LaTeX documents for the North-Holland series
in Statistics and Probability
W.j. Maas

Some Generalizations of Correspondence Analysis

Jan de Leeuw

Department of Mathematics UCLA, 405 Hilgard Avenue, Los Angeles, CA 90024-1555

Abstract

Correspondence analysis of a bivariate table has many different optimality properties. For instance, the scores computed by correspondence analysis linearize both regressions, and maximize the correlation coefficient. We try to generalize both properties to the multivariate situation, using the concept of maximizing any aspect of the correlation matrix, and the concept of simultaneously linearizing all bivariate regressions.

1. Introduction

Correspondence Analysis (CA) is a technique which has been reinvented many times, in many countries, and in many disciplines. Historical reviews are, for instance, in Nishisato [34], Tenenhaus and Young [38], Gifi [20], De Leeuw [12], and Benzécri [2]. Some of the reasons for the multiple discovery phenomenon are implicit in our first sentence. Until quite recently there was not much communication between countries, and almost no communication between disciplines. There is another reason, however, which is more interesting from a scientific point of view. The equations defining CA can be derived in many different ways, from many different starting points. We give a brief and sadly incomplete overview.

In the French approach to CA, initiated by Benzécri around 1965, a distance measure is defined on the rows and/or columns of a table, and these distances are approximated by Euclidean distances in a low-dimensional representation of the table. Excellent descriptions of this approach are in Greenacre [25] or Benzécri [3]. Thus CA is a form of *metric multidimensional scaling.* The Anglo approach to CA quantifies the row and column categories of a table in such a way that some optimality criterion is satisfied. As we shall indicate below, early work on this approach was done

by Pearson [35] and Hartley [27], but Fisher [17],[18] and Maung [32],[33] were the first to apply the technique to real data.

In psychometrics, Guttman [26] invented *multiple correpondence analysis* (MCA) using the optimal scoring approach. MCA applies the optimal scoring approach to multivariate tables, generalizing (in some respects) principal component analysis (PCA). Guttman went on to generalize optimal scoring to other forms of multivariate analysis (compare [10] for an overview of his contributions). Guttman's approach to multivariate analysis was imported in Japan by Hayashi in the early 1950's, and rediscovered by Burt in England around the same time. The optimal scoring approach, or optimal scaling approach, was systematized from a programmatic and computational point of view, first by De Leeuw [10], and on the basis of this by De Leeuw, Young, and Takane in a series of papers in Psychometrika in the late 1970's. These ALSOS (Alternating Least Squares with Optimal Scaling) papers are summarized in Young [42], and eventually lead to the book by Gifi [20]. In a parallel development, optimal scaling using smoothers was systematically applied by Breiman and Friedman [5] and their students, using the acronym ACE (Alternating Conditional Expectations). Compare Buja [6] for a recent review. Finally, in another Anglo tradition, CA was rediscovered by Goodman and Haberman, when they were extending log-linear analysis to deal with tables having ordered categories. Recent papers in this tradition are [21], [22], [23], [24],[40],[39].

The classical work on CA, and the more recent work linking CA with log-linear modelling, concentrate on the bivariate situation, in which we have a single cross-table. In this paper, we shall talk mainly about the multivariate case, which is, in a sense, more challenging.

The basic motivation for developing correspondence analysis techniques for multivariate data, according to Gifi [20], is that there is a wide gap in MVA between the multinomial and the multinormal. There are discrete numerical variables, ordinal variables, and non-normal numerical variables. In applied work in the social, behavioural and life sciences, discrete numerical and ordinal variables seem to be the rule rather than the exception. Yet most MVA techniques are designed for either purely normal data or purely nominal data. The multinormal is obviously too strong a model for most applications, and the multinomial (log-linear) models are too weak for high-dimensional situations. Thus we need to develop a class of techniques that is in between the two. With the multinormal MVA tradition we have in common that we only use the bivariate marginals of the table, with the multinomial tradition we share the emphasis on nonparametric modeling.

Given the French geometric approach to CA, and the Anglo optimal scaling approach, it is not surprising that MCA and its various generalizations have also been discussed in a number of essentially different ways.

Benzécri and Greenacre continue to use chi-square distances, defined on a cross table of indicator matrices. Since zero-one matrices are not frequencies, the chi-square metric is not very natural, and the approach more or less breaks down. Gifi [20] emphasizes the specific geometry of MCA, and this geometry is even more central in [16]. In this paper we emphasize optimal scaling, which means at the same time that we stress the relationships between CA and other forms of MVA. This paper extends and summarizes [1], [11], [13],[14].

2. Formalism

We give some notation and terminology first. Suppose we have m random variables $\underline{x}_j$, with finite variances, all defined on the same probability space $(X, \mathcal{B}, P)$. The (real, separable Hilbert) space of all finite-variance random variables on this space is $\mathcal{L}$. We define $\mathcal{L}_j$ as the subspace of all measurable transformations of $\underline{x}_j$ with finite variance. Suppose $\mathcal{K}_j$ is a subspace of $\mathcal{L}_j$, of dimension p, and $\{g_{js}\}$ is a basis for $\mathcal{K}_j$. We assume $\underline{x}_j \in \mathcal{K}_j$.

For ease of notation, we suppose all bases have the same dimension p, but this is no real restriction. Also, for most results there is no reason to exclude the case $p = \infty$, because the infinite sums replacing our matrix operations will converge in mean square in $\mathcal{L}$. By another slight, but inconsequential, misuse of matrix notation we use $z_j = G_j\alpha_j$ to describe transformations in $\mathcal{K}_j$, where the "matrix" G_j has $\dim(\mathcal{L})$ rows and p columns, and contains the elements of the basis $\{g_{js}\}$ as columns.

We write $C_{j\ell} = G_j'G_\ell$ for the cross products of the bases. Also, D_j is short for C_{jj}. Observe that for orthogonal bases D_j is diagonal, for an orthonormal basis it is the identity. We can collect the $C_{j\ell}$ in an $mp \times mp$ supermatrix C, which is called the *Tableau de Burt* in the French CA literature, after Burt [7].

The cross product of any two transformations of the form $z_j = G_j\alpha_j$ and $z_\ell = G_\ell\alpha_\ell$ can simply be written as $\alpha_j'C_{j\ell}\alpha_\ell$. If $\mathcal{K}_j$ only has centered random variables, then this is the covariance of z_j and z_ℓ, if in addition $\alpha_j'D_j\alpha_j = \alpha_\ell'C_\ell\alpha_\ell = 1$, then it defines the correlation.

For illustrative purposes we mention two examples of the general framework. Ordinary contingency tables are usually dealt with by taking as a basis the *indicator matrices* or *dummies* coding the categories. This makes $C_{j\ell}$ equal to the cross-table of variables j and ℓ. It makes D_j equal to the diagonal matrix with univariate marginals. If the joint distribution of the $\underline{x}_j$ is a standardized multivariate normal, then the basis we use are the normalized Hermite-Chebyshev polynomials. In that case the $C_{j\ell}$ are diagonal, with on the diagonal the successive powers of $\rho_{j\ell}$, the correlation coefficient between variables j and ℓ in the multivariate normal.

This particular way of treating finite contingency tables and continuous multivariate distributions by basically the same formalism was first suggested in this context by Fisher (communication to Maung [32]). It has been used succesfully by Lancaster [30],[31], and by the ACE group. Of course it was already old hat in functional analysis by that time. There have been several interesting generalizations of the framework. Dauxois and Pousse [9] allow for an infinite number of variables, by using the continuous direct sum of the subspaces of the underlying Hilbert space. Koster [29] extends the framework by replacing the subspaces by convex cones.

Let us also define linear regression in this context. We say that $\underline{z} \in \mathcal{L}$ has a linear regression on $\underline{x}_j$, if the projection of $\underline{z}$ on $\underline{x}_j$ is the same as the projection of $\underline{z}$ on $\mathcal{L}_j$. This means that we must have $\mathcal{P}_j(\underline{z}) = \lambda \underline{x}_j$. In matrix form this is $G_j(G_j'G_j)^{-1}G_j'z = \lambda x_j$. Now suppose $\underline{z} = \underline{x}_\ell = G_\ell \alpha_\ell$, and $\underline{x}_j = G_j\alpha_j$. Then $\underline{x}_\ell$ has linear regression on $\underline{x}_j$ if $G_j(G_j'G_j)^{-1}G_j'G_\ell = \lambda G_j\alpha_j$, i.e. if $C_{j\ell}\alpha_\ell = \lambda D_j\alpha_j$.

For completeness we define MCA. We solve the generalized eigenvalue problem $C\alpha = \lambda D\alpha$, or, in more detail,

$$\sum_{\ell=1}^{m} C_{j\ell}\alpha_\ell = \lambda D_j\alpha_j.$$

If we compare this we the previous paragraph, we see that MCA finds scores such that the "average regression" between the transformed variables is linear.

3. Linearizing the regressions

In 1906 Pearson published a paper [35] in which he proved the following result. At least we can *interpret* his paper as proving the following result.

There are two categorical variables, with indicator bases G_1 and G_2. Assign scores to the rows and columns of their cross table $C = G_1'G_2$, with marginals in the diagonal matrices D and E. Suppose the scores a and b are in deviations from the mean, with unit variance. Thus the correlation *induced* by the scores a and b is $r(a,b) = a'Cb$.

Now perturb the scores, again with vectors in deviations from the mean δ_a and δ_b. Then

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{r(a+\epsilon\delta_a, b+\epsilon\delta_b) - r(a,b)}{\epsilon} &= \\ &= \delta_a'(Cb - r(a,b)Da) + \delta_b'(C'a - r(a,b)Eb).\end{aligned}$$

The interpretation of this formula is quite simple. If both regressions *are* linear, then the right-hand side is zero, i.e. if both regressions are linear

the correlation coefficient is relatively insensitive to small modifications of the scoring. For score-changes of order $O(\epsilon)$ the correlation-change is $O(\epsilon^2)$. It is clear why Pearson was interested in this result. We get roughly the same value of the correlation coefficient, even if we are not sure about the scoring. Pearson was mainly interested in interchanging two columns and/or rows, which was interpreted as an example of a small change. It is a somewhat unfortunate example, because of its discreteness, but it is not too complicated to fit it into the general result. The explicit version of the argument for general random variables is in [12].

