Nonlinear Multivariate Analysis with Optimal Scaling

Jan de Leeuw
Department of Data Theory FSW
University of Leiden

Paper presented at the NATO Advanced Research Workshop on Numerical Ecology. Station Marine de Roscoff, Brittany, France. June 3-11, 1986.

Introduction

It has already been pointed out by many authors that multivariate analysis is the natural tool to analyze ecological data structures. Gauch summarizes the reasons for this choice in a clear and concise way. "Community ecology concerns assemblages of plants and animals living together and the environmental and historical factors with which they interact. ... Community data are multivariate because each sample site is described by the abundances of a number of species, because numerous environmental factors affect communities, and so on. ... The application of multivariate analysis to community ecology is natural, routine, and fruitful." (Gauch, 1982, p. 1). Legendre and Legendre discuss the ecological hyperspace implicit in Hutchinson's concept of a *fundamental niche*. "*Ecological data sets* are for the most part *multidimensional:* the ecologist samples along a number of axes which, depending on the case, are more or less independent, with the purpose of finding a structure and interpreting it." (Legendre and Legendre, 1983, p. 3).

A number of possible ecological applications of multivariate techniques are mentioned in the following quotation from the recent book by Gittins (1985). "Ecology deals with relationships between plants and animals and between them and the places where they live. Consequently, many questions of interest to ecologists call for the investigation of relationships between variables of two distinct but associated kinds. Such relationships may involve those, for example, between the plant and animal constituents of a biotic community. They might also involve, as in plant ecology, connections between plant communities and their component species, on the one hand, and characteristics of their physical environment on the other. As another example, comparative relationships among a number of affiliated species or populations with respect to a particular treatment regime in a designed experiment might be studied. In more general terms, the question which arises calls for the exploration of relationships between *any* two or more sets of variables of ecological interest." (l.c., page 1).

It is of some importance to observe that Gittins gives a somewhat limited

description of the possibilities of multivariate analysis here. The reason being, of course, that his book is about canonical analysis, a rather specific class of multivariate techniques. We can study relationships *between* sets of variables, as in the various form of canonical analysis, but also relationships *within* a single set of variables, as in the various forms of clustering and component analysis. In classification and ordination, for example, we usually deal with a single set of variables. Each species in the study defines a variable, assigning abundance numbers to a collection of sites. It may seem natural to relate sets of variables if we want to study abundance or behaviour of species in relation to the environment, but it would be more appropriate to analyze the within-structure of a single set if we describe the structure of a single community or location. And if we want to study the interaction between members of a community, under various circumstances, it may be even more appropriate to use techniques derived from multidimensional scaling, for which the basic data are square interaction or association matrices and the basic units are pairs of individuals.

Forms and problems of multivariate analysis

As indicated in the introduction, multivariate analysis studies the relationships between a number of variables which are defined for each of a number of objects. We shall formalize this below, but the intuitive meaning is probably clear. The objects can be samples or sites, and the variables can be species with varying degree of abundance in each of the sites or they can be physical characteristics of the sites. Or the objects can be pairs of individuals of a certain species, and the variables can be measures of interaction between the pairs. In this section we argue that multivariate analysis consists of a very large variety of models and techniques, in fact a far greater variety than one could ever hope to discuss in a single paper or book. Nevertheless some classes of techniques can be distinguished, and we shall briefly discuss them in order to delineate the class we shall be talking about in the sequel. A more extensive treatment of the same classificatory problem

is in Gifi (1981), and in Gnanadesikan and Kettenring (1984).

In mathematical statistics the notion of a *model* plays a very prominent part. In fact the model is usually the starting point of a statistical analysis. The assumption is that the data are realizations of random variables, whose distribution, except possibly for some unknown constants, is described by the model. In multivariate analysis by far the most prominent model is the multivariate normal distribution (Anderson, 1958, Muirhead, 1982). The multivariate measurements are assumed to be realizations of independent random vectors, each with the same multivariate normal distribution. Statistical techniques estimate unknown parameters and test hypotheses on the basis of this multinormal model, usually employing the likelihood function. The multivariate normal distribution has numerous technical and interpretational advantages, which are mostly due to its intimate connections with Euclidean geometry.

In recent years another model has gained some prominence, mainly in discrete multivariate analysis. This is the multinomial model, usually presented in the form of loglinear analysis (Bishop, Fienberg, and Holland, 1975, Haberman, 1979, also compare Legendre and Legendre, 1983, chapter 4). Again the basic assumption is that we are dealing with realizations of independent and identically distributed random vectors, but in multinomial analysis no additional parametric assumptions are made. Because the data vectors are discrete, and each variable assumes only a finite number of values, it is possible to use such a nonparametric approach. The main difference between the multinormal and the multinomial model is that in the multinormal case we only have to model the first order interactions between the variables. Because the means and covariances are a complete set of sufficient statistics, they contain all information in the data, and we can ignore all higher order moments. In the multinomial model all higher order interactions have to be taken into account. This often leads to serious interpretational problems, and it makes analysis with a moderate number of variables already quite impractical. It is consequently not surprising that much effort in the recent statistical literature is expended on the development of models which combine features of multinomial and multinormal modelling (Agresti, 1983). In a sense the techniques we shall present below can also

be interpreted as such combinations.

In another sense, however, there are important differences between the classical statistical modelling techniques and our multivariate data analysis methods. As we have seen above, the notion of a probabilistic model is basic in classical statistics. From the model we derive the technique, and the results then tell us if the model is appropriate or not. In multivariate data analysis we work differently. We do not make explicit assumptions about the process that has generated the data, because very often it is not at all clear how realistic such assumptions are, and in many cases it is even clear that the usual assumptions are not satisfied at all. Multivariate normality and complete independence are quite rare in practice. Thus in stead of starting with a model and trying to fit in the data, we start with the data and we try to find a structure or model that can describe or summarize the data. These two approaches correspond, of course, with the age-old distinction between *induction* and *deduction*, between *empiricism* and *rationalism*. In recent discussions the concepts of *exploration* and *confirmation*, and of *description* and *inference*, are often contrasted. Data analysts generally feel that the models of classical statistics are much too strong and too unrealistic to give good descriptions of the data. And, of course, mathematical statisticians feel that the techniques of data analysis very often lead to unstable results, that are difficult to integrate with existing prior knowledge. It will not come as a surprise, that we think that both approaches have their value. If there is strong and reliable prior knowledge, then it must be incorporated in the data analysis, because it will make the results more stable and more easy to interprete. But if this prior knowledge is lacking, it must not be invented just for the purpose of being able to use standard statistical methodology. And, certainly, we must not make assumptions which we know to be not even approximately true. Finally there are many situations in which good statistical procedures can in principle be applied, on the basis of firm prior knowledge, but in which there simply are not enough data to make practical application possible. In such situations a data analytical compromise is needed too.

There are some interesting problems in the application of various multivariate analysis techniques to ecology. They have been admirably reviewed by

Noy-Meir and Whittaker (1978). We mention them briefly here, but we shall also encounter them again in our more formal development below. The distinction between *R and Q techniques* has been discussed extensively by psychometricians such as Cattell and Stephenson. It is based on the fact that we think as the species as ordening the samples, but also as the samples as ordening the species. In a given data structure we have to decide what the variables are, and what the units are on which the variables are defined. Sometimes the choice is clear and unambiguous, sometimes the situation is more complicated. Compare Heiser (1986) for additional discussion of this problem. As a second problem Noy-Meir and Whittaker mention *data transformation* and the *choice of similarity measures*. We could generalize this somewhat to the problem of *data definition and expression.* This has as special cases the choice of centering and standarization, but also taking logarithms or using any of the other reexpression techniques discussed by Legendre and Legendre (1983, p. 11-18). The nonlinear multivariate techniques explained in our paper take a radical point of view, by assuming that the expression of the variable in the data matrix is essentially conventional, merely a *coding*. Thus the reexpression problem does not have to be solved before the technique is applied, but it is an important part of our multivariate techniques to find appropriate reexpressions. The third problem is the distinction between the *discrete* and the *continuous*, or between *ordination* and *classification*. This has also been discussed extensively in the psychometric multidimensional scaling literature. Compare Carroll and Arabie (1980), De Leeuw and Heiser (1982). In this paper we take the point of view that continuous representation, if applied carefully, will often show discontinuities in the data. Assuming discontinuity right away, and applying a classification or cluster method, in many cases imposes too much a priori structure. A final problem mentioned by Noy-Meir and Whittaker is that of *non-linearity* and *axes interpretation*. This is perhaps especially relevant in connection with the component analysis or correspondence analysis of abundance matrices, in which we invariably find the *horseshoe* or *Guttman effect* (Heiser, 1986). Again the nonlinear multivariate analysis techniques discussed below take a radical stand in this problem. Nonlinearities due to the coding of the variables are avoided by finding optimal

transformations, and nonlinearities that occur in the representation can be eliminated by imposing restrictions on the representation, somewhat as in *detrended correspondence analysis* (Hill and Gauch, 1980).

Noy-Meir and Whittaker come to the following conclusion in their useful review paper. "After twenty-five years of development of continuous multivariate techniques in ecology, some of the early optimistic promises, as well as some of the sceptical criticisms, seem to have been overstated " (Noy-Meir and Whittaker, 1978, p. 329). The nonlinear multivariate data analysis techniques developed in this paper may contribute additional useful procedures and possibilities. But they must be seen in the proper perspective. If there is strong prior knowledge, either of a structural or of a probabilistic nature, then it must be incorporated in the analysis. Sometimes our techniques have options which make it possible to build in suitable restrictions, but if the information is very specific, then one must switch to a specific technique. If it is known that species distributions are Gaussian, then one should use Gaussian ordination, and not correspondence analysis. Our techniques are most useful in the areas in which there is not much prior knowledge, or in which the ratio of amount of data to amount of theory is large.

Multivariables

We start our formal developments in this paper by providing some definitions. In multivariate analysis we always study a number of *variables*, defined on a set of *objects*. More precisely, a *variable* is a function . Legendre and Legendre use a slightly different terminology. "Any ecological study, classical as well as numerical, is based on *descriptors*. In this text the term *descriptor* will be used for the attributes, variables, or characters (also called items in the social sciences) that describe or compare the *objects of the study*. The *objects* that the ecologists compare are the samples, locations, quadrats, observations, sampling units or subjects which are defined *a priori* by the sampling desing, before making the observations." (Legendre and Legendre, 1983, p. 8). For variables we use the

familiar notation $\phi : \Omega \rightarrow \Gamma$. Here Ω is the *domain* of the variable, consisting of the objects, and Γ is its *target*, containing the possible *values* of the variable. Elements of the target are also called the *categories* of a variable. A variable ϕ associates with each $\omega \in \Omega$ a category $\phi(\omega) \in \Gamma$. In practical applications and in actual data analysis the domain Ω will be a finite set $\{\omega_1,...,\omega_n\}$. For theoretical purposes the domain can be infinite. If Ω is a probability space, for instance, and ϕ is measurable, then our variable is a *random variable*. Targets can be finite or infinite as well. In many cases the target is the reals or the integers, i.e. $\Gamma = \mathbb{R} =]-\infty,+\infty[$, or $\Gamma = \mathbb{N} = \{0,1,2,....\}$. But it is also possible that Γ = {short grass, short grass /thicket, tall grass with thicket} or Γ = {close, moderate, distant}.

Table 1.5 from Legendre and Legendre (1983, p. 9), that we copy here, shows the types of targets we can expect to encounter. Most of the terminology will probably be clear, but we refer to Legendre and Legendre (1983, p. 10-11) for further explanation.

Descriptor type	Examples
Binary (two states, presence-absence)	species present or absent
Multi-state (many states)	
| nonordered (qualitative, nominal, attributes)	geological group
| ordered	
| | semi-quantitative (rank-ordered, ordinal)	importance or abundance scores
| | quantitative (measurement)	
| | | discontinuous (meristic, discrete)	equidistant abundance classes
| | | continuous	temperature, length

Most of the techniques of multivariate analysis have been developed for continuous variables such as temperature and length. As shown by Gittins (1985), for example, nonnumerical multi-state variables can be incorporated in some techniques. In analysis of variance, for example, the design matrices consist of *dummies*, which are codings of nonordered multi-state variables. In discriminant

analysis a similar dummy is used to code class membership. It remains true, however, that the models of classical continuous multivariate analysis are entirely in terms of multinormal variables. Dummies are used only as coding devices, to indicate that objects are sampled from different populations. In nonlinear multivariate analysis as discussed in this paper we use dummies and coding in a much more constructive way. A good starting point is the following quotation. "Coding is a technique by which raw data can be transformed into other values that can then be used in the analysis. All types of descriptors can be recoded but non-numerical descriptors must be coded before they can be analyzed numerically." (Legendre and Legendre, 1983, p. 10). The coding problem is thus related to the reexpression problem discussed above. If variables are numerical we often use *transformation*, if they are non-numerical we use *quantification* , but in all cases the coding we use is a real-valued function on the target set of the variable. Real-valued codings of non-numerical variables are often called *scalings*. Coding in many cases is dictated by conventional considerations. Thus {close,moderate,distant} is often coded as {1,2,3}, but in nonlinear multivariate analysis we look specifically for codings (or transformations, or quantifications, or scalings) which are *optimal* in a well-defined sense.

In multivariate analysis we analyze several variables at the same time. This requires some additional terminology. A *multivariable* is a set of variables with a common domain. We use the notation $\Phi = \{\phi_t \mid t \in T\}$, where $\phi_t : \Omega \to \Gamma_t$, and where T is the *index set* of the multivariable. Thus the variables in Φ have the common domain Ω, but they have possibly different targets Γ_t. *Multivariate analysis* studies the structure of multivariables.

A simple example is perhaps useful here. We have taken some classical data of Mayr (1932). The domain of the five variables in this example consists of twelve races of the bird *Pachycephala Pectoralis*. Specifically Ω = {dahli, chlorura, vitiensis, bougainvillei, torquata, melanota, melanoptera, sanfordi, ornata, bella, optata, graeffii}. Variable 1 is called THROAT and maps Ω into Γ_1 =

{yellow,white}. Variable 2 is called BREAST BAND and maps Ω into Γ_2 = {present,absent}. COLOR OF BACK maps Ω into Γ_3 = {olive,black}, FOREHEAD maps into Γ_4 = {yellow,black}. Variable 5, finally, is called WING, and maps the races into Γ_5 = {colored,black}. All variables are binary. The multivariable is defined by the following table, in which we have numbered races and variables, and in which we have used simple abbreviations for the values of the variables.

	1	2	3	4	5
01	W	P	O	B	C
02	W	P	O	B	C
03	W	P	O	B	C
04	Y	P	O	B	C
05	Y	P	O	B	C
06	Y	P	B	B	B
07	Y	P	O	B	B
08	Y	A	O	B	C
09	W	P	B	B	C
10	Y	P	O	Y	C
11	Y	P	O	Y	C
12	Y	A	O	Y	C

Table 1:
bird data from Mayr

Thus, for instance, ϕ_3(bella) = olive and ϕ_2(sanfordi) = absent.

The example can also be used to illustrate *interactive coding* of variables. COLOