{ik}. If we multiply (1.13) by the scalar $n^{1/2}$, then its elements equal the difference between the observed and expected frequencies, divided by the root of the expected frequency. This implies that the sum of squares of a non-diagonal submatrix of (1.13) equals $\chi_{ik}^2n^{-1}$ the Pearson χ^2-statistic divided by n.

For diagonal submatrices we have

$$D_i^{-1/2}(D_i-D_iuu'D_in^{-1})D_i^{-1/2}=I-D_i^{1/2}uu'D_i^{1/2}n^{-1}.$$

The sum of squares of this idempotent matrix equals its trace: k_j-1.

The total sum of squares of (1.12) is therefore

$$\sum_i(k_i-i)+\sum\sum_{i\neq k}\chi_{ik}^2n^{-1}=SSQ(W\Psi^2W')=\sum_j\psi_j^4.$$

Since $\sum_j\psi_j^2=\sum_j k_j-m$, we may now write

$$\sum\sum_{i\neq k}\chi_{ik}^2=n\sum_j(\psi_j^4-\psi_j^2).$$

As $\sum_j 1=\sum_j k_j-m$, we also have

$$\sum\sum_{i\neq k}\chi_{ik}^2=n\sum_j(\psi_j^4-2\psi_j^2+1)=n\sum_j(\psi_j^2-1)^2. \tag{1.14}$$

In case of independently distributed variables, the statistic

$$\sum_{i<k} \chi_{ik}^2 = \tfrac{1}{2}n \sum_j (\psi_j^2 - 1),$$

converges to a χ^2-distribution with $\mathrm{df} = \frac{1}{2}\{\left(\sum_j k_j - m\right)^2 - \sum_j (k_j - 1)^2\}$ (cf. de Leeuw, 1973).

1.3.2. *The geometry of MCA*

As was the case with PCA, the column vectors of X form an orthonormal basis of a vector space, in which all original variables, the column vectors G, and transformed variables, $G_j y_{js}$, are contained. For,

$$G = XY'D, \quad \text{and} \quad G_j y_{js} = XY_j' D_j y_{js}, \tag{1.15}$$

where Y_j is a matrix of order $(k_j \times \sum k_j)$, with y_{js} $(s = 1, \ldots, \sum k_j)$ as columns. Since the transformed variables $G_j y_{js}$ are in deviations from the mean for non-trivial solutions, the trivial vectors of X and Y_j can be removed without causing any trouble. The transformed variables can thus be represented by vectors in a vector space, of which the x_s $(S = 1, \ldots, (\sum k_j - m))$ form an orthonormal basis.

The squared norms of the transformed variables are usually called discrimination measures: $y_{js}' D_j y_{js}$; the norm of a transformed variable $G_j y_{js}$ is larger as the discrimination between the categories, according to the quantifications y_{js}, is larger.

The sum total of the discrimination measures of all variables, for one solution s, equals $y_s' D y_s = \psi_s^2$. So, for every solution s, the sum total of the discrimination measures is maximized. At the same time the discrimination measure equals the squared correlation between x_s, and $G_j y_{js}$. Namely,

$$y_s = D^{-1} G' x_s, \quad \text{and} \quad y_{js} = D_j^{-1} G_j' x_s, \tag{1.16}$$

from which we may infer that the projection of x_s on the subspace spanned by the column vectors of G_j, is identical to the transformed variable:

$$G_j D_j^{-1} G_j' x_s = G_j y_{js}.$$

As x_s is normalized such that $x_s' x_s = 1$, we have the result that the squared correlation between x_s and $G_j y_{js}$ equals the squared norm of $G_j y_{js}$, which is the same as the discrimination measure.

This derivation indicates that we could interpret MCA as follows. Find, in the space spanned by the column vectors of G, a vector x_1 for which the sum of the squared norms of the projections on the m subspaces G_j is maximized. Having found such a vector x, find another one, subject to the restriction $x_2' x_1 = 0$, etc. The projections of the trivial solution x_0 on the various subspaces G_j all equal x_0

itself; x_0 is contained in the intersection of the subspaces spanned by $G_j (j = 1, \ldots, m)$.

Another description is the following. As we have $y_s = D^{-1} G' x_s$, this means that the object scores, corresponding to a certain category of a variable, have a centre of gravity that coincides with the quantification of that category; these points are usually called barycentra. If we replace the object scores by the barycentra of a given variable, the dispersion of the points will be smaller than before replacement. This reduction in dispersion is due to the fact that the dispersion of the object scores around their barycentra has not been taken into account. In this context, the discrimination measure of a variable gives the percentage of dispersion explained by the barycentra. So, for the first dimension, the dispersion of object scores around their centres of gravity is minimized for all variables simultaneously. The second dimension is the best orthogonal dimension, etc.

We can also conceive of the category scores as vectors. For we can write $GD^{-1} = XY'$, and the projections of these vectors on the space spanned by x_1 and x_2, can be completely represented by the category quantifications of these dimensions. Bearing in mind that all column vectors of $GD^{-1/2}$ have norms equal to unity, it is evident that, considering the norms of the column vectors of GD^{-1}, categories with low frequencies are represented by vectors with large norms. Since the projections on the trivial dimension x_0, all have norms equal to one: $u'GD^{-1} = u'DD^{-1} = u'$, these differences in norms will be present in the nontrivial space as well. As there is no reason why those differences in norms should not be present in the first two or more dimensions, we generally expect categories with low frequencies to have extreme positions, when represented as points in two (or more) dimensional space.

Because $GD^{-1} = XY'$, the norms of the projections of the object scores, the rows of X, on the category scores, the rows of Y, have to be proportional to the elements in GD^{-1}, or G. Since the mean is not mapped into non-trivial space, we expect objects corresponding to a certain category of a variable to have a position in the direction of the related category, while others have a position in the opposite direction. Obviously this will be approximately true for two dimensions. But, as we have seen, category scores are in the centre of gravity of object scores and this relates to another possible introduction, or interpretation, of MCA.

1.3.3. *The method of reciprocal averages*

This method begins with the notion that the category quantifications and the object scores in a way should be proportional to one another. For example, the objects should be located in the centre of the categories to which they correspond. This means that $x = Gy/m$. Or, conversely, the situation we already had at hand, the categories should be located in the centre of the objects: $y = D^{-1} G' x$. For non-trivial solutions these two requirements are inconsistent, and so we only require the following proportionalities:

$$x \div Gy/m, \quad \text{and} \quad y \div D^{-1}G'x. \tag{1.17}$$

Consequently we must have,

$$x \div GD^{-1}G'x.$$

From which it is evident that x should be a latent vector of $GD^{-1}G'$. This means that we are dealing with MCA.

Conditional to the normalizations $x'x=1$, or $y'Dy=1$, we find $y=D^{-1}G'x$, and $x\psi^2/m=Gy/m$, or, respectively, $x=Gy/m$, and $y\psi^2/m=D^{-1}G'x$.

1.3.4. *PCA revisited*

As we have seen MCA and PCA are closely related. Beginning with a similar loss function, the presentation of both techniques can be analogous. Partly based on this analogy we can relate PCA and MCA in another way. Every MCA solution generates m transformed variables $G_j y_{js} (j=1, \ldots, m)$. It must therefore be possible to apply PCA to the correlation matrix of these transformed variables. In addition we can apply PCA to every non-trivial MCA solution. In doing so we obtain $\sum k_j - m$ correlation matrices, each with m PCA solutions. Thus we find no less than m $(\sum k_j - m)$ different solutions. Albert Gifi (1981a) refers to this phenomenon as 'data production' as opposed to 'data reduction'

It would make sense to investigate whether or not redundant information is present in these correlation matrices. Although this topic will be discussed extensively in the following sections, we can already give a relation between the MCA and related PCA solutions.

Using the loss function (1.1) it is easy to see that the first non-trivial MCA solution (i.e. the solution having the largest latent root) equals the first PCA solutions of the associated correlation matrix. The transformed variables of the first MCA solutions, $G_j y_{j1}$ $(j=1, \ldots, m)$, all are in deviations from the means and have norms equal to the root of their discrimination measures, $v_{j1}^{1/2}$ say. The loss function

$$\sigma(x_1; y_1) = m^{-1} \sum_j SSQ(x_1 - G_j y_{j1}),$$

has a non-trivial minimum. PCA applied to these transformed variables means minimizing the loss function,

$$\sigma(x; a) = m^{-1} \sum_j SSQ(x - G_j y_{j1} v_{j1}^{-1/2} a_j). \tag{1.18}$$

Clearly then, a minimum will be attained for $x=x_1$ and $a_j=v_{j1}^{1/2}$, and we have a solution identical to the one found for MCA. We may say that the first MCA-solution is found for a transformation which maximizes the largest latent root of the associated correlation matrix. Also the discrimination measures of the first MCA-solution are in fact, using PCA terminology, the squared component loadings of the first component. Of course, the same unambiguous relation holds

for the MCA solution with the smallest latent root and the smallest latent root of the corresponding correlation matrix.

For the intermediate MCA solutions things are more complicated. Subsequent MCA solutions are found subject to the restrictions of orthogonality. Applying PCA to the correlation matrix of transformed variables then means substituting the transformations and discrimination measures into (1.18), and any x that minimizes (1.18) will do, i.e. restrictions of orthogonality no longer exist.

Nevertheless, the intermediate MCA solutions do correspond to a principal component of the associated correlation matrix. This component, however, does not necessarily correspond to the largest or smallest latent root. In order to prove this we let v_s denote the vector of discrimination measures and let V_s denote the diagonal matrix comprising the discrimination measures in its diagonal, and $Q_s = \{G_j y_{js}\}$ the matrix of transformed variables. From Section 1.3.2, we know that the correlation between transformed variables and object scores equals the root of the discrimination measures. Then clearly

$$V_s^{-1/2} Q_s' x_s = v_s^{1/2}.$$

Hence,

$$\begin{aligned} R_s v_s^{1/2} &= R_s V_s^{1/2} u = V_s^{-1/2} Q_s' Q_s V_s^{-1/2} V_s^{1/2} u = V_s^{-1/2} Q_s' G y_s \\ &= V_s^{-1/2} Q_s' x_s \psi_s^2 = v_s^{1/2} \psi_s^2. \end{aligned}$$

Obviously $v_s^{1/2}$ is a latent vector of the corresponding correlation matrix; and so $a_j = v_{js}^{1/2}$ generates one of the successive solutions of (1.18).

1.4. Non-metric principal component analysis: NCA

As we have shown, MCA can be regarded as a direct generalization of PCA. Both techniques can be described in terms of a similar loss function, while MCA has an extended class of admissible transformations.

NCA will now be presented as a technique for which the same class of admissible transformations is used as was used for MCA, but which differs from the latter because it starts off with a different loss function.

Nevertheless NCA can still be regarded as a direct generalization of PCA because both loss functions produce the same results in case of linear weighting.

1.4.1. *PCA based on another loss function*

PCA was introduced by using loss function (1.1). The difference among the transformed variables was measured by the mean squared distance to one hypothetical variable. Of course, we could as well base our loss function on the distance of the transformed variables to a plane, or more generally, to a p-dimensional vector space. By doing so, the objects or individuals can be characterized by object scores in more dimensions, although the variables are transformed only once. We can depict the p-dimensional space by a matrix of

basis vectors X of order $(n \times p)$, with $X'X = I$. A vector in this space is given by Xa, where a is a vector of p weights, or coordinates. A possible loss function is

$$\sigma(X, a, \phi) = m^{-1} \sum_j SSQ(Xa_j - \phi_j(h_j)). \tag{1.19}$$

For PCA we have $\phi_j(h_j) = h_j y_j$. In order to avoid meaningless solutions, we have to normalize again. The normalization $X'X = I$ is not enough, for both a_j and y_j are not normalized. A possible normalization is to have $y_j' h_j' h_j y_j = 1$; this normalization determines the vectors $h_j y_j$ completely in the matrix $HD^{-1/2}$. If we collect the vectors a_j as rows in the $(m \times p)$ matrix A, we could rewrite (1.19) as

$$\sigma(X, A) = m^{-1} SSQ(HD^{-1/2} - XA'). \tag{1.20}$$

How to minimize this function is known from the Eckart and Young (1936) theorem. If we define the singular value decomposition $HD^{-1/2} = V \Psi W'$, the loss function is minimal for

$$XA' = \sum_i v_i \psi_i w_i'. \tag{1.21}$$

This simultaneous solution for more dimensions is similar to the successive solution described before. This means that the simultaneous solutions for different p are nested; that is to say that the solution for $p = k$ corresponds to the first k principal components found for $p \geq k$.

1.4.2. *Principal components and non-linear weighting*

For NCA the transformed variables are given by $G_j y_j$, just as they were for MCA. Now the loss function (1.19) reads

$$\sigma(X, a, y) = m^{-1} \sum_j SSQ(Xa_j - G_j y_j). \tag{1.22}$$

Again we use the normalizations $X'X = I$ and $y_j' D_j y_j = 1$, the transformed variables are normalized to unit length. However, this does not determine the transformed variables $G_j y_j$. The $G_j y_j$ are to be found so as to maximize the sum of the p largest latent roots of the corresponding correlation matrix. For PCA, or linear weighting, we found the same correlation matrix for every 'transformation'. Now they are different. This also means that the correlation matrix found for $p = k$ will usually be different from the one found for $p \neq k$. Consequently, the solutions are no longer nested. We may observe that NCA for $p = 1$, corresponds to the first MCA solution. However, in general it is not true that NCA with $p = k$ corresponds to the first k MCA solutions. We shall come back to this point in the following sections.

For the minimization of (1.22) we have to use an iterative procedure. For example:

(1) Take an arbitrary vector y for which $y_j' D_j y_j = 1$, and $u' G_j y_j = 0$.

(2) For fixed Gy we can apply principal component analysis of which only the first p solutions will be used.
(3) Subsequently the vectors Xa_j are projected onto the subspaces G_j. The projections are $G_jz_j = G_jD_j^{-1}G_j'Xa_j$. Every vector in G_j will be orthogonal to the vector $G_jz_j - Xa_j$, and so must be $G_jy_j - G_jz_j$. Since $G_jy_j - Xa_j = (G_jz_j - Xa_j) + (G_jy_j - G_jz_j)$, we have

$$SSQ(G_jy_j - Xa_j) = SSQ(G_jz_j - Xa_j) + (y_j - z_j)'D_j(y_j - z_j).$$

The first term on the right-hand side is fixed and so we only have to minimize the second term. This can be done simply by setting y equal to z. In the case of ordinal data we can carry out a monotonic regression.
(4) Return to the second iteration step or stop.

Loss function (1.22) is used by Kruskal and Shepard (1974), Tenenhaus (1977) and Young, de Leeuw and Takane (1976). De Leeuw and van Rijckevorsel (1980) and Gifi (1982) use another loss function, for two reasons: (i) the treatment of missing data becomes more simple, (ii) both variables with a single quantification or transformation, and variables with multiple quantifications can be analysed simultaneously: a combination of NCA and MCA.

This alternative loss function is given by:

$$\sigma_{\text{alt}}(X, a, y) = m^{-1} \sum_j SSQ(G_jy_ja_j' - X). \tag{1.23}$$

In this function the rank-one matrix y_ja_j' can be replaced for certain variables by a matrix Y_j, not necessarily of rank one, in order to treat variables with multiple quantifications.

Both loss functions (1.22) and (1.23) give the same results.

$$\sigma(X, a, y) = m^{-1} \sum_j a_j'a_j + 1 - 2m^{-1} \sum_j a_j'X'G_jy_j$$

$$\sigma_{\text{alt}}(X, a, y) = m^{-1} \sum_j \text{tr}(a_ja_j') + p - 2m^{-1} \sum_j \text{tr}(X'G_jy_ja_j')$$

$$= m^{-1} \sum_j a_j'a_j + p - 2m^{-1} \sum_j a_j'X'G_jy_j.$$

So that $\sigma_{\text{alt}} = \sigma + (p-1)$. Thus it is evident that both loss functions attain minima for the same X, A, Y (for missing data the solutions will usually not be the same). Again we use an interative procedure

(1) Take a matrix X of order $(n \times p)$, with $X'X = I$ and $u'X = 0$.
(2) Project the column vectors of X onto the subspaces G_j. The projections are $G_jZ_j = G_jD_j^{-1}G_j'X$, and

$$SSQ(G_jy_ja_j' - X) = SSQ(G_jZ_j - X) + SSQ(G_jy_ja_j' - G_jZ_j)$$

For multiple quantifications, the second term on the right-hand side can be made equal to zero by taking $Y_j = Z_j$. For nominal data the solution for y_j and a_j' can be found by taking the dominant singular value solution of $D_j^{1/2} Z_j$. This solution can also be found by an iterative procedure, which can also be used to treat ordinal data: the so-called inner iterations.

(2a) Take a vector y_j with $y_j' D_j y_j = 1$.
(2b) Now project $G_j Z_j$ on $G_j y_j$. This gives b_j. Thus

$$G_j y_j b_j' = G_j y_j y_j' G_j G_j Z_j, \text{ and}$$

$$SSQ(G_j y_j a_j' - G_j Z_j) = SSQ(G_j y_j b_j' - G_j Z_j) + SSQ(a_j - b_j);$$

set $a_j = b_j$
(2c) Project the rows of Z_j onto a_j: $a_j z_j' = (a_j' a_j)^{-1} a_j a_j' Z_j'$,

and

$$SSQ(G_j y_j a_j' - G_j Z_j) = SSQ(G_j z_j a_j' - G_j Z_j) + a_j' a_j SSQ(G_j(y_j - z_j))$$

The minimum of the second term on the right-hand side can be found, depending on the type of data, by linear regression, monotonic regression, or simply by setting $y_j = z_j$. After normalization of $G_j y_j$, we can return to (2b) or go on to step 3. Before doing so, we may observe that $G_j z_j$ is in deviations from the mean:

$$\begin{aligned} u' G_j z_j &= u' D_j Z_j a_j (a_j' a_j)^{-1} = u' D_j D_j^{-1} G_j' X a_j (a_j' a_j)^{-1} \\ &= u' X a_j (a_j' a_j)^{-1} = 0. \end{aligned}$$

(3) For fixed $G_j y_j a_j'$ we now have to find a new space X. If we let $Q_j = G_j y_j a_j'$, then we have to minimize the loss function

$$\sigma(X) = m^{-1} \sum_j (Q_j - X) \tag{1.24}$$

where $X'X = I$. This is a Procrustus problem (Cliff, 1966). Minimizing (1.24) is the same as maximizing,

$$\sum_j \operatorname{tr}(X' Q_j) = \operatorname{tr}(X' \sum_j Q_j) = \operatorname{tr}(X' Q).$$

where $Q = \sum_j Q_j$. If we define the SVD-solution $Q = K \Lambda L'$ then

$$\operatorname{tr}(X' Q) = \operatorname{tr}(X' K \Lambda L') = \sum_i (l_i' X' k_i) \lambda_i.$$

As $k_i' X l_i \leq 1$, the above expression reaches a maximum for $k_i X l_i = 1$. This means $X = KL'$. As the vectors of Q_j, and thus the vectors of Q, are in deviations

from the mean, also X is in deviations from the mean. We can now return to the second iteration step.

1.4.3. *The geometry of NCA*

For every variable we have p component loadings and the transformed variables can be mapped into a subspace or 'X-plane' as vectors Xa_j. As was the case for PCA, the squared component loadings $(a_{js})^2 (s=1, \ldots, p)$ can be interpreted as explained variances per variable. Usually these quantities are referred to as measures of 'single fit'. Analogous to MCA, the barycentra $Z_j = D_j^{-1} G_j' X$ can be mapped into the 'X-plane' by projecting the vectors $G_j D_j^{-1}$ onto the 'X-plane': $XZ_j' = XX'G_jD_j^{-1}$. And, also analogous to MCA, usually categories with low frequencies will have extreme positions, when they are represented by Z_j. Where MCA uses the term 'discrimination measure' for indicating how close the object scores are located around their barycentra, NCA uses 'multiple fit'.

As opposed to MCA the multiple fit is not maximized for all variables per dimension. It just indicates the dispersion of the categories per dimension.

However, for nominal data (with single quantifications) we do have a maximum dispersion of the categories in the direction of Xa_j. So we first have to project the object scores onto Xa_j, instead of x_1 and x_2 as was the case for MCA. The corresponding measure is the single fit $(a_{js})^2$ summed over all dimensions. For all variables together:

$$\sum_i \psi_i^2 (\psi_i^2: \text{latent roots of the correlation matrix: } i=1, \ldots, p).$$

For ordinal data also the order of the categories has to be correct, so that the single fit might be smaller than a measure of dispersion of the barycentra (when projected on Xa_j) would indicate.

2. A REFORMULATION

In this section we reformulate multiple correspondence analysis and non-metric principal component analysis in terms of optimal scaling.

2.1. Optimal scaling

Linear multivariate analysis can be generalized in several directions to non-linear multivariate analysis. An important generalization is given by non-linear multivariate analysis using optimal scaling, where optimality is defined in terms of the correlation matrix of scaled (or quantified or transformed) variables. We could describe this generalization as follows. Suppose we are free to choose m elements (vectors or random variables) q_j from m subsets L_j of a linear space L. For every choice $(q_1, \ldots, q_m)$, $q_j \in L_j (j=1, \ldots, m)$, we can compute a correlation matrix $R(q_1, \ldots, q_m)$ (we do not consider those $(q_1, \ldots, q_m)$ for which this

computation is impossible). We also have an objective function μ, defined on the set of all possible correlation matrices. Our non-linear multivariate analysis technique then consists of choosing the $q_j \in L_j$ in such a way that the function $\mu(R(q_1, \ldots, q_m))$ is maximized. More generally, we could say that we are interested in computing some, or all, of the stationary points of $\mu(R(q_1, \ldots, q_m))$ on $L_1 \times \ldots \times L_m$.

In defining our technique this way, we must consider two aspects: the form of the subspaces L_j and the nature of the objective function. Different choices in respect of one or both aspects will usually result in different techniques of analysis. The question is whether different choices would, in special situations, yield the same results. In the following sections we hope to give an answer for this question, in particular for two different techniques: MCA and NCA. Before presenting these two techniques in terms of optimal scaling, we first give some examples of possible subspaces L_j and objective functions μ.

Optimal scaling may be used to compute quantifications for observations that are missing. In this case the subspaces L_j are formed by elements with values that equal the observed values but which are arbitrary for observations that are missing. Using these subspaces, we can find optimal quantifications for the missing data (cf. Wold and Lyttkens, 1969).

Another application is given by the analysis of ordinal data. In this case the subspaces L_j are convex cones: if $q_j, p_j \in L_j$, and $\alpha > 0$, then $\alpha q_j \in L_j$ and $q_j + p_j \in L_j$. These subspaces are applied by Kruskal (1965), Kruskal and Shepard (1974), de Leeuw, Young and Takane (1976), Young, de Leeuw and Takane (1976), Young, Takane and de Leeuw (1978), and many others. Of course, for nominal data, we can also form subspaces. Again each variable corresponds to a subspace and an element corresponds to an arbitrary quantification of the categories of a variable. Obviously these spaces are linear. This definition too has been used in the articles of de Leeuw, Young and Takane mentioned above. Other choices of subspaces are polynomials of a certain degree, or polynomial splines (van Rijckevorsel, 1987). See also Chapters 2 and 3.

As regards the objective function we first mention the situation where there exists a partitioning of the variables into two sets. As objective function we may then use the canonical correlation; special, that is linear, applications are given by Anova, Manova, discriminant analysis and multiple regression (cf. van de Geer, 1986). When there is no prior partitioning of the variables into subsets, we treat the variables in a symmetric way and the objective function will then be invariant under permutations of the variables. Such a function is given for example by the sum of the correlations (Horst, 1965). Another class of objective functions is given by the symmetric functions of the latent roots (ψ_i) of the correlation matrix. We mention the determinant of the correlation matrix ($\prod \psi_i$), used by Chang and Bargmann (1974) and the sum of squares of the correlations ($\sum \psi_i^2$), suggested by Kettenring (1971). Another important function is the largest latent root of the correlation matrix, proposed by Horst (1965) and Carroll (1968).

The most popular programs for non-linear principal component analysis use the sum of the p largest latent roots of the correlation matrix as their objective function. Of course, maximizing the sum of the p largest latent roots is equivalent to minimizing the sum of the $m-p$ smallest latent roots. We shall denote these particular functions as μ_p, with the understanding that different choices of p result in different analyses.

2.2. Another definition of NCA

If we denote the intersection of a subspace L_j and the unit sphere S by L_jS, then all elements $q_j \in L_jS$ are normalized to unit length and for each choice $(q_1, \ldots, q_m)$, we can define a matrix $R(q_1, \ldots, q_m)$ with elements $r_{jl}(q_1, \ldots, q_m) = q_j'q_l$, the ordinary inner product of q_j and q_l. The matrix $R(q_1, \ldots, q_m)$ can be regarded as a correlation matrix, in the sense that it is positive semi-definite and that it has diagonal elements equal to one. The problem of non-metric principal component analysis (NCA) is to find $y_j \in L_jS$ in such a way that the sum of the p largest latent roots of the matrix $R(q_1, \ldots, q_m)$ is maximized. More generally, we shall be interested in all solutions of the stationary equations corresponding to the maximization of $\mu_p(R(q_1, \ldots, q_m))$. We consider the special case where the subspaces L_j are linear and of finite dimension k_j.

If the dimensionality of the total space L equals n, then orthonormal bases of $L_j (j = 1, \ldots, m)$ can be depicted by matrices F_j of order $(n \times k_j)$, and $F_j w_j \in L_jS$ if and only if $w_j'w_j = 1$. If we set $B_{jl} = F_j'F_l$, then,

$$r_{jl}(q_1, \ldots, q_m) = w_j' B_{jl} w_l.$$

We collect the matrices F_j in a supermatrix $F = (F_1, \ldots, F_m)$. The matrix $B = F'F$ is called the Burt table, named after Burt (1950). This latter matrix is of course dependent on the choice of a basis. If we define W as the direct sum $W = w_1 \# \ldots \# w_m$, so that $W'W = I$, then clearly,

$$R = W'BW, \tag{2.1}$$

$$\mu_p(R(q_1, \ldots, q_m)) = \operatorname{tr}(A'RA), \tag{2.2}$$

where the matrix A varies over all matrices of order $(m \times p)$, for which $A'A = I$. For a maximum of (2.2) the vectors of A must be latent vectors of R. If we write $\sigma_p(L_1, \ldots, L_m)$ for a maximum of $\mu_p(R(q_1, \ldots, q_m))$ with $q_j \in L_jS$, then,

$$\sigma_p(L_1, \ldots, L_m) = \max\{\operatorname{tr}(A'W'BWA)\} = \max\{\operatorname{tr}(T'BT)\}, \tag{2.3}$$

with A and W varying over matrices of the prescribed form, or T varying over matrices $T = WA$ of order $(\sum k_j \times p)$, for which $T'T = I$ and T consists of matrices $T_j = w_j a_j'$ of order $(k_j \times p)$, where a_j is row j of matrix A. Thus T is blockwise of rank one; each subspace L_j defines a block. The maximum (2.3) is found by differentiation. We first write:

$$\operatorname{tr}(A'W'BWA) = \sum_l \sum_j \sum_s a_{js} w_j' B_{jl} w_l a_{ls}$$

$$= \sum_l \sum_j \gamma_{jl} w_j' B_{jl} w_l, \tag{2.4}$$

where γ_{jl} are elements of $\Gamma = AA'$. A maximum subject to $w_j' w_j = 1$ is found by differentiating the function

$$\sum_l \sum_j \gamma_{jl} w_j' B_{jl} w_l - \sum_j \lambda_j (w_j' w_j - 1),$$

where λ_j are Lagrange multipliers. The maximum is found for

$$w_j' w_j = 1,$$

$$\sum_l \gamma_{jl} B_{jl} w_l = \lambda_j w_j.$$

As mentioned above, for a maximum of (2.4) the matrix A must comprise latent vectors of R, and thus the stationary equations are given by

$$RA = A\Omega, \tag{2.5}$$

$$\sum_l \gamma_{jl} B_{jl} w_l = \lambda_j w_j, \tag{2.6}$$

where Ω is a diagonal matrix of order $(p \times p)$ comprising latent roots of R.

From the equations, we can derive a relation between Ω and λ_j. An element of R is given by $r_{jl} = w_j' B_{jl} w_l$. An element of AA' is given by γ_{jl}. Both are symmetric matrices of order $(m \times m)$, so that for a diagonal element of RAA', we have on the one hand, according to (2.6),

$$\sum_l \gamma_{jl} w_j' B_{jl} w_l = \lambda_j,$$

on the other hand, according to (2.5),

$$RAA' = A\Omega A'.$$

Thus, we must have that λ_j is a diagonal element of $A\Omega A'$. Consequently,

$$\sum_j \lambda_j = \operatorname{tr}(\Omega), \quad \text{or} \quad \operatorname{tr}(\Lambda) = \operatorname{tr}(\Omega).$$

Although (2.5) and (2.6) can be used to construct convergent algorithms, they give little insight into the mathematical structure of the NCA problem.

It is not clear how many solutions there are of (2.5) and (2.6) nor is it clear how these various solutions might be related. However, there is one fortunate exception: $p = 1$.