There are a number of ways in which we want to extend Pearson's result. In the first place he starts with scores linearizing the regressions, and looks in neighborhood of these scores. It is not entirely clear under what circumstances such linearizing scores actually exist. In the second place we would like to generalize the result to more than two variables. And, finally, there is no reason to restrict our attention to categorical variables with indicator bases. The generalization to other examples is more or less immediate, using the notation from the previous section. In order to get at the existence question, we rewrite our result as

$$\begin{aligned}\frac{\partial r}{\partial a} &= Cb - r(a,b)Da,\\ \frac{\partial r}{\partial b} &= C'a - r(a,b)Eb,\end{aligned}$$

and we take the next step in the history of CA.

4. Maximizing the Correlation

In 1935 Hirschfeld (who later changed his name to Hartley) published [27] in which he proved (quite explicitly) the following.

Suppose we want to find scores that linearize the regressions in a cross table C. Thus we want

$$\begin{aligned}Cb &= \rho Da,\\ C'a &= \rho Eb.\end{aligned}$$

This system always has $p-1$ non-trivial solutions, given by the generalized singular values and singular vectors of the triple (C,D,E). The vectors of scores are mutually orthogonal, etc. Moreover (generalizing Pearson) these scores give maxima, and saddle points, and minima of the correlation coefficient. Hartley knew about the work of Hilbert and Schmidt on what is effectively the singular value decomposition, and he could consequently provide the existence theorem that had eluded Pearson. Also, the singular

value decomposition provided an expansion of the bivariate distribution, which was a special case of the expansions studied by Schmidt and Mehler.

Hartley's result for finite tables was generalized to some extent by Fisher and Maung around 1940, and by Lancaster et al. since 1955, see [31], to general bivariate distributions. The idea of using the maximum of the correlation coefficient over scores as a measure of association is due to Gebelein [19], and it has been studied in detail by Renyi [36].

Again, we would like to find out what happens if $m > 2$. Can we find scores which linearize *all* the bivariate regressions, or perhaps even all the multivariate regressions as well. And if we can't in general, under what conditions do such scores exist.

5. More than two variables

It is clear that for $m > 2$ variables things are not so simple any more. In general, not all bivariate regressions (let alone all multivariate regressions) can be linearized by scoring (or, if you prefer, transformation). Let us call a multivariate distribution *bi-linearizable* if all bivariate regressions can be linearized by scoring. In obvious notation there exist m different vectors $\{\alpha_1, \cdots, \alpha_m\}$ such that

$$C_{j\ell}\alpha_\ell = \rho_{j\ell} D_j \alpha_j .$$

Although each of these equations can be solved easily, and has multiple solutions, by Hartley's result, in general the solution for α_j from equations (j, ℓ) and (j, ν) will be different. We see that the condition means that the singular value decompositions of the bivariate distributions are linked, in the sense that decompositions with an index in common also have a singular vector in common. This can also be written by defining the matrices

$$T_{jl} = D_j^{-1} C_{j\ell} D_\ell^{-1} C_{\ell j} .$$

For each j the m matrices $T_{j\ell}$ must have an eigenvector in common.

We give some simple examples of bi-linearizable distributions.

- all variables are binary,
- there are only two variables,
- special cases, such as the multinormal (or elliptical),
- the *strained multinormal* in the sense of Yule.

The first example is trivial. Obviously we can draw a line through two points. The second example is Hirschfeld's theorem. Regressions are linear in multinormal of elliptical distributions, so obviously they can be linearized. The strained multinormal will be discussed in more detail below. We see that assuming that the multivariate distribution is bi-linearizable is

an important generalization from assuming it to be multinormal (or elliptical). Cuadras [8] shows that bi-linearizable distributions with arbitrary marginals exist.

Observe that we do not assume that *all* regressions can be linearized. Considering all multivariate regressions would take us into the realm of higher-dimensionsal tables again, and we would run into the empty-cell problem (also known as the *curse of dimensionality*). We concentrate on properties of the bivariate marginals, which will be reasonably well-filled even in high-dimensional situations.

6. Strained Multinormals

The notion of a strained multinormal is not very well known, so we describe it a bit more in detail.

Suppose $\underline{x} = (\underline{x}_1, \cdots, \underline{x}_m)$ is multivariate normal. Now suppose ϕ_j are strictly increasing, and define $\underline{y} = (\phi_1(\underline{x}_1), \cdots, \phi_m(\underline{x}_m))$. Then $\underline{y}$ is strained multinormal. Thus the marginal normality is destroyed by applying separate transformations to all variables. Obviously we can *unstrain* $\underline{y}$ by applying the inverse transformations ϕ_j^{-1}. The notion is due to Yule, in his discussion of Pearson's tetrachoric correlation coefficient [43].

> We assume, then, not necessarily, that the correlation surface is normal, but that it is "strained normal," as we may term it, and it is conceivable that "strained normal" may cover markedly skew correlation tables (l.c., page 141).

We can write down expressions for the distribution and density of a strained multinormal quite easily. Use Ψ and ψ for the standard multinormal distribution and density. The distribution is

$$F(z_1, \cdots, z_m) = \Psi(\phi_1^{-1}(z_1), \cdots, \phi_m^{-1}(z_m)).$$

and thus the density is given by

$$f(z_1, \cdots, z_m) = \frac{\partial^m F}{\partial z_1 \cdots \partial z_m} = \psi(\phi_1^{-1}(z_1), \cdots, \phi_m^{-1}(z_m)) \prod_{j=1}^{m} \frac{\partial \phi_j^{-1}}{\partial z_j}.$$

This creates a fairly general family of multivariate distributions. In a sense it generalizes the approach to transformations popularized by Box and Cox [4]. In a strained multinormal we can of course linearize *all* regressions (not only the bivariate ones) by unstraining. Thus assuming strained multinormality is stronger than assuming bi-linearizability.

There is, by the way, a condition logically in between strained normality and bi-linearizability: suppose orthonormal systems $\{\alpha_{j1}, \cdots, \alpha_{jp}\}$ exists such that

$$C_{j\ell}\alpha_{\ell s} = \rho_{j\ell s} D_j \alpha_{js}.$$

This could be called bi-linearizable of order p. We have bi-linearizability of order p if the for each j the m matrices $T_{j\ell}$ have a complete system of eigenvectors in common, which happens if and only if they commute. The standardized multivariate normal is bi-linearizable of all orders, because the Hermite-Chebyshev polynomials can be chosen as the common eigensystem.

7. Some questions

We have defined bi-linearizable distributions, and we studied some special cases. Immediately we are stuck with a number of questions about the linearizing transformations.

- If they exist, how do we find them ? (estimation)
- Do they exist ? (test of fit)
- What do they do to the standard errors ? (precision)
- How do they look ? Are they useful ? (data analysis)

In the rest of this paper we shall try to answer the first three of these questions. The fourth one can only be answered by looking at many examples, and for this we refer to the book by Gifi [20]. We have already discussed the existence question above, algebraically, but we have not translated our results into a statistical test.

8. LPV diagonalization

We start with the question on how to find linearizing scores. There is a straightforward direct approach. We have cross-tables $C_{j\ell}$ and univariate marginals in diagonal matrices D_j. For standardized scores the correlations are

$$\rho_{j\ell} = \alpha_j' C_{j\ell} \alpha_\ell,$$

and the correlation-ratios are

$$\eta^2_{j\ell} = \alpha_j' C_{j\ell} D_\ell^{-1} C_{\ell j} \alpha_j.$$

Obviously

$$\rho^2_{j\ell} \leq \eta^2_{j\ell},$$

with equality if and only if the regression of variable ℓ on variable j is linear. This indicates one straightforward way of finding the scores, if they exist. Minimize

$$\tau(\alpha_1, \cdots, \alpha_m) = \sum_{j=1}^{m} \sum_{\ell=1}^{m} (\eta_{j\ell}^2 - \rho_{j\ell}^2).$$

This loss function can be minimized quite easily by changing one set of scores at the times, and cycling through the m sets iteratively. The subproblem of finding an optimal set of scores for variable j, with the other $m-1$ sets fixed at their current values, is a small generalized eigenvalue problem.

A more general approach, which can be used for bi-linearizability of higher orders, is taken in [1] and [11]. The approach gives us a lot of insight into the MCA problem, and consequently it is not merely a computational tool. We collect the $C_{j\ell}$ in the $mp \times mp$ Burt table, and the D_j in the corresponding diagonal matrix. Multiple correspondence analysis solves the generalized eigenvalue problem $C\alpha = \lambda D\alpha$, which generally has mp solutions. Or, to put it differently, we look for an orthonormal K of order mp that diagonalizes $H = D^{-1/2}CD^{-1/2}$. We can compute K by the usual techniques for solving the symmetric eigenvalue problem. But instead of doing that, we shall try to build up K from simpler components in three steps.

Start with m orthonormal L_j of order p such that all

$$U_{j\ell} = L_j' D_j^{-1/2} C_{j\ell} D_\ell^{-1/2} L_\ell$$

are diagonal. Such L_j need not exist, but if they do, and we collect them in the direct sum $L = L_1 \oplus L_2 \oplus \cdots \oplus \mathrm{L}_m$, then $U = L'HL$ has submatrices which are all diagonal. It is now possible to find a permutation matrix P such that $P'UP = R_1 \oplus R_2 \oplus \cdots \oplus R_p$, where the R_j are of order m. We find R_1 by selecting all the $(1,1)$ elements of the $U_{j\ell}$, R_2 by selecting the $(2,2)$ elements, and so on. Now $P'UP = P'L'HLP$, and we have found the orthonormal matrix LP which transforms H to direct sum form. But obviously there exist orthonormal V_r which diagonalize the R_r. If V is their direct sum, then LPV diagonalizes H.

The computations in the previous paragraph can be carried out exactly if and only if we can find orthonormal L_j such that $U_{j\ell}$ are diagonal. In general, we cannot (because it would mean that the distributions are indeed bi-linearizable of order p). What we can do is minimize the sun of squares of the off-diagonal elements of the $U_{j\ell}$. A convenient and rapid way to do this is by using Jacobi-like plane rotations to build up the L_j. In [1] two

strategies are discussed. The first one minimizes the sum of squares of *all* off-diagonal elements. The second one minimizes the sum of squares of the elements in the first row and column of all $U_{j\ell}$ only. It is easy to see that we can make this particular sum of squares equal to zero if and only if the distributions are bi-linearizable. Moreover, the sum of squares is exactly equal to the sum of differences between the correlation ratios and the squared correlation coefficients $\tau(\alpha_1,\cdots,\alpha_m)$ we used a few paragraphs ago. If we are done, we fix the first row of the L_j, and start on the second one. And so on.

In any case, we have build up an orthonormal LPV which approximately diagonalizes H. And, of course, we also have an orthonormal K which exactly diagonalizes H. The point made in [11] is that the LPV approach in many cases gives much more insight into the MCA problem. In order to illustrate this, think of MCA as a form of nonlinear component analysis. Each system of scores, i.e. each column of K, can be used to compute an *induced correlation matrix*, and each induced correlation matrix can be submitted to a regular PCA. But there are mp such correlation matrices, and we consequently find m^2p principal components. Gifi calls this "data production." Now suppose α_j is a bi-linearizing set of scores. Let us look at vectors of the form $\theta_j\,\alpha_j$ and substitute them in the MCA equations. They become

$$\sum_{\ell=1}^m \rho_{j\ell}\theta_\ell = \lambda\theta_j\,.$$

This has m solutions for θ, the eigenvectors of $R=\{\rho_{j\ell}\}$. Thus each bi-linearizing solution produces m solutions to the MCA equations, each with the same induced correlation matrix. Less data production, consequently.

Moreover suppose $\theta_{js}\alpha_j$ are the MCA solutions corresponding with the linearizing scores, and suppose γ_j are the scores for another MCA solution. If the eigenvalue for scores γ_j is different from the eigenvalue of the correlation matrix induced by the α_j, then

$$\sum_{j=1}^m \theta_{js}\alpha_j'\gamma_j = 0$$

for all $s=1,\cdots,m$, which will generally be the case only if $\alpha_j'\gamma_j=0$ for all $j=1,\cdots,m$. Thus all other MCA solutions are *strongly orthogonal* to the α_j, in the sense that each piece is orthogonal. A single set of bi-linearizing scores already means that we can use L and P to transform H to the form $R_1\oplus R_2$, with R_1 the correlation matrix induced by the scores, and with R_2 the "residuals". Bi-linearizability of order p means that there are only

p induced correlation matrices, with only mp principal components. LPV exactly diagonalizes H.

If we want to compare the MCA solution to the approximate LPV solution, then we can compute $K'LPV$, which will have correlations between the two systems of solutions. There are examples in [11]. We find remarkable results, which can be understood quite easily by keeping the standardized multinormal in the back of our minds. The first set of bi-linearizing scores are the zero-degree Hermite polynomials. They give an induced correlation matrix with all elements equal to one (not really a correlation matrix, because the corresponding transformed variables are not centered). This bi-linearizing set occurs in any MCA, and gives one eigenvalue equal to m and $m-1$ eigenvalues equal to zero. In the multinormal the second set of bi-linearizing scores are the first degree polynomials, i.e. the identity transformation. The induced correlation matrix is the correlation matrix of the underlying multinormal, and we have m eigenvalues taken from that correlation matrix. Then the second degree polynomials, corresponding with quadratic transformations, induce the correlation matrix $R^{(2)}$, which consists of the squares of the correlation coefficients. And so on.

Empirically we find that the MCA solutions corresponding with the largest eigenvalues, and those corresponding with the smallest eigenvalues, are found by both eigen-analysis and LPV diagonalization. But LPV gives the eigenvalues in a natural order. First we get the m trivial ones, corresponding with the zero-degree polynomial. Then the m eigenvalues corresponding with the first induced correlation matrix, and so on. In an ordinary MCA we typically find that the largest nontrivial transformation corresponds with the dominant eigenvalue of the first induced correlation matrix, while the second largest eigenvalue is actually the largest eigenvalue of the second induced correlation matrix. If we plot the two transformations, we see a quadratic structure, the famous *horseshoe,* or, in French, the *effect de Guttman.* Compare [41] for more information, and [37] for a nonparametric (ordinal) explanation of horseshoes. Horseshoes are not inevitable. There are multiunormal examples in which the first two eigenvalues come from the same (linear) correlation matrix. If we mix two multinormals with correlations that are opposite in sign, then the eigenvalues corresponding with the odd powers of the correlation coefficients disappear, and the dominant solutions can be both quadratic.

9. Functions of Correlation Coefficients

In the bivariate case we could find the bi-linearizing scores by maximizing the correlation coefficient. In the case of $m > 2$ variables we can find the scores, if they exist, by making the correlation ratios equal to the squared

correlation coefficients, or by using the LPV plane rotations to eliminate the appropriate off-diagonal elements.

But let us go back now to the situation in which we do not necessarily assume bi-linearizability. It may still be interesting, for data analysis reasons, to find systems of scores with various optimal properties. We discuss a general algorithm which can be used for this purpose. It is explained in greater detail in [14].

For given scores α_j we can compute induced correlation coefficients between our m variables. Take any function $\tau(\bullet)$ of these correlation coefficients $\rho_{k\ell}$, i.e. any function of the correlation matrix. In [14] this is called an *aspect* of the correlation matrix. We define optimal scaling techniques by maximizing (or minimizing) aspects $\tau(\bullet)$ over the scores α_j. Each aspect defines a different technique, and each choice of the subspaces $\mathcal{K}_j$ defines a different special case of a technique. The $\mathcal{K}_j$ can be defined by polynomials, or splines, or dummies, with varying degrees on various knot-sequences. Actually, let us make the easy generalization here to convex cones $\mathcal{K}_j$.

From the algorithmic point of view quite a few things can be said about the problem of maximizing $\tau(\bullet)$. First let us suppose that the aspect is convex as a function of R. Then, with γ another set of scores,

$$\tau(R(\alpha)) \geq \tau(R(\gamma)) + \text{ tr } G(\gamma)(R(\alpha) - R(\gamma)),$$

where $G(\bullet)$ is the matrix of partial derivatives (or an element of the subgradient) of $\tau(\bullet)$. Now let us maximize, in each step of the algorithm, the right hand side of this expression over α, taking $\gamma = \alpha^{(s)}$, our current best guess of the scores. The maximizer is $\alpha^{(s+1)}$. It follows that

$$\begin{aligned}\tau(R(\alpha^{(s+1)})) &\geq \tau(R(\alpha^{(s)})) + \text{ tr } G(\alpha^{(s)})(R(\alpha^{(s+1)}) - R(\alpha^{(s)})) \\ &\geq \tau(R(\alpha^{(s)})) + \text{ tr } G(\alpha^{(s)})(R(\alpha^{(s)}) - R(\alpha^{(s)})) = \tau(R(\alpha^{(s)})).\end{aligned}$$

Thus we increase the aspect, and we can use general results [15] to show that this leads to a convergent algorithm.

In each sub-step of the algorithm we have to maximize $\alpha'\tilde{C}^{(s)}\alpha$, where $\tilde{C}_{j\ell}^{(s)} = g_{j\ell}(\alpha^{(s)})C_{j\ell}$, over all α with $\alpha_j' D_j \alpha_j = 1$, and perhaps cone-constraints of the form $\alpha_j \in \mathcal{K}_j$. We cycle over the α_j, and update each in turn by

$$\alpha_j \leftarrow \mathcal{P}_j(D_j^{-1}\sum_{\ell \neq j} C_{j\ell}^{(s)}\alpha_\ell),$$

where $\mathcal{P}(\bullet)$ projects on the cone.

Ways to speed up and simplify this basic algorithm are discussed in [14]. Clearly it can be used on a very general class of aspects. We can show, for

instance, that the sum of the p largest eigenvalues, the squared multiple correlation coefficient of one variable with the rest, the log-determinant, and many other aspects are indeed convex functions of the correlation matrix.

10. Consequences of bi-linearizability

Let us forget about cone constraints for the time being, and generalize the class of aspects to functions of both the correlation coefficients and the correlation ratios. The stationary equations for maximizing $\tau(\bullet)$ are (assuming differentiability)

$$\sum_{\ell\neq j}^{m}\frac{\partial\tau}{\partial\rho_{j\ell}}C_{j\ell}\alpha_\ell+\sum_{\ell\neq j}^{m}\frac{\partial\tau}{\partial\eta^2_{j\ell}}C_{j\ell}D_\ell^{-1}C_{lj}\alpha_j=\lambda_j D_j\alpha_j.$$

The λ_j are Lagrange multipliers, taking care of the normalization of the scores. If the scores α_j linearize the bivariate regressions, then they solve the stationary equations with

$$\lambda_j=\sum_{\ell\neq j}^{m}\frac{\partial\tau}{\partial\rho_{j\ell}}\rho_{j\ell}+\sum_{\ell\neq j}^{m}\frac{\partial\tau}{\partial\eta^2_{j\ell}}\rho^2_{j\ell}.$$

This result is interesting, because it shows that bi-linearizing systems give stationary points, no matter what the aspect is (it does not even have to be a convex function of the correlations). Or, to put it differently, if bi-linearizing scores exists, then optimizing any aspect will find them. We need some qualifications here, because the stationary point may not be an actual maximum, but essentially this strong corollary of bi-linearizability guarantees a solution with is invariant over choice of aspect. There are already many programs which maximize functions of the form $\tau(\bullet)$, such as MCA, ACE, etc. If bi-linearizable scorings exist, they will find them.

11. Model oriented approach

The correspondence analysis based techniques are primarily exploratory in character, at least if one believes in the distinction between exploratory and confirmatory [20]. Nevertheless, bi-linearizability and strained multinormality are restrictive models, which can be good or bad descriptions of an observed Burt matrix. It consequently makes sense to look at the fit of the model with the usual statistical large sample techniques.

Suppose the α_j linearize the bivariate regressions. Complete the scores to matrices $A_j = (\ \alpha_j \quad | \quad \overline{A}_j\ )$, such that $A_j' D_j A_j = I$. Then

$$A_j' C_{j\ell} A_\ell = \begin{pmatrix} \rho_{jl} & 0 \\ 0 & \overline{A}_j' C_{j\ell} \overline{A}_\ell \end{pmatrix}.$$

Solving the equations gives us the parametric model (for the joint bivariate marginals)

$$D_j^{-1} C_{j\ell} D_\ell^{-1} = A_j \begin{pmatrix} \rho_{j\ell} & 0 \\ 0 & ?_{jl} \end{pmatrix} A_\ell'.$$

This can be done by weighted least squares, applied directly to the bivariate marginals. We fit the parameters A_j, as well as the parameters ρ_{jl} and $?_{j\ell}$. For bi-linearizability of order p we can strengthen the model to

$$D_j^{-1} C_{j\ell} D_\ell^{-1} = A_j \Delta_{j\ell} A_\ell',$$

with $\Delta_{j\ell}$ a diagonal matrix. For strained multinormality we combine this with no higher-order interactions, and we can even use likelihood methods. Although these techniques are fairly straightforward in principle, they involve tedious delta-method type calculations, and they have not been implemented so far.

12. Two-step techniques

There is another way in which we can combine classical inference with optimal scaling. A useful statistical procedure seems to be the two-step technique. First we scale the variables by trying to bi-linearize the regressions. Then we apply standard techniques to the induced correlation coefficients. Such standard technqiues can be multiple regression, principal component analysis, LISREL, etc. But what about the asymptotic normal distribution of the induced correlations ?

The nice result, again generalizing Pearson [35], is that for linearizable distributions the asymptotic normal distribution is the same as the one we would derive if the scores had been known (fixed, not dependent on the data). This is discussed in detail in [13]. The reason is simple. Let us look at the discrete case. The Burt table is a function of the profile frequencies, i.e. of the cell entries in the multivariate table. Collect them in the vector p. We compute our statistics on the basis of the correlation coefficients, which depend on the scores, which depend on the p. Suppose statistic $\tau(\bullet)$ is differentiated with respect to p. We find

$$\frac{\partial \tau}{\partial p} = \sum_{j=1}^m \sum_{\ell=1}^m \frac{\partial \tau}{\partial \rho_{j\ell}} \frac{\partial \rho_{j\ell}}{\partial p}.$$

Now

$$\frac{\partial \rho_{j\ell}}{\partial p} = \frac{\partial \rho_{j\ell}}{\partial \alpha_j}\frac{\partial \alpha_j}{\partial p} + \frac{\partial \rho_{j\ell}}{\partial \alpha_\ell}\frac{\partial \alpha_\ell}{\partial p} + \frac{\partial \rho_{j\ell}}{\partial C_{j\ell}}\frac{\partial C_{j\ell}}{\partial p}.$$

If the scores bi-linearize the regressions, then the first two terms on th right hand side disappear, and the last expression becomes simply

$$\frac{\partial \rho_{j\ell}}{\partial p} = \alpha_j' \frac{\partial C_{j\ell}}{\partial p}\alpha_\ell.$$

But this means that the partials of the scores with respect to p do not enter the delta-method calculations, and thus we can treat (for statistical purposes) the scores as fixed and known. We know since Isserlis [28] how to compute the asymptotic distribution of correlation coefficients. This means that standard error calculations from ordinary regression, factor analysis, and LISREL programs are still (first-order) correct.

So let us consider any method which consists of scaling the variables first, using any technique which optimizes an aspect of the correlation coefficients and correlation ratios, followed by a classical multivariate analysis technique on the scaled variables. Such a two-step method gives unbiased estimates of the structural parameters under the assumption of bilinearizability, while the usual methods to compute standard errors are still asymptotically valid. Moreover they give the same result as if we had fixed and know scores. If we compare this with assuming multivariate normality, we gain a lot in terms of bias, and we do not seem to lose anything in terms of precision. There is a (first-order) free lunch here. It will be interesting to find out in how far these results are borne out by small-sample comparisons.

References

[1] P. Bekker and J. de Leeuw. Relation between variants of nonlinear principal component analysis. In J.L.A. van Rijckevorsel and J. de Leeuw, editors, *Component and Correspondence Analysis*. Wiley, Chichester, England, 1988.

[2] J.P. Benzécri. *Histoire et préhistoire de l'Analyse des Données*. Dunod, Paris, France, 1982.

[3] J.P. Benzécri. *Correspondence analysis handbook*. Marcel Dekker, Inc., New York, New York, 1992.

[4] G. E. P. Box and D. R. Cox. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society*, B26:211–252, 1964.

[5] L. Breiman and J.H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–619, 1985.

[6] A. Buja. Remarks on functional canonical variates, alternating least squares methods, and ace. *Annals of Statistics*, 18:1032–1069, 1990.

[7] C. Burt. The factorial analysis of qualitative data. *British Journal of Statistical Psychology*, 3:166–185, 1950.

[8] C.M. Cuadras. Probability distributions with given multivariate marginals and given dependence structure. *Journal of Multivariate Analysis*, 42:51–66, 1992.

[9] J. Dauxois and A. Pousse. *Les analyses factorielles en calcul des probabilites et en statistique: essai d'étude synthetique*. PhD thesis, Université Paul-Sabatier, Toulouse, France, 1976.

[10] J. de Leeuw. *Canonical analysis of categorical data*. PhD thesis, University of Leiden, The Netherlands, 1973. Republished in 1985 by DSWO-Press, Leiden, The Netherlands.

[11] J. de Leeuw. Nonlinear principal component analysis. In H. Caussinus et al., editor, *COMPSTAT 1982*, Vienna, Austria, 1982. Physika Verlag.

[12] J. de Leeuw. On the prehistory of correspondence analysis. *Statistica Neerlandica*, 37:161–164, 1983.

[13] J. de Leeuw. Multivariate analysis with linearizable regressions. *Psychometrika*, 53:437–454, 1988.

[14] J. de Leeuw. Multivariate analysis with optimal scaling. In S. Das Gupta and J. Sethuraman, editors, *Progress in Multivariate Analysis*, Calcutta, India, 1990. Indian Statistical Institute.

[15] J. de Leeuw. Block-relaxation algorithms in statistics. Preprint 120, UCLA Statistics, Los Angeles, 1993.

[16] J. de Leeuw and J. van Rijckevorsel. Beyond homogeneity analysis. In J.L.A. van Rijckevorsel and J. de Leeuw, editors, *Component and Correspondence Analysis*. Wiley, Chichester, England, 1988.

[17] R.A. Fisher. *Statistical Methods for research workers*. Oliver and Boyd, London, England, 1925.

[18] R.A. Fisher. The precision of discriminant functions. *Annals of Eugenics*, 10:422–429, 1940.

[19] H. Gebelein. Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *Zeitschrift fuer Angewandte Mathematik and Mechanik*, 21:364–379, 1941.

[20] A. Gifi. *Nonlinear multivariate analysis*. Wiley, Chichester, England, 1990.

[21] Z. Gilula and S.J. Haberman. Canonical analysis of contingency tables by maximum likelihood. *Journal of the American Statistical Association*, 81:780–788, 1986.

[22] Z. Gilula and Y. Ritov. Inferential ordinal correspondence analysis: motivation, derivation and limitations. *International Statistical Review*, 58:99–108, 1990.

[23] L.A. Goodman. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables (with discussion). *International Statistical Review*, 54:243–309, 1986.

[24] L.A. Goodman. Measures, models, and graphical displays in the analysis of cross-classfied data (with discussion). *Journal of the American Statistical*

Association, 86:1085–1138, 1991.

[25] M.J. Greenacre. *Theory and applications of correspondence analysis*. Academic Press, New York, New York, 1984.

[26] L. Guttman. The quantification of a class of attributes: A theory and method of scale construction. In P. Horst et al., editor, *The prediction of personal adjustment*. Social Science Research Council, New York, New York, 1941.

[27] H.O. Hirschfeld. A connection between correlation and contingency. *Proceedings Cambridge Philosophical Society*, 31:520–524, 1935.

[28] L. Isserlis. On certain probable errors and correlation coefficients of multiple frequency distributions with skew regression. *Biometrika*, 11:185–190, 1916.

[29] J.T.A. Koster. *Mathematical aspects of multiple correspondence analysis for ordinal variables*. PhD thesis, Erasmus University, Rotterdam, The Netherlands, 1989. Also published in 1989 by DSWO-Press, Leiden, The Netherlands.

[30] H.O. Lancaster. The structure of bivariate distributions. *Annals of Mathematical Statistics*, 29:719–736, 1958.

[31] H.O. Lancaster. *The chi-squared distribution*. Wiley, New York, New York, 1969.

[32] K. Maung. Discriminant analysis of Tocher's eye colour data for Scottish school children. *Annals of Eugenics*, 11:64–76, 1941.

[33] K. Maung. Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colour of Scottish school children. *Annals of Eugenics*, 11:189–223, 1941.

[34] S. Nishisato. *Analysis of categorical data: dual scaling and its applications*. University of Toronto Press, Toronto, Canada, 1980.

[35] K. Pearson. On certain points connected with scale order in the case of a correlation of two characters which for some arrangement give a linear regression line. *Biometrika*, 5:176–178, 1906.

[36] A. Renyi. On measures of dependence. *Acta Mathemtaica Academy Sciences Hungary*, 10:441–451, 1959.

[37] B. F. Schriever. *Order Dependence*. PhD thesis, University of Amsterdam, The Netherlands, 1985. Also published in 1985 by CWI, Amsterdam, The Netherlands.

[38] M. Tenenhaus and F.W. Young. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50:91–119, 1985.

[39] P. G. M. van der Heijden, A. de Falguerolles, and J. de Leeuw. A combined approach to contingency table analysis with correspondence analysis and log-linear analysis (with discussion). *Applied Statistics*, 38:249–292, 1989.

[40] P. G. M. van der Heijden and J. de Leeuw. Correspondence analysis used complimentary to loglinear analysis. *Psychometrika*, 50:429–447, 1985.

[41] J. L. A. van Rijckevorsel. *The application of fuzzy coding and horseshoes in multiple correspondence analysis*. PhD thesis, University of Leiden, The Netherlands, 1987. Also published in 1987 by DSWO-Press, Leiden, The

Netherlands.

[42] F.W. Young. Quantitative analysis of qualitative data. *Psychometrika*, 46:357–388, 1981.

[43] G. Udny Yule. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75:107–170, 1912.