18
Correspondence Analysis Techniques

JAN DE LEEUW and DEBORAH Y. WANG University of California, Los Angeles, California

GEORGE MICHAILIDIS University of Michigan, Ann Arbor, Michigan

1. CORRESPONDENCE ANALYSIS

Correspondence analysis can be introduced in many different ways, which is probably the reason why it was reinvented many times over the years. We do not repeat the various derivations in this chapter; instead we refer to the extensive discussions in the books by Greenacre (1984), Gifi (1990), and Benzécri (1992).

Usually, correspondence analysis is motivated in graphical language. It is often said, in this context, that "A picture is worth a thousand numbers." Complicated multivariate data are made more accessible by displaying the main regularities of the data in scatterplots. This graphical approach is outlined in considerable detail in the books mentioned above. We merely give a brief introduction, which differs in some important aspects from earlier ones because it emphasizes the *graph plot* and the *star plots* (defined below). This type of introduction, which discusses the data analysis technique as a graph-layout method, was first dicussed in the review articles by Hoffman and de Leeuw (1992) and Michailidis and de Leeuw (1998). We think it nicely captures the essential geometric characteristics of the technique.

We have to choose one of the many names the technique has had over the years (see de Leeuw (1973, 1983) for a historical overview). The most widely used name seems to be (multiple) correspondence analysis or MCA, and this is what we shall use in this review as well. By considering various generalizations, we actually review a very broad class of techniques under the MCA label.

1.1 Data

MCA starts with n observations on m categorical variables, where variable j has k_j categories (possible values). Using categorical variables causes no real loss of generality: so-called *continuous* variables are merely categorical variables with a large number of numerical categories. We use K for the total number of categories over all variables.

The data are coded as m *indicator matrices* or *dummies* G_j, where G_j is a binary $n \times k_j$ matrix with exactly one nonzero element in each row i (indicating in which category of variable j observation i falls). The $n \times K$ matrix $G = (G_1 | \dots | G_m)$ is called the *indicator supermatrix*.

1.2 Graph Layout

One can represent all information in the data by a bipartite graph with $n + K$ vertices and nm edges. Each edge connects an object and a category. Thus the n vertices corresponding to the objects all have degree m, and the K vertices corresponding to the categories have varying degrees, equal to the number of objects in the category. The indicator supermatrix G is the *adjacency matrix* of the graph. The idea of presenting categorical data as a graph is not new, of course, but the idea of interpreting a data analysis technique as a graph layout method does not seem to have been studied before.

We can make a drawing or layout of the graph by placing the vertices at $n + K$ locations in the plane, or, more generally, in R^p. If we then draw the nm edges, the resulting picture will generally be more informative and more aesthetically pleasing if the edges are short. In other words, if objects are close to the categories they fall into, and categories are close to the objects falling in them. Thus we want to make a *graph plot* that "minimizes the amount of ink," i.e. the total length of all edges.

There is a substantial literature in computer science about methods and criteria to draw graphs (di Battista et al., 1994). Graph drawing algorithms for bipartite graphs that emphasize minimizing edge crossing are discussed by Eades and Wormald (1994). MCA, i.e., our "minimum ink" criterion, is closely related to the force-directed or spring algorithms first introduced by Eades (1984). Many of the criteria discussed in the computer science literature lead to NP-complete problems, i.e., they are computationally infeasible even for fairly small problems. Our edge-length algorithm is designed to be practical even for very large bipartite graphs.

Actually, we will minimize the total *squared* length of the edges. The reasons for choosing the square are the classical ones.

> Of all the principles that can be proposed for this purpose, I think there is none more general, more exact, or easier to apply, that that which we have used in this work; it consists of making the sum of squares of the errors a *minimum*. By this method, a kind of equilibrium is established among the errors which, since it prevents the extremes from dominating, is appropriate for revealing the state of the system which most nearly approximates the truth.
>
> Legendre (1805), quoted by Stigler (1986, p. 13)

In order to implement a useful algorithm, we also need a *normalization constraint*. This is needed because we want distances between vertices that are connected to be small, but we do not require distances between edges that are not connected to be large. Merely minimizing the amount of ink, without requiring a normalization, can be done by collapsing the drawing into a single point.

1.3 Least-Squares Criterion

To formalize our "minimum ink" criterion in a convenient way, we use the indicator matrices G_j. If the $n \times p$ matrix X has the locations of the object vertices in $\mathbb{R}^p$, and Y_j has the location of the k_j category vertices of variable j, then the squared length of the n edges for variable j is

$$\sigma_j(X, Y_j) = \mathbf{SSQ}(X - G_j Y_j) \tag{1}$$

where **SSQ**() is short for the sum of squares. The corresponding graph drawing, with $n + k_j$ vertices and n edges, is known as the *star plot* for variable j. The *graph plot* is the union (overlay) of the m star plots.

The squared edge length over all variables is

$$\sigma(X, Y) = \sum_{j=1}^{m} \mathbf{SSQ}(X - G_j Y_j) \tag{2}$$

and this is the function we want to minimize. The book by Gifi (1990) is mainly about many different versions of this minimization problem, where the differences are a consequence of various *restrictions* imposed on the quantifications Y_j.

As we said earlier, minimizing equation (2) without any restrictions on the vertex locations is not possible. Or, more appropriately, it is too easy. We just collapse all vertices into a single point, and we use no ink at all. This means that in order to get a nontrivial solution, we have to impose some form of *normalization*. There are two obvious ways to normalize, defining, say, MCA_1 and MCA_2. In MCA_1 we require that the columns of X add up to zero, and are *orthonormal*, i.e., satisfy $mX'X = I$. In MCA_2 we normalize Y, i.e., we require that $u'DY = 0$ and $Y'DY = I$. Here u is a vector with all elements equal to +1, and D is the $K \times K$ diagonal matrix with the marginal frequencies of all m variables on the diagonal.

We emphasize that there are no compelling reasons, except for computational convenience, to choose these particular normalizations. Specifically, introducing additional dimensions by requiring orthonormality is in many respects not completely satisfactory. This was indicated in Guttman's classical 1941 MCA paper.

> we should be tempted to try a "multiple factor" analysis. But the present rationale was devised specifically for a "single factor" analysis and does not necessarily carry over to the other case. It may be quite a different task to devise a rationale for "multiple factor" analysis of attributes
>
> (Guttman, 1941, p. 332)

1.4 Eigenvalue Formulations

One of the reasons why squared edge lengths are so appealing is that the MCA_1 and MCA_2 problems we are trying to solve are basically eigenvalue problems. We discuss this in some detail, following Gifi (1990).

First we define some useful matrices. Define the $k_j \times k_\ell$ matrix $C_{j\ell} = G_j'G_\ell$. Matrix $C_{j\ell}$ is the *cross table* or *contingency table* of variables j and ℓ. Thus $D_j = C_{jj}$, where D_j is the diagonal matrix with the univariate marginals of variable j on the diagonal. The $K \times K$ supermatrix C is known in the correspondence analysis literature as the *Burt matrix*, after Burt (1950). Write $CY = mDY\Xi$ for the generalized eigenvalue problem associated with the Burt matrix.

We also define $P_j = G_jD_j^{-1}G_j'$, then P_j is the *between-category* projector, which transforms each vector in $\mathbb{R}^n$ into a vector in $\mathbb{R}^n$ with category means. Moreover $Q_j = I - P_j$ transforms each vector into a *within-category* vector of deviations from category means.Write $P_\star$ for the average of the P_j, and write Θ for the diagonal matrix of eigenvalues of $P_\star$.

THEOREM 1: *Suppose $(\hat{X}, \hat{Y})$ solves either the MCA_1 or the MCA_2 problem. Then*

$$P_{\star}\hat{X} = \hat{X}\Lambda \tag{3a}$$

$$C\hat{Y} = mD\hat{Y}\Lambda \tag{3b}$$

where $\Theta = \Xi = \Lambda$.

Proof. We first analyze MCA_1, in which X is normalized by $mX'X = I$, and Y is free. Define $\sigma(X, \bullet)$ as the minimum of $\sigma(X, Y)$ over all Y. Clearly the minimum is attained for

$$\hat{Y}_j = D_j^{-1}G_j'X \tag{4}$$

i.e., by locating a category quantification in the centroids of the objects in that category. We see that

$$\sigma(X, \bullet) = m \,\mathbf{tr}\, X'(I - P_{\star})X \tag{5}$$

Clearly we minimize $\sigma(X, \bullet)$ over $mX'X = I$ by choosing $\hat{X}$ equal to the eigenvectors corresponding to the p largest eigenvalues of $P_{\star}$. Thus $P_{\star}\hat{X} = \hat{X}\Theta$ for MCA_1. Also, from equation (4), we see $G\hat{Y} = mP_{\star}\hat{X} = m\hat{X}\Theta$ and thus $C\hat{Y} = mG'\hat{X}\Theta = mD\hat{Y}\Theta$. This proves equation (3b), with $\Theta = \Lambda$.

We now travel the other route, and tackle MCA_2. Define $\sigma(\bullet, Y)$ as the minimum of $\sigma(X, Y)$ over all X. This minimum is attained by

$$\hat{X} = \frac{1}{m}\sum_{j=1}^{m} G_j Y_j \tag{6}$$

i.e., each object is located in the centroid of the m categories that it is in. Then

$$\sigma(\bullet, Y) = \mathbf{tr}\, Y'\left(D - \frac{1}{m}C\right)Y \tag{6}$$

and the minimum over $Y'DY = I$ is attained by finding $\hat{Y}$, the eigenvector corresponding with the largest eigenvalues of the eigen-problem $CY = mDY\Xi$. Now we use equation (6) to derive $mG'\hat{X} = C\hat{Y} = mD\hat{Y}\Xi$ and thus $mP_{\star}\hat{X} = G\hat{Y}\Xi = m\hat{X}\Xi$. This proves equation (3a), with $\Xi = \Lambda$.

□

There are several aspects of the proof which deserve some additional attention. Equation (4) is called the *first centroid principle*, and equation (6) is the *second centroid principle.* The first centroid principle shows clearly how the star plots get their name in MCA_1. Category vertices are in the centroid of the vertices of the objects in the category, and if we have a clear separation of the k_j categories, we see k_j stars in $\mathbb{R}^p$. This also shows that in MCA$_1$ the category vertices are in the convex hull of the object vertices; they form a more compact cloud. In MCA$_2$, it is the other way around.

Of course the theorem does not say the solutions to MCA$_1$ andMCA$_2$ are *identical*, they are in fact merely *proportional*. In MCA$_1$ we see from equation (4) that $\hat{Y}'D\hat{Y} = m\hat{X}'P_\star\hat{X} = \Lambda$. In the same way, in MCA$_2$, $m\hat{X}'\hat{X} = \Lambda$.

In fact, this leads to one last important construct in MCA$_1$. The matrix $\hat{Y}_j'D_j\hat{Y}_j = \hat{X}'P_j\hat{X}$ is known as the discrimination matrix. It is equal to the between-category dispersion matrix of variable j, i.e., to the size of the stars for that variable. The average *discrimination matrix* is equal to Λ, the diagonal matrix of eigenvalues. Since $P_\star$ is the average of m orthogonal projectors, we have that $\Lambda \leq I$. This can also be seen from the fact that each element of Λ is the average, over variables, of the ratio of the between-category variance and the total variance.

A more direct proof of the equivalence of the eigenvalue problems for MCA$_1$ and MCA$_2$ is possible by starting with the singular value decomposition problems $GY = mXM$ and $G'X = DYM$, which immediately gives $CY = mDYM^2$ and$P_\star X = XM^2$. In the French approach, this is expressed by saying that MCA is correspondence analysis applied to the "tableau disjonctif complet" G.

2. ASPECTS OF MULTIVARIABLES

We think the graphical or geometric approach to MCA outlined above is a valid and interesting way to introduce and discuss the technique, but in this paper we go back to the more analytical formulations first proposed by Guttman (1941). We unify and extend results in previous papers (de Leeuw, 1982, 1983, 1988, 1990, 1993).

One major reason for preferring the analytical approach is that it generalizes easily to different criteria and to more general (infinite-dimensional) situations. Another reason is that a more convincing treatment of multidimensional quantification becomes possible.

In the alternative (nongeometric) formulation we discuss here, the emphasis is on finding transformations of variables and on the construction of scales. This makes it easier to relate correspondence analysis to classical multivariate analysis techniques, such as principal components analysis.

2.1 Aspects and Feasible Transformations

In this paper, a variable is an element h of some Hilbert space $\mathcal{H}$. This could just be the space of vectors with n elements, but it could also be the space of random variables with finite variances. In fact, this is precisely the reason why we use $\mathcal{H}$; our subsequent formulas apply without modification to the "population" case in which our variables can be continuous random variables.

A *multivariable* is just a mapping of an index set $\mathcal{J}$ into $\mathcal{H}$, i.e., each $j \in \mathcal{J}$ corresponds with a variable h_j. A finite number of the h_j can be collected in a "matrix" $H = \{h_1, \dots, h_J\}$. Observe that H, interpreted as a matrix, does not have a well-defined number of rows. Nevertheless it is straightforward to define the matrix operations we need in a consistent way. Matrix $H'H$ contains inner products of the elements of H, while $Hu \in \mathcal{H}$ is a linear combination of the h_j. If $f \in \mathcal{H}$, then $f'H$ has inner products of f with the elements of H.

Informally, an *aspect* of a multivariable is a well-defined function that is used to measure how well the multivariable sätisfies some, presumably desirable, criterion. Because there are many such criteria, there are many different aspects.

DEFINITION 1: An *aspect* of a multivariable is a real-valued function ϕ defined on the set of multivariables on a given index set $\mathcal{J}$.

The idea of using aspects to define multivariate analysis techniques was introduced in de Leeuw (1990). The basic idea is simple. In many situations, especially in the social and behavioral sciences, we do not know precisely how to *express* our variables. Thus in a regression model the dependent variable "income," for instance, might be dollar-income, but it might also be log-dollar-income, or even some unknown monotonic *transformation* of dollar-income. For other variables there may be missing information on some of the observations, and we get different expressions for different *imputations.* In yet another scenario, there may be *latent variables,* which are completely unobserved and only defined by their place in the model. Finally, some variables may be *ordinal* or *nominal,* and they can be incorporated in a correlational analysis only after *quantification.*

Of course not all quantifications are *feasible.* If we impute missing data, we want the imputed variable to be equal to the data in the nonmissing part. If we transform or quantify an ordinal variable, we want the transformation to be monotone. If we quantify the nominal variable "religion," we want all protestants to get the same value, all buddhists to get the same value, and so on.

Thus the basic problem in this particular approach to multivariate analysis is to select an aspect, and to investigate how this aspect varies over all feasible transformations, quantifications, or imputations of the variables. One particular approach is to study what the maximum value of the aspect is. We will formalize this optimization problem for the case in which the feasible transformations are finite-dimensional subspaces of $\mathcal{H}$. This can easily be generalized to infinite-dimensional subspaces, and even to convex cones (see de Leeuw, 1990).

In the psychometric literature this approach is known as "optimal scaling," a term due to Darrell Bock. Thus optimality is defined in terms of the aspect, and each aspect defines its own form of optimal scaling.

2.2 Examples of Aspects

In a well-known paper, Kettenring (1971) uses a similar "aspect" approach to extend canonical correlation analysis to three or more sets of variables. Versions of these ideas were proposed even earlier by Steel (1951) and Horst (1961, 1965). Kettenring's contribution is also discussed in the book by Gnanadesikan (1977, pp. 69–81). From a slightly different (psychometric) angle, other related aspects were discussed by van de Geer (1984) and ten Berge (1988).

Proposer	Name	Description
Horst	SUMCOR	Sum of $r_{j\ell}$
Kettenring	SSQCOR	Sum of $r_{j\ell}^2$
Horst	MAXVAR	Largest eigenvalue of R
Kettenring	MINVAR	Smallest eigenvalue of R
Steel	GENVAR	Determinant of R

All of the correlational aspects in the table, except SUMCOR, are actually also *eigenvalue aspects*, i.e., they are functions of the eigenvalues of the correlation matrix R. SSQCOR is the sum of squares of the eigenvalues of the correlation matrix. Because the sum of the eigenvalues of the correlation matrix is a constant, using the SSQCOR aspect is also identical to looking at the variance of the eigenvalues. GENVAR, the determinant of the correlation matrix, is the product of the eigenvalues.

The correlational aspects in the table are all measures of "interdependence" of the variables, there is no notion of dependence or causal ordering in any of them. De Leeuw (1990) observed that we may as well include aspects such as the squared multiple correlation (SMC) coefficient between one variable and the rest, or we could use the sum or sum-of-squares of one or more canonical correlation coefficients for a given partitioning of the variables into sets. Also, generalizations such as the sum of the correlation coefficients to the power s, or the absolute value of the correlation coefficients to the power s could be considered. MAXVAR and MINVAR can be generalized by considering the sum of the p largest or smallest eigenvalues. The multinormal negative log-likelihood can also be analyzed as a correlational aspect. This is

$$\phi(R) = \min_{\Gamma} \log|\Gamma| + tr\ \Gamma^{-1}R \tag{8}$$

where the minimization is over a set of model-constrained correlation matrices (for instance, all matrices satisfying the Spearman two-factor model). We do not go into details here, but clearly the notion of a correlational aspect is very general.

In de Leeuw (1988) another aspect, which is noncorrelational, was studied in some detail. We form the difference of the sum of all correlation ratios and corresponding squared correlation coefficients. Thus the aspect is

$$\phi(H) = \sum_{j=1}^{m}\sum_{\ell=1}^{m}\{\eta_{j\ell}^2 - r_{j\ell}^2\} \tag{9}$$

Minimizing such an aspect means aiming for transformations that maximize linearity of the bivariate regressions.

Another noncorrelational aspect is the Box–Cox version of equation (8). This adds a penalty term to the log-likelihood equal to the logarithm of the transformation Jacobian. It penalizes for making the transformations too flat. Because of this there is no need to normalize, and we can use covariances instead of correlations. The aspect is given by

$$\phi(S) = \min_{\Sigma} \log|\Sigma| + tr\ \Sigma^{-1}S - 2\sum_{j=1}^{m}\log \mathcal{D}h_j \tag{10}$$

The original reference is Box and Cox (1964), and use of this aspect in multivariate analysis has been analyzed in detail in Meijerink (1996). One reason why it makes sense for us to look at aspects at this level of generality is because there is a simple algorithm that allows us to optimize many of them. We will discuss this below. The other reason is that in some theoretically and perhaps also practically interesting situations we can show that the transformations we find are independent of the choice of the aspect. This will be discussed in Section 6.

3. MAXIMIZING ASPECTS

We have seen that different multivariate analysis techniques are associated with different aspects. Canonical analysis looks at aspects defined in terms of the canonical correlations, principal component analysis looks at eigenvalue aspects. Multiple regression uses the SMC, path analysis uses the sum of a number of SMCs, and so on. Some of the aspects we have considered do not correspond with classical techniques at all. It is, of course, interesting to discuss the problem of how to choose an aspect, but this is not what this chapter is about. We deal with the situation in which the client arrives in our office with an aspect, and asks us for a method to optimize it.

We need some additional notation. Suppose G_j is a basis for the subspace $\mathcal{H}_j$ of feasible transformations of variable j. Thus G_j consists of a finite number of elements, say k_j elements, of $\mathcal{H}$, collected in the "matrix" H. Previously, G_j was the indicator matrix of variable j, now it is more general. Suppose D_j is the diagonal matrix of order k_j with the squared lengths of the elements of G_j on the diagonal. An element of $\mathcal{H}_j$ can obviously be written as a linear combination of the elements of G_j, i.e., in the form $h_j = G_j\theta_j$. We write $\mathbf{NORM}(\theta_j)$ for the normalization of θ_j that satisfies $\theta_j'D_j\theta_j = 1$.

3.1 Majorization

The algorithms proposed in this paper are all of the majorization type. In a majorization algorithm we want to maximize $\phi(\theta)$ over $\theta \in \Theta$. Suppose $\psi(\theta, \xi)$ on $\Theta \times \Theta$, which we call the *majorization function*, satisfies

$$\phi(\theta) \geq \psi(\theta, \xi) \qquad \text{for all} \qquad \theta, \xi \in \Theta \tag{11a}$$

$$\phi(\theta) = \psi(\theta, \theta) \qquad \text{for all} \qquad \theta \in \Theta \tag{11b}$$

Thus, for a fixed ξ, $\psi(\bullet, \xi)$ is below ϕ, and it touches ϕ at the point $(\xi, \phi(\xi))$.

There are two key theorems associated with these definitions.

THEOREM 2: *If ϕ attains its maximum on Θ at $\hat{\theta}$, then $\psi(\bullet, \hat{\theta})$ also attains its maximum on Θ at $\hat{\theta}$.*

Proof. Suppose $\psi(\tilde{\theta}, \hat{\theta}) > \psi(\hat{\theta}, \hat{\theta})$ for some $\tilde{\theta} \in \Theta$. Then, by equations (11a and 11b), $\phi(\tilde{\theta}) \geq \psi(\tilde{\theta}, \hat{\theta}) > \psi(\hat{\theta}, \hat{\theta}) = \phi(\hat{\theta})$, which contradicts the definition of $\hat{\theta}$ as the maximizer of ϕ on Θ. □

THEOREM 3: *If $\tilde{\theta} \in \Theta$ and $\hat{\theta}$ maximizes $\psi(\bullet, \tilde{\theta})$ over Θ, then $\phi(\hat{\theta}) \geq \phi(\tilde{\theta})$.*

Proof. By equation (11a) we have $\phi(\hat{\theta}) \geq \psi(\hat{\theta}, \tilde{\theta})$. By the definition of $\hat{\theta}$ we have $\psi(\hat{\theta}, \tilde{\theta}) \geq \psi(\tilde{\theta}, \tilde{\theta})$. And by equation (11b) we have $\psi(\tilde{\theta}, \tilde{\theta}) = \phi(\tilde{\theta})$. Combining the three results gives the desired conclusion. □

If ϕ is bounded above on Θ, then the algorithm generates a bounded increasing sequence of function values, and thus it converges. Some mild continuity considerations are needed to actually show that the sequence of θ values converges as well (see de Leeuw (1990), or for a general discussion the book by Zangwill (1969)).

3.2 General Aspects

We shall maximize the aspect

$$\phi(G_1\theta_1, \ldots, G_m\theta_m)$$

over the θ_j with the normalizations conditions $\theta_j'D_j\theta_j = 1$ for all j. The notion of a general aspect, natural as it is, has not been discussed before.

THEOREM 4: *If $(\theta_1, \ldots, \theta_m)$ maximizes the aspect $\phi(H)$ over the normalized θ_j then*

$$G_j'\frac{\partial \phi}{\partial h_j} = \lambda_j D_j \theta_j \tag{12}$$

Proof. This just applies the chain rule to the Langrangian of the optimization problem. □

It follows that the Lagrange multipliers λ_j are given by

$$\lambda_j = h_j'\frac{\partial \phi}{\partial h_j} \tag{13}$$

THEOREM 5: *Suppose $\phi(H)$ is a convex and differentiable function of H which is bounded above. Then the algorithm $\mathcal{A}$ defined by*

$$\overline{\theta}_j^{(k)} = D_j^{-1}G_j'\frac{\partial \phi}{\partial h_j}\bigg|_{\theta=\theta^{(k)}} \tag{14a}$$

and

$$\theta_j^{(k+1)} = \mathbf{NORM}(\overline{\theta}_j^{(k)}) \tag{14b}$$

converges from any starting point.

Proof. Convexity implies that for all θ and $\tilde{\theta}$ we have the majorization

$$\phi(H(\theta)) \geq \phi(H(\tilde{\theta})) + \sum_{j=1}^{m} (\theta_j - \tilde{\theta}_j)' G_j' \frac{\partial \phi}{\partial h_j}\bigg|_{\theta=\tilde{\theta}} \tag{15}$$

Maximizing the majorization function on the right-hand side gives the algorithm in the theorem. □

3.3 Correlational Aspects

For correlational aspects we use the fact that the covariance of feasible transformations of variables j and ℓ can be written as the simple bilinear-form $\theta_j' C_{j\ell} \theta_\ell$. The variances are given by $\theta_j' D_j \theta_j$ and $\theta_\ell' D_\ell \theta_\ell$. We now state two theorems very similar to those in the previous section. These results are formalizations of the discussion of correlational aspects by de Leeuw (1990).

THEOREM 6: *If* $(\theta_1, \ldots, \theta_m)$ *maximizes the aspect* $\phi(R)$ *over all normalized* θ, *then*

$$\sum_{\ell=1}^{m} \frac{\partial \phi}{\partial r_{j\ell}} C_{j\ell} \theta_\ell = \mu_j D_j \theta_j \tag{16}$$

Proof. Use the chain rule, just as before. □

The Lagrange multipliers μ_j are now given by

$$\mu_j = \sum_{\ell=1}^{m} \frac{\partial \phi}{\partial r_{j\ell}} r_{j\ell} \tag{17}$$

Unfortunately, the majorization algorithm for correlational aspects is somewhat less simple. The main difference is that we now have to update a single θ_j at a time, and then recompute the aspect and its derivatives before we update the next θ_j.

THEOREM 7: *Suppose* $\phi(R)$ *is a convex and differentiable function of* R *which is bounded above. Then the algorithm* $\mathcal{A}$ *defined by*

$$\overline{\theta}_j^{(k)} = D_j^{-1} \left\{ \sum_{\ell=1}^{j-1} \frac{\partial \phi}{\partial r_{j\ell}} C_{j\ell} \theta_\ell^{(k+1)} + \sum_{\ell=j+1}^{m} \frac{\partial \phi}{\partial r_{j\ell}} C_{j\ell} \theta_\ell^{(k)} \right\} \tag{18a}$$

and

$$\theta_j^{(k+1)} = \textbf{NORM}\left(\overline{\theta}_j^{(k)}\right) \tag{18b}$$

converges from any starting point.

Proof. Suppose we update θ_j.The convexity of ϕ gives the majorization

$$\phi(R) \geq \phi(\tilde{R}) + \sum_{\substack{\ell=1 \\ \ell \neq j}}^{m} \frac{\partial \phi}{\partial r_{j\ell}} (r_{j\ell} - \tilde{r}_{j\ell})$$

Thus it obviously suffices to maximize

$$\sum_{\substack{\ell=1 \\ \ell \neq j}}^{m} \frac{\partial \phi}{\partial r_{j\ell}} \theta_j' C_{j\ell} \theta_\ell$$

over the normalized θ_j. This gives the update in the theorem. □

In de Leeuw (1990) several ways are discussed to simplify the above algorithm for correlational aspects. If we assume that $\partial\phi/\partial R$ is positive semidefinite for all R, then we can apply majorization a second time, and we find an algorithm that can update all θ_j in a single step, without having to recompute aspects and derivatives.

3.4 Convexity

The above theorems are useful if the aspects we study are convex, either in the transformed variables H or in the correlation matrix R. In de Leeuw (1990), it is shown that most interesting correlational and eigenvalue aspects are, indeed, convex. The key result used to prove convexity is the following lemma.

LEMMA 1: *Suppose $\psi(\theta, \xi)$ is convex in θ for every $\xi \in \Xi$. Then*

$$\phi(\theta) = \sup_{\xi \in \Xi} \psi(\theta, \xi)$$

is convex in θ as well.

Proof. See, for instance, Rockafellar, 1970, pp. 102–111. □

In particular, the first three aspects in Table 7 are convex in R, and the last two (which we usually want to minimize) are concave in R. Norms of R are convex, the SMC is convex, the sum of the p largest eigenvalues is convex, the multinormal log-likelihood in equation (8) is convex, and so on.

3.5 Canonical Correlation Aspects

It is observed in de Leeuw (1990) that the canonical correlations and most aspects based on them are not convex functions of the correlation coefficients. Thus, although they are correlational aspects, we cannot use the results based on convexity. Nevertheless successful algorithms based on canonical correlation aspects were tried out by Tijssen and de Leeuw (1989).

To some extent, however, reintroducing enough convexity is merely a matter of redefining the problem. Suppose we have three sets of variables. We then set up the analysis as if there are only three subspaces of $\mathcal{H}$, and we compute the correlations between elements of those three subspaces. Thus we do not transform each variable separately, we transform each set of variables with a feasible transformation. This brings us back to the correlational aspects, in particular those in the Kettenring table, applied to the 3×3 correlation matrix of the sets.

This is precisely the way in which the generalized canonical correlation program OVERALS, discussed in van der Burg et al. (1988), is fitted into the MCA loss function equation (2). We code the variables in the sets interactively, and then impose additivity restrictions on the quantifications. So far, applications of the generalized canonical correlation aspects are rare, and it is difficult to choose a natural set of invariants (and a corresponding aspect) if there are more than two multivariables.

4. THE LARGEST EIGENVALUE ASPECT

Suppose the aspect we want to maximize is the largest eigenvalue $\lambda_{max}(R)$. Thus, in a sense, we want to scale the variables in such a way that they are as one-dimensional as possible. It is of some interest that in recent numerical analysis and mathematical programming literature there is a great deal of interest in minimizing the largest eigenvalue of a parameter-dependent matrix. Compare, for example, various papers by Michael Overton and co-workers (Overton, 1988; Overton and Womersley, 1993; Haeberly and Overton, 1994). In fact, it might be interesting in general to look at the *range* of aspects, i.e., to compute both the minimum and the maximum over all monotone transformations.

To compute the derivatives of the aspect, we need a general result on the derivatives of eigenvalues. This result is classical. Background, and rigorous proofs, can be found in Kato (1976) or Baumgaertel (1985).

LEMMA 2: *Suppose $Rz = \lambda z$, where $z'z = 1$ and the eigenvalue λ is unique. Then*

$$\frac{\partial \lambda}{\partial R} = zz'$$

This result is powerful enough to implement the algorithm of the previous section, but actually this algorithm can be simplified considerably in this case.

THEOREM 8: *If (θ, μ) corresponds to a stationary value of the maximum eigenvalue aspect, then $y_j = z_j\theta_j$ and $\lambda = \sum_{j=1}^{m} \mu_j$ satisfy the generalized eigenvalue problem*

$$\sum_{\ell=1}^{m} C_{j\ell} y_\ell = \lambda D_j y_j \tag{19a}$$

where

$$\sum_{j=1}^{m} y_j' D_j y_j = 1 \tag{19b}$$

*Conversely, any eigenvalue–eigenvector pair (y, λ) of this generalized eigenvalue problem defines a stationary value of the maximum eigenvalue aspect with $\mu_j = \lambda z_j^2$ and $\theta_j =$ **NORM**(y_j).*

Proof. The stationary equations (3a and b) in this case are given by

$$\sum_{\ell=1}^{m} z_j z_\ell C_{j\ell} \theta_\ell = \mu_j D_j \theta_j \tag{20}$$

which implies that

$$\mu_j = \sum_{\ell=1}^{m} z_j z_\ell r_{j\ell} = \lambda z_j^2 \tag{21}$$

Substituting this and defining $y_j = z_j\theta_j$ gives the generalized eigenvalue problem. A similar substitution establishes the converse. □

This eigenvalue–eigenvector problem is, of course, precisely equation (3b). Thus finding quantifications which maximize the largest eigenvalue is the same thing as finding the dominant eigenvalue in MCA. Again this shows that the first dimension of MCA is special. The remaining dimensions, from this point of view, merely give the remaining stationary values of the maximum eigenvalue aspect.

5. NOTIONS OF MULTIDIMENSIONALITY

The technique we have discussed in the previous section can be extended, or "made multidimensional," in at least two different ways. The fact that there are two forms of multidimensionality in this context has created some confusion, but it also provides a framework in which the Guttman quotation from Subsection 1.3 and the "horseshoe" effect (in French, the "effect Guttman") can be discussed (see Section 7).

5.1 Multiple Quantifications

In our first multidimensional extension, we can compute additional solutions to the generalized eigenvalue problem $Cy = \lambda Dy$. Each one of these defines a stationary value of our maximum eigenvalue aspect, and a corresponding system of feasible transformations. This type of multidimensionality is used in *multiple correspondence analysis*. From the point of view of maximizing aspects it is not very natural to go this way, as Guttman already indicated a long time ago.

Also observe that each additional dimension produces a set of quantifications, which can be used to construct an "induced" correlation matrix. We have $m \times k$ nontrivial solutions with m variables coded with subspaces over dimension k. This produces a lot of correlation matrices (and each of these could be subjected to a principal component analysis, for instance). Gifi uses the expression "data production" in this context.

5.2 Multidimensional Aspects

In the second multidimensional extension, we use as a different aspect the sum of the first p largest eigenvalues. This is used in *nonlinear principal component analysis*, which is discussed in considerable detail in the multidimensional scaling literature. A classic reference is Young et al. (1978), and more recently the technique has been discussed in the ACE framework by Koyak (1987).

The aspect is a convex correlational aspect, but it does not now have a simple relationship with a single fixed generalized eigenvalue problem. Thus the computational problem is inherently more complicated. It is based on the obvious generalization of Lemma 2.

LEMMA 3: *Suppose $RZ = Z\Lambda$, where $Z'Z = 1$ and the p dominant eigenvalues are in the diagonal matrix Λ. If $\lambda_p > \lambda_{p+1}$ then*

$$\frac{\partial \sum_{s=1}^{p} \lambda_p}{\partial R} = ZZ'$$

The majorization algorithm alternates between finding the dominant eigenvalues and their eigenvectors with optimal transformation of the variables. Or, alternatively, we alternate a single simultaneous iteration for the eigenvectors with optimal scaling of the variables. Convergence of these algorithms follows from the general theory.

Observe that these multidimensional aspects give rise to stationary equations that in general also have more than one solution. Such additional solutions, corresponding with other stationary values, have not been studied, except in some very special cases.

6. LINEARIZING ALL BIVARIATE REGRESSIONS

In this section we discuss a very interesting robustness result which was first mentioned in de Leeuw (1982). It says that under some circumstances, the choice of the aspect does not matter. We find the same quantifications, no matter which aspect we maximize.

THEOREM 9: *Suppose we can find* $(\theta_1, ..., \theta_m)$ *that make all bivariate regressions linear. Then* $(\theta_1, ..., \theta_m)$ *satisfies the stationary equations* (3.5), *independently of the aspect.*

Proof. The bivariate regressions are linear if

$$C_{j\ell}\theta_\ell = r_{j\ell}D_j\theta_j \tag{22}$$

If we substitute this in equation (3.5) we find

$$\sum_{\ell=1}^{m} \frac{\partial \phi}{\partial r_{j\ell}} r_{j\ell}D_j\theta_j = \mu_j D_j\theta_j \tag{23}$$

which is an identity because of equation (17). □

Clearly, for such a bilinearizable multivariate distribution, nonlinear principal component analysis will give the same quantifications as the first multiple correspondence analysis dimension.

The question is, however, in how far we can expect to observe approximate bilinearizability in real data. It seems intuitively obvious that we may be able to find it in ordinal variables, such as attitude scales, but we are unlikely to observe it with purely nominal variables which do not have any obvious order on the categories. Also, more than one system of quantifications linearizing the regressions may exist.

7. SOME GAUGES

According to Gifi (1990), a gauge is a dataset whose structure we know. Thus we know what to expect from an analysis of such a gauge, and if a technique does not represent the essential features of the data, it fails the gauge. The notion is used in Gifi's book to broaden the relationship between models and techniques beyond the classical optimality relationship of mathematical statistics.

We shall discuss some gauges to show that bilinearizability occurs in at least three common situations.

7.1 A Single Bivariate Table

In a single bivariate table both regressions can be linearized by performing a correspondence analysis on the table. The row and column scores of the correspondence analysis linearize the regressions. In fact this is one way to define correspondence analysis: find row and column scores for a table which makes both regressions linear.

7.2 Binary Variables

If all variables are binary, then any set of scores linearizes the regressions, because two points are always on a line.

7.3 The Multivariate Normal

Suppose the multivariate distribution we analyze is a multivariate normal, with standard normal marginals and correlations $\rho_{j\ell}$. Of course if we *know* that the distribution is multivariate normal, we generally do not apply any "optimal scaling" technique. In that case classical multinormal multivariate analysis applies.

For the multinormal, the Hermite polynomials of degree s in $h_s(x_j)$ are a bilinearizing system. In fact

$$C_{j\ell}h_s(x_\ell) = \rho_{j\ell}^s D_j h_s(x_j) \tag{24}$$

Thus we have a denumerable system of bilinearizing systems, one for each polynomial degree. Each system induces a correlation matrix which is an element-wise (Hadamard) power of the correlation matrix R. The eigenvalues of MCA are the m eigenvalues of R, the m eigenvalues of $R^{(2)}$, and so on. Transformations corresponding to $R^{(s)}$ are all Hermite polynomials of degree s, weighted by the coefficient from the eigenvector.

This polynomial bilinearizibility, discussed for the bivariate case in great detail by Lancaster and his students, (Lancaster, 1969), is one possible explanation of the (in)famous "horseshoe" of correspondence analysis. If the largest eigenvalue of $R^{(2)}$ is larger than the second eigenvalue of $R = R^{(1)}$, which will happen for dominant first dimensions, then all second-dimension transformations are quadratic functions of the first-dimension transformations. Two dimensional transformation plots will look like horseshoes.

Observe that basically the same result applies to what Yule calls *strained multinormals*. In this case the variables we observe are smooth monotonic ("strained") transformations of a number of standard joint multinormals. Applying MCA (or any other aspect-based transformation technique) will "unstrain" the distribution by finding the inverse transformation, which linearizes all bivariate regressions. Some statistical consequences of bilinearizability are discussed by de Leeuw (1988).

7.4 KPL Diagonalization

The structure in our three gauges can be described nicely in terms of diagonalizing the Burt matrix. Let us look at the Burt matrix of a multinormal, for instance. We continue to use matrix notation, even though some of our operators are infinite-dimensional. Thus each $C_{j\ell}$ is a standard bivariate normal, with correlation $\rho_{j\ell}$. Collect the Hermite polynomials as columns in the "matrix" K. Then $K'C_{j\ell}K = \Lambda_{j\ell}$, where $\Lambda_{j\ell}$ is a diagonal matrix which contains the powers of $\rho_{j\ell}$. Thus, if we use the direct sum $\mathcal{K} = K \oplus \cdots \oplus K$ in $\mathcal{K}'C\mathcal{K}$, then all blocks will be diagonal. Thus there is a permutation matrix P, such that $P'\mathcal{K}'C\mathcal{K}P$ is the direct sum $R \oplus R^{(2)} \oplus \cdots$. Construct the direct sum $\mathcal{L} = L \oplus L_{(2)} \oplus \cdots$ which contains the eigenvalues of the $R^{(s)}$. Then $\mathcal{L}'P'\mathcal{K}'C\mathcal{K}P\mathcal{L}$ is diagonal, i.e., the matrix $\mathcal{K}P\mathcal{L}$ has the eigenvalues of the Burt matrix C.

This KPL structure for the eigenvectors also occurs (trivially) in our two other gauges (see de Leeuw (1982) and Bekker and de Leeuw (1988) for details). More importantly, however, there is an approximate KPL structure for the MCA eigenvalues in many actual examples with attitude or diagnostic scales. It provides a much more compact decomposition of the Burt matrix, and it shows in which respects an MCA of the matrix is redundant.

Of course assuming KPL is stricter than assuming bilinearizibility, because if KPL applies, each dimension is a bilinearizing system (with some of the bilinearizing systems inducing the same correlation matrix).

8. DISCUSSION

If we compare the geometrical approach to multivariate analysis with optimal scaling and the aspect-based approach, we see that using aspects gives us a great deal of additional generality, and that we stay relatively close to classical multivariate analysis. The same criteria are studied, and the notion of an aspect suggests many generalizations (and majorization can be used to produce simple algorithms).

Choice of a particular aspect is a major practical problem, similar to the problem of choosing a loss function for regression. We have seen that bilinearizability and KPL diagonalizability provide a partial solution to the dilemma, at least as far as the optimal transformations are concerned.

Bilinearizability is a very powerful property. It guarantees that different aspects will lead to the same transformations, and it guarantees a certain statistical robustness which entails that classical asymptotic tests and estimates can still be used (see de Leeuw, 1988). Although perfect bilinearizability only occurs in some of the standard gauges, it seems to occur approximately in many practical situations.

REFERENCES

Baumgaertel, H. (1985). *Analytic Perturbation Theory for Matrices and Operators*. Basel, Boston, Stuttgart: Birkhauser.

Bekker, P. and de Leeuw, J. (1988). Relation between variants of nonlinear principal component analysis. In: J. van Rijckevorsel and J. de Leeuw, eds. *Components and Correspondence Analysis*. Chichester: Wiley.

Benzécri, J. (1992). *Correspondence Analysis Handbook*. New York: Marcel Dekker.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. R. Stat. Soc.*, **B26**, 211–252.

Burt, C. (1950). The factorial analysis of qualitative data. *Br. J. Stat. Psychol.*, **3**, 166–185.

de Leeuw, J. (1973). Canonical analysis of categorical data. PhD Thesis, University of Leiden. Republished in 1985 by DSWO-Press, Leiden.

de Leeuw, J. (1982). Nonlinear principal component analysis. In: H. Caussinus et al., ed. *COMPSTAT 1982*, Vienna: Physika Verlag.

de Leeuw, J. (1983). On the prehistory of correspondence analysis. *Stat. Neerlandica*, **37**, 161–164.

de Leeuw, J. (1988). Multivariate analysis with linearizable regressions. *Psychometrika*, **53**, 437–454.

de Leeuw, J. (1990). Multivariate analysis with optimal scaling. In: S. D. Gupta and J. Sethuraman, eds. *Progress in Multivariate Analysis*. Calcutta: Indian Statistical Institute.

de Leeuw, J. (1993). Some generalizations of correspondence analysis. In: C. M. Cuadras and C. R. Rao, eds. *Multivariate Analysis: Future Directions 2*. Amsterdam, London, New York, Tokyo: North-Holland.

di Battista, G., Eades, P., Tamassia, R., and Tollis, I. (1994). Algorithms for automatic graph drawing: An annotated bibliography. *Comput. Geom.*, **4**, 235–282.

Eades, P. (1984). A heuristic for graph drawing. *Congressus Numerantium*, **42**, 149–160.

Eades, P. and Wormald, N. C. (1994). Edge crossings in drawings of bipartite graphs. *Algorithmica*, **11**, 379–403.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Chichester: Wiley.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. New York: Academic Press.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In: P. Horst, ed. *The Prediction of Personal Adjustment*. New York: Social Science Research Council.

Haeberly, J.-P. A. and Overton, M. (1994). A hybrid algorithm for optimizing eigenvalues of symmetric definite pencils. *SIAM J. Matrix Anal. Appl.*, **15**, 1141–1156.

Hoffman, D. L. and de Leeuw, J. (1992). Interpreting multiple correspondence analysis as a multidimensional scaling method. *Marketing Lett.*, **3**, 259–272.

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

Horst, P. (1965). *Factor Analysis of Data Matrices*. New York: Holt, Rinehart & Winston.

Kato, T. (1976). *Perturbation Theory for Linear Operators*, 2nd edn. Berlin, Heidenlberg, New York: Springer.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, **58**, 433–451.

Koyak, R. (1987). On measuring internal dependence in a set of random variables. *Ann. Stat.*, **15**, 1215–1228.

Lancaster, H. (1969). *The Chi-Squared Distribution*. New York: Wiley.

Legendre, A. M. (1805). *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*. Paris: Courcier.

Meijerink, F. (1996). *A Nonlinear Structural Relations Model*. Leiden: DSWO Press.

Michailidis, G. and de Leeuw, J. (1998). The Gifi system for descriptive multivariate analysis. Statistical Science, 13.

Overton, M. (1988). On minimizing the maximum eigenvalue of a symmetric matrix. *SIAM J. Matrix Anal. Appl.*, **9**, 256–268.

Overton, M. and Womersley, R. (1993). Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Math. Prog.*, **62**, 321–357.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton: Princeton University Press.

Steel, R. G. D. (1951). Minimum generalized variance for a set of linear functions. *Ann. Math. Stat.*, **22**, 456–460.

Stigler, S. M. (1986). *The History of Statistics*. Cambridge: Belknap Press.

ten Berge, J. M. F. (1988). Generalized approaches to the MAXBET problem and the MAXDIFF problem, with applications to canonical correlations. *Psychometrika*, **53**, 487–494.

Tijssen, R. and de Leeuw, J. (1989). Multi-set nonlinear canonical analysis via the burt-matrix. In: R. Coppi and S. Bolasko, eds. *Multiway Data Analysis*. Amsterdam, New York: North Holland.

van de Geer, J. P. (1984). Linear relations among K sets of variables. *Psychometrika*, **49**, 79–94.

van der Burg, E., de Leeuw, J., and Verdegaal, R. (1988). Homogeneity analysis with K sets of variables: An alternating least squares approach with optimal scaling features. *Psychometrika*, **53**, 177–197.

Young, F., Takane, Y., and de Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, **45**, 279–281.

Zangwill, W. I. (1969). *Nonlinear Programming. A Unified Approach.* Englewood Cliffs: Prentice-Hall.