2.3. Another definition of MCA

In the special case where $p=1$, the solution of (2.5) and (2.6) becomes much simpler and the problem of multiple correspondence analysis (MCA) is to compute some or all solutions. If $p=1$, the matrix T comprises one vector only and the restriction that the blocks of T must be of rank one is trivially satisfied. The solution is then found by maximizing $t'Bt$ subject to $t't=1$, and this is simply a latent root problem, leading to the equation,

$$Bt=t\mu. \tag{2.7}$$

Of course, we could decompose t into subvectors $t_j=w_ja_j'$, where a_j' is a scalar $a_j=(t_j't_j)^{1/2}$, and $w_j=t_j(t_j't_j)^{-1/2}$ (if $t_j't_j=0$, then w_j is arbitrary with $w_j'w_j=1$). The relations between w and λ_j in (2.5) and (2.6) are then as follows:

$$\lambda_j=\sum_l a_ja_lw_j'B_{jl}w_l=\sum_l t_j'B_{jl}t_l=t_j't_j\mu,$$

and

$$\sum_j \lambda_j=\Omega.$$

Hence $\Omega=\mu$. If μ is the largest, or smallest, latent root of B, then μ must also be the largest, or smallest, latent root of R. For intermediate latent roots of B, we can only state that they must also be latent roots of the corresponding matrix R.

In order to emphasize the relationship between NCA and MCA, we could define the latter as the maximization of,

$$\rho_p(L_1,\ldots,L_m)=\max.\{\mathrm{tr}(V'BV)\}, \tag{2.8}$$

where V varies over matrices of order $(\sum k_j\times p)$ and $V'V=I$. The main difference with NCA, (2.3), is that the blocks of V need not be of rank one. As a consequence and as contrasted with NCA, MCA is nested, i.e. the solution for $p=k$ is the same as the first k solutions found for $p>k$. In addition, a MCA solution, according to (2.8), generates p correlation matrices, whereas a NCA solution generates only one correlation matrix. Clearly, we have,

$$\rho_p(L_1,\ldots.L_m)\geq\sigma_p(L_1,\ldots,L_m), \tag{2.9}$$

with the equality holding if and only if the p dominant latent vectors of B have blocks of rank one.

3. CORRESPONDING MCA AND NCA SOLUTIONS

In this section, we shall derive a condition under which both MCA and NCA find the same solution. The emphasis will be on the possible interpretations of this situation.

3.1. A special condition

At the end of the previous section, we gave a condition under which the maxima (2.3) and (2.8) are equal: $\rho_p = \sigma_p$. The rigidity of this condition can be lessened by demanding that a solution of (2.3) equals a stationary value of (2.8). Then we can still say that the two solutions are equivalent. This new condition means that there are latent vectors of B which are block-wise of rank one, but which need not be necessarily the latent vectors corresponding to the p largest latent roots of B. In particular, the situation for which there are m such latent vectors is interesting, because then the relation between MCA and NCA holds for any p.

Therefore we assume that there are m latent vectors of B which can be decomposed in the manner of equation (2.3) to give: $V = WA$, $W = w_1 \# \ldots \# w_m$, $W'W = I$ and $A'A = I$ (because A is of order $(m \times m)$, we must also have $AA' = I$). This means that,

$$BWA = WA\Omega, \tag{3.1}$$

where Ω is again a diagonal matrix comprising latent roots. For these m different MCA solutions (by which we mean separate latent vector solutions and not m solutions of (2.8)), we find only one correlation matrix.

$$R = W'BW.$$

We already know that each MCA latent root corresponds to a latent root of the associated correlation matrix. In this case all latent roots of R are latent roots of B:

$$RA = W'BWA = W'WA\Omega = A\Omega. \tag{3.2}$$

We can now derive an important relation,

$$BW = BWAA' = WA\Omega A' = WRAA' = WR,$$

so that,

$$BW = WR, \quad \text{or,} \quad B_{jl}w_l = w_j r_{jl}. \tag{3.3}$$

Conversely, when there are vectors $w_j (j = 1, \ldots, m)$ satisfying (3.3), we can combine these vectors with p latent vectors of R, in order to form p latent vectors of B which are block-wise of rank one. For,

$$BWA = WRA = WA\Omega.$$

Although we can form m such latent vectors of B, it follows from (3.1) that $\rho_p (p < m)$ has a stationary value. Also σ_p has a stationary value. This follows, in particular, from the fact that the stationary equations (2.5) and (2.6) are satisfied:

(2.5): $RA = A$

(2.6): $$\sum_l \gamma_{jl} B_{jl} w_l = \sum_l \gamma_{jl} w_j r_{jl} = w_j \left(\sum_l \gamma_{jl} r_{jl} \right) = w_j \lambda_j.$$

Thus, if condition (3.3) is satisfied we can form m different MCA solutions and $C(m, p)$ stationary NCA solutions (using definition (2.8). We could also say that there are $C(m, p)$ stationary MCA solutions).

Besides MCA and NCA, condition (3.3) can also be related to other techniques of analysis. In fact, all (differentiable) functions $\mu(R(q_1, \ldots, q_m))$ have a stationary value if condition (3.3) is satisfied. A stationary value of $\mu(R(q_1, \ldots, q_m))$, where $q_j = F_j w_j$ is subject to $w_j' w_j$, is found by differentiating.

$$\mu(R(q_1, \ldots, q_m)) - \sum_j \lambda_j (w_j' w_j - 1), \tag{3.4}$$

where λ_j are Lagrange multipliers. As $r_{jl} = w_j' B_{jl} w_l$, we can use the chain-rule to find:

$$w_j' w_j = 1,$$

$$\sum_l \frac{\partial \mu}{\partial r_{jl}} B_{jl} w_l = \lambda_j w_j. \tag{3.5}$$

Clearly, if condition (3.3) is satisfied, then (3.5) is also satisfied, i.e. all functions $\mu(R(q_1, \ldots, q_m))$ have a stationary value. Of course, this is due to the fact that $R(q_1, \ldots, q_m)$ itself is stationary if (3.3) is satisfied. Conversely, if $R(q_1, \ldots, q_m)$ is stationary then (3.3) is satisfied.

Condition (3.3) corresponds to another desirable property. Suppose v is a MCA solution, and thus a latent vector of B, but not one of the m solutions that can be formed by using (3.3). As the latent vectors of B are orthogonal (when the latent roots are equal, they can be chosen to be orthogonal), we may write,

$$v'WA = (0, \ldots, 0), \quad \text{or,} \quad v'WAA' = (0, \ldots, 0), \quad \text{or}$$

$$v'W = (0, \ldots, 0), \quad \text{or,} \quad v_j' w_j = 0. \tag{3.6}$$

This means that the quantifications are orthogonal for each variable separately. In the terminology suggested by Dauxois and Pousse (1976), the solutions are not only weakly orthogonal, because they are latent vectors of B, but actually strongly orthogonal. Thus $F_j v_j$ and $F_j w_j$ are orthogonal, but $F_j v_j$ is also orthogonal to the transformations of the other variables $F_l w_l (l = 1, \ldots, m)$. For,

$$v_j' F_j' F_l w_l = v_j' B_{jl} w_l = v_j' w_j r_{jl} = 0.$$

This means that the space Z spanned by the vectors $F_j w_j (j = 1, \ldots, m)$ is orthogonal to the transformed variables of all other MCA solutions. In particular this holds for all other solutions of (3.3):

$$B_{jl} v_l = v_j s_{jl}.$$

If all subspaces L_j have the same dimensionality k, then the maximum number of solutions of (3.3) equals k. We can depict the condition for the existence of these k solutions as follows. If we collect the solutions $w_{js} (s = 1, \ldots, k)$ in an

orthonormal matrix K_j, and the correlations r_{jl} for the k different solutions in a diagonal matrix R_{jl}, then the total condition is given by:

$$B_{jl}K_l = K_j R_{jl}. \tag{3.7}$$

If the dimensionalities of the subspaces are different, we can maintain the formulation of (3.7), with the understanding that the matrices K_j are of order $(k_j \times k_j)$, where k_j is the dimensionality of subspace L_j, and R_{jl} is a matrix of order $(k_j \times k_l)$ having non-zero elements in the positions (i, i) where $(i = 1, \ldots, \min(k_j, k_l))$.

A solution of (3.7) for which all submatrices B_{jl} are diagonalized simultaneously, generates a number of orthogonal spaces $Z_s (s = 1, \ldots, \max(k_j))$. Each space Z_s is spanned by m_s vectors: those transformed variables for which $k_j \geq s$. Each space generates m_s separate MCA solutions and $C(m_s, p)$ NCA solutions (if $m_s < m$, we may use arbitrary quantifications for variables with $k_j < s$, while, at the same time, the weights a_j equal zero; for then the stationary equations are still satisfied). This amounts to $\sum m_s = \sum k_j$ different MCA solutions.

Before discussing the interesting interpretations of the conditions (3.3) and (3.7) we would like to make a remark on the possible nestedness of the NCA solutions. As we have already observed, the NCA solutions are not nested in general, i.e. correlation matrices formed at a maximum of σ_p are different for different p. If (3.7) is satisfied then the correlation matrix found at a maximum of σ_k also generates stationary values of σ_p, where $p \neq k$. However, it is not all that clear whether this correlation matrix corresponds to a maximum of σ_p, $(p \neq k)$.

With respect to the MCA solutions we can say that, although it is not impossible, the largest p latent roots are generally not generated by the same correlation matrix. Only if the p largest latent roots come from the same correlation matrix, do we have $\rho_p = \sigma_p$ (see (2.9)).

3.2. A general interpretation

If we represent (3.3) in words, it says that if two of the matrices C_{jl} have a subscript in common, then they have a singular vector in common. Condition (3.7) then says that they share all singular vectors corresponding to the common index. This is of course a very strong condition which greatly restricts the form of the matrices. The point of interest at the moment, however, is the interpretation. One possible interpretation says that the (normalized) quantifications of the variables remain invariant under removal of other variables from the analysis.

As far as we are considering the (normalized) quantifications, we could say that the analyses are nested with respect to the variables. This also holds for the condition (3.3) if we restrict our attention to the MCA and NCA solutions generated by this single system.

Obviously this situation always exists for numerical data. In that case the linear subspaces are one-dimensional and the condition (3.3) is a trivial one. We

also expect approximations of this ideal situation to be much better for ordinal data than for nominal data, the former being more 'one-dimensional' than the latter. We could say that the nominal variables measure more than one thing. Consequently, in a homogeneity analysis, we expect a nominal variable to act upon the other variables in two ways, the first of which is 'how well is something measured', the second is 'what is measured'. Therefore, we do not expect the conditions (3.3) or (3.7) to be satisfied, or even nearly satisfied for nominal data: variables will measure different things before and after removal of a variable from the analysis, that is to say, the (normalized) quantifications will be different.

Another interpretation is given by a reformulation of (3.3) in terms of projections. We can rewrite (3.3) as

$$F_j F_j' F_l w_l = F_j w_j r_{jl}. \tag{3.8}$$

This indicates that the projection of the transformed variable, $q_l = F_l w_l$, onto the subspace L_j, spanned by the orthonormal basis F_j, coincides with the projection on the one-dimensional space spanned by $q_j = F_j w_j$. This holds for all mutual projections. As a consequence, the projections of all vectors in the space Z (spanned by $q_l = F_l w_l (l = 1, \ldots, m)$) onto the subspaces L_j will coincide with the projections on the one-dimensional spaces q_j. Conversely, projections of vectors in L_j onto Z coincide with projections on q_j. Namely, if we denote a vector in L_j by Fa, where a is a vector containing zeros with the exception of those k_j elements that correspond to the k_j vectors of F_j, then the projection of Fa onto the space Z, spanned by FW, is given by:

$$\begin{aligned} FW(W'BW)^{-1}W'F'Fa &= FWR^{-1}W'Ba = FWR^{-1}RW'a \\ &= FWW'a = F_j w_j (w_j' a_j), \end{aligned} \tag{3.9}$$

which is a vector in the space spanned by $q_j = F_j w_j$. In the next subsection, we shall see that these projections have a special meaning if we use bases of indicator functions. Then we also have special interpretations of the conditions (3.3) and (3.7).

3.3. The interpretation for bases of indicator functions

In the first section, where we used indicator functions, we derived a relation between the optimal MCA-quantifications y_{js} and the object scores:

$$y_{js} = D_j^{-1} G_j' x_s.$$

This indicated that the projection of x_s on the subspace spanned by G_j coincides with the projection on the one-dimensional space of the transformed variable $G_j y_{js}$. As the vectors of G_j are indicator functions, we have a special interpretation for the operator $D_j^{-1} G_j'$. Variable h_j partitions the objects, or individuals, into k_j subgroups, indicated by the categories of variable h_j. The number of objects in

these groups can be found in the diagonal of $D_j = G_j'G_j$. The averages of a variable, for instance x_s, for these k_j groups can then easily be formed by:

$$D_j^{-1}G_j'x_s.$$

The averages y_{js} lie on the regression line of the linear regression of x_s on y_{js} (or $G_j y_{js}$). In fact, linear regression means projecting a vector on another vector, and a non-linear regression could be defined as the projection of a vector on the space of all non-linear transformations of another vector. In the present case of discrete data, non-linear regression means projecting onto a subspace L_j or G_j. If these two regressions, linear and non-linear, coincide, we say that the regression is linearized. In the present case this means that no other transformation of h_j, than $G_j y_{js}$, gives better predictions of x_s. The variance between the groups can be fully explained by a linear function of $G_j y_{js}$.

For stochastic variables the 'percentage of variance between groups' is expressed in the correlation ratio:

$$\text{c.r.}(\mathbf{x}_1\mathbf{x}_2) = \frac{\text{Var}(E(\mathbf{x}_1|\mathbf{x}_2))}{\text{Var}(\mathbf{x}_1)}, \tag{3.10}$$

which is the ratio of the variance of the conditional expectation of $\mathbf{x}_1$ given $\mathbf{x}_2$, and the variance of $\mathbf{x}_1$. If the conditional expectation lies on the regression line, then the correlation ratio equals the squared correlation r^2. For discrete variables the correlation ratio can be simply related to the projection on a subspace G_j.

Returning to MCA, the relations between the transformed variables can be formed by using (1.10):

$$\sum_l D_j^{-1}C_{jl}y_l = y_j\psi^2. \tag{3.11}$$

This indicates that the average regression of y_l on y_j is linear: the summation of the projections of all $G_l y_l (l = 1, \ldots, m)$ on G_j is in the one-dimensional space of $G_j y_j$. Let us now consider the situation where (3.3) or (3.7) is satisfied. The Burt table is given by:

$$B = D^{-1/2}CD^{-1/2}. \tag{3.12}$$

If we normalize the transformed variables, $y_j D_j y_j = 1$, then y_j can be written as $y_j = D_j^{-1/2}w_j$. Condition (3.3) is then given by:

$$D_j^{-1}C_{jl}y_l = y_j r_{jl}. \tag{3.13}$$

We now see that all regressions between the transformed variables are linearized. Hence, all correlation ratios equal the squared correlations. These findings are important because now the correlation matrix accounts for the whole bivariate relationship amongst the transformed variables, whereas it usually only gives linear relations.

4. SOME THEORETICAL EXAMPLES

Some reflection shows that the ideal conditions (3.3) and (3.7), which we discussed in the previous section, will usually not be met in practice. Very often the ideal situation can only be approximated. However, in this section we shall discuss some examples for some of which the conditions underlying this ideal situation are always met.

4.1. The trivial solutions

When discussing MCA in the first section, we observed that indicator functions, used as bases of the subspaces of non-linear transformations, generated meaningful solutions, as well as a number of trivial solutions. As we shall see, these trivial solutions have a quite natural place within the framework of the previous section. In fact, they give a trivial example where the condition (3.3) is always met.

We have already noticed that, by using indicator functions, condition (3.3) can be transformed as follows.

$$D_j^{-1} C_{jl} y_l = y_j r_{jl}.$$

This condition has a trivial solution given by: $r_{jl}=1$ and $y_j = un^{-1/2} (j, l = 1, \ldots, m)$ where the elements of u are units. All trivially quantified variables $G_j un^{-1/2} = un^{-1/2}$ have a length equal to unity and they coincide completely. They span a one-dimensional space Z_0. All elements of the so-called correlation matrix are units:

$$R = Y'CY = uu' \qquad (Y = y_1 \# \ldots \# y_m) \tag{4.1}$$

This matrix has one latent root equal to m, and $(m-1)$ latent roots equal to zero. These are the trivial MCA latent roots. The consequent strong orthogonality implies that the space Z_0 is orthogonal to all other transformed variables. Thus all other, meaningful transformed variables are in deviations from the mean, and the matrices R_s can be regarded as correlation matrices in the usual sense.

4.2. *Analyse des correspondances:* $m=2$

If we apply MCA to a data set of only two variables, the supermatrix of bimarginals has only one contingency table as submatrix. It is well known that, if we apply a similar analysis to this contingency table instead of to the supermatrix of bimarginals, the solutions agree up to a normalization and the latent roots can be directly related to one another (cf. Gifi, 1981a). This technique of analysis is called *analyse des correspondances* and it is discussed by many authors; see for instance Benzécri *et al.* (1973), Nishisato (1980), Gifi (1981a) or Greenacre (1984).

In this case of two variables, condition (3.7) is always met. In fact, when all submatrices B_{jl} can be diagonalized simultaneously, then condition (3.7) is satisfied:

$$K_j' B_{jl} K_l = R_{jl},$$

and this will always be possible when there are two variables, since then only one submatrix B_{12} has to be diagonalized. In that case a solution for (3.7) can be found easily by taking singular vectors of B_{12}. The orthogonal spaces Z_s are now spanned by two vectors, and the corresponding correlation matrices R_s have only one subdiagonal element which equals the singular value r_{12} of B_{12}. The MCA latent roots are consequently one plus the singular value and one minus the singular value.

In the case of *analyse des correspondances*, the regressions between the transformed variables are always linearized and the quantifications are always strongly orthogonal. Of course, in this case, NCA is not very meaningful.

4.3. Dichotomous variables: $k_j = 2$

For dichotomous or binary variables, the regressions are by definition linear, because a straight line can always be drawn through two points. If $k_j = 2$, the subspaces L_j are two-dimensional. However, one dimension is due to the trivial solution Z_0, which means that the variables can be quantified in deviations from the mean in only one direction. The frequencies directly induce a quantification, and similar to the case of numerical data, there is nothing to quantify. We simply compute the product-moment correlations (phi-coefficients, or point-correlations) and perform a PCA of the correlation matrix.

4.4. Normally distributed variables

Hitherto, we have assumed the dimension of a subspace L_j to be finite. In this section we shall consider subspaces of infinite dimensions. We have seen that the indicator functions are a perfectly satisfactory basis if L_j is the space of all non-linear transformations of a discrete variable which assumes only a finite number of values. However, if the number of categories of the variables is very large and close to the number of observations, difficulties might arise. In that case the use of indicator matrices is no longer satisfactory: for they will be close to a permutation of the identity matrix. Next to G_j, this will also be true for the matrices D_j and C_{jl}. Consequently, all latent roots of the multiple correspondence problem will be close to either zero or one, and the latent vectors will be very unstable and rather uninteresting.

This solution occurs for continuous variables. For, in practice, dealing with continuous variables really means dealing with discrete variables with a very large number of categories, close to the number of observations. In these cases the

space of all non-linear transformations is simply too big, because it is approximately equal to the whole space L, and if each L_j is approximately equal to L then non-linear PCA does not make sense. Thus, we want to approximate the space of all non-linear transformations by using small-dimensional subspaces. Indicator functions correspond with approximations of non-linear functions by step functions. In the theory of the approximation of functions, it is well known that step functions give poor approximations of continuous or smooth mappings. A classical alternative is given by polynomials. Although polynomials have many attractive theoretical properties, they are not very suitable for approximating general continuous functions. The basic problem is that polynomials are too rigid; if we change a coefficient then the whole polynomial changes. Therefore polynomials of a very high degree are needed for the approximation of functions which do not behave in the same way over the whole range. A more satisfactory alternative is given by the so-called B-splines (De Boor, 1978). We shall not go further into this matter, this being dealt with in the Chapters 2 and 3. However, we will give an example in which polynomials are used. The theoretical example of this section is given by the multinormally distributed variables, for which it is known that all bivariate regressions are linearized, that is to say, condition (3.7) is satisfied. Suppose the random variables $\mathbf{x}_j$, $(j=1, \ldots, m)$ are jointly multivariate normal, with zero means, unit variances and correlations ρ_{jl}. The subspaces of non-linear transformations are given by:

$$L_j = \{\phi_j(\mathbf{x}_j) \,|\, \text{var}\{\phi_j(\mathbf{x}_j)\} < \infty\}. \tag{4.2}$$

The mapping ϕ_j are assumed to be measurable. As a basis for L_j we use Hermite–Chebyshev polynomials $\psi_\nu(\,.\,)$, of degree $\nu = 0, 1, \ldots$ which are orthonormal on the normal distribution:

$$\int \psi_\nu(x)\psi_\chi(x)N(x)\,dx = \delta(\nu, \chi) \tag{4.3}$$

where $\delta(\nu, \chi)$ is the Kronecker delta. We can now expand $\phi_j(\mathbf{x}_j)$ as,

$$\phi_j(\mathbf{x}_j) = \sum_\nu a_{j\nu}\psi_\nu(\mathbf{x}_j). \tag{4.4}$$

For the covariance of $\phi_j(\mathbf{x}_j)$ and $\phi_l(\mathbf{x}_l)$ we have,

$$\text{cov.}\,(\phi_j(\mathbf{x}_j), \phi_l(\mathbf{x}_l)) = \sum_\nu \sum_\chi a_{j\nu}a_{l\chi}\text{cov.}(\psi_\nu(\mathbf{x}_j), \psi_\chi(\mathbf{x}_l)). \tag{4.5}$$

The following identity is due to Mehler (cf. Lancaster, 1969),

$$\text{cov.}\,(\psi_\nu(\mathbf{x}_j), \psi_\chi(\mathbf{x}_l)) = \delta(\nu, \chi)\,(\rho_{jl})^\nu, \tag{4.6}$$

where ρ_{jl} is the correlation between $\mathbf{x}_j$ and $\mathbf{x}_l$. Clearly then,

$$\text{cov.}\,(\phi_j(\mathbf{x}_j), \phi_l(\mathbf{x}_l)) = \sum_\nu a_{j\nu}a_{l\nu}(\rho_{jl})^\nu. \tag{4.7}$$

With respect to (4.6) we could say, in the terminology of the previous section, that polynomials of the same degree, $\psi_\nu(\mathbf{x}_l), \ldots, \psi_\nu(\mathbf{x}_m)$, span a space Z_ν, orthogonal to all other spaces Z_χ, spanned by polynomials of another degree, $\chi \neq \nu$. This means that we could as well have taken the polynomials of a certain degree directly as the (normalized) non-linear transformations. For then condition (3.7) is satisfied. The MCA and NCA solutions are again generated by the latent root solutions of the correlation matrices:

$$R^{(\nu)}a_\nu = \omega a_\nu, \qquad \nu = 1, 2, \ldots, \tag{4.8}$$

where $R^{(\nu)}$ has elements $(\rho_{jl})^\nu$. Thus, for every matrix $R^{(\nu)}$ we find m MCA solutions and $C(m,p)$ NCA solutions. It is implied by general results on Hadamard products (Styan, 1973) that the largest latent root of $R^{(1)}$ cannot be smaller than the largest latent root of $R^{(2)}$, which in its turn cannot be smaller than the largest latent root of $R^{(3)}$, etc. For the smallest roots the converse holds. The largest and smallest MCA latent roots are, consequently, the largest and smallest roots of $R^{(1)}$, and both correspond to linear transformations.

The second largest latent root is more of a problem. It can either be the second largest latent root of $R^{(1)}$, or the largest latent root of $R^{(2)}$. In the former case, both the first and second MCA dimensions correspond to linear transformations; in the latter case the first MCA dimensions corresponds to a linear transformation, whereas the second dimension corresponds to a quadratic transformation.

For homogeneous variables, the largest latent root of $R^{(1)}$ will be considerably larger than the second latent root of $R^{(1)}$, and usually also the largest latent root of $R^{(2)}$ will be larger than the second root of $R^{(1)}$. Thus, homogeneous variables approximating the normal distribution usually have mappings (into the first two dimensions) which approximate parabolas: horseshoes.

For general ordinal conditions for data matrices to have horseshoe mappings we refer to Schriever (1986). The question, however, is whether this second, quadratic transformation contributes to our knowledge about the relations amongst the variables. In his doctoral thesis de Leeuw (1973) writes: 'As pointed out by Bartlett [1953] and McDonald [1968], in the classical case we suppose, more or less implicitly, that the component scores are stochastically independent'. In the present case there is a simple non-linear relation between the components and, obviously, the dimensions are not independent. As we noticed before, correlations can only fully account for the bivariate relations amongst the variables if the regressions are linearized. In our case, the regressions between the linear transformations and the quadratic transformations are not linearized at all; they are 'parabolized'. The non-correlatedness of the dimensions says nothing about the non-relatedness of the dimensions. Thus, in case one finds a horseshoe mapping, it is best to consider only the correlation matrix of the first MCA solution, or NCA solution. See also van Rijckevorsel (1987).

The NCA solutions are always generated by one of the $R^{(1)}$, $R^{(2)}, \ldots,$ matrices. In practice, it is usually the first matrix $R^{(1)}$, which is the original, non-transformed, correlation matrix.

We discussed some situations where (3.3) and (3.7) hold. Data analytically the situations in which these properties are approximately true are just as interesting. Often we do not know nor can we find out beforehand how the data are distributed or whether there exist any other dependencies. If simultaneous diagonalization is approximately true the correlations in the off-diagonal blocks must be small; the percentage of corresponding dimensions between the eigenvectors of $R^{(1)}, R^{(2)}, \ldots, R^{(q)}$ and the observed complete set of homogeneity analysis scores is a measure for the existence of the blockwise rank-one structure. One can compute the correlations between the approximately diagonalizing eigenvectors and the actual homogeneity analysis transformations. De Leeuw (1982) and Bekker (1983) respectively constructed an algorithm and a computer program (PREHOM) to this purpose. The practical value of the program is not that obvious, but an application provides a tangible illustration of the blockwise rank one approximation. To this purpose we show the following example by van Rijckevorsel (1987) using the Holmquist data.

Pathologists study biopsy slides of the uterine cervix in order to classify carcinoma *in situ* and related lesions into five classes:

1 = negative
2 = atypical squamous hyperplasia
3 = carcinoma *in situ*
4 = squamous carcinoma with early stromal invasion
5 = invasive carcinoma

cf. Holmquist, McMahan and Williams (1967).

The major decision to give treatment or not is usually based on the fact whether a slide belongs to the classes 1 or 2 (= no treatment) or to the classes 3, 4 or 5 (= treatment). If several pathologists study the same set of slides we would like their judgements to be consistent at least with respect to give treatment or not. In practice this is an ideal and unrealistic situation because there exists no absolute true scale of slides that is perfectly partitioned diagnostically. Even for the best pathologist in the world there occasionally exist serious doubts about the right classification of certain slides.

Slides and subsets are quantified by the parameters x and y respectively, and the problem is to find a common scale x for slides and scores y for categories such that the common scale is maximally consistent with all weighted judgements of each pathologist $G_j y_j$ simultaneously.

These data are of the rating scale type and they show a non-perfect horseshoe in the first two dimensions of a homogeneity analysis. We do not report on this analysis here. The eigenvalues with their approximations and the correlations between the corresponding eigenvectors per axis are shown in Table 1.1

Table 1.1. The eigenvalues of and the correlations between the homogeneity analysis solution and its block diagonal approximation of the Holmquist data

	Eigenvalues				
	θ_1	θ_2	θ_3	θ_4	θ_5
Actual	5.56	5.20	2.75	2.09	1.68
approximation	5.50	5.24	2.50	2.21	1.53
Correlations between the corresponding eigenvectors	0.92	0.92	0.84	0.82	0.87

Without any extra assumptions the axes of the homogeneity analysis have approximately a blockwise rank-one structure, at least in the first five dimensions. This is more or less to be expected because of the type of data, the number of variables, the number of categories and the occurrence of a horseshoe in the first two dimensions. We observed a particular bad fit of pathologist no. 6 in the preceding homogeneity analysis; this is also reflected in a bad approximation regarding the size of the off-diagonal correlations between dimensions for this pathologist (not displayed). However, it is more efficient to look at the percentage of overlap between corresponding observed and approximated dimensions per pathologist. This can be expressed in chi square per pathologist between observed and approximate transformations (Bekker, 1983) see Table 1.2. Pathologist no. 6 has the smallest overlap and is hence the least diagonizable. Tentatively this could mean that no. 6's response is not so regularly distributed, is less of a rating scale type or is less order dependent than the responses of the other pathologists. To locate this odd man out by block diagonalization or by testing for order dependence in another way than by ordinary homogeneity analysis is respectively too cumbersome or downright impossible (there exist no statistical tests for order dependence (cf. Schriever, 1986)).

Table 1.2. The percentage of chi square in corresponding dimensions of the Holmquist data

Pathologist							*Averaged*	
							1	94
2	97						2	97
3	93	98					3	96
4	98	98	97				4	96
5	98	97	96	96			5	97
6	81	93	97	90	95		6	92
7	100	98	93	98	99	96	7	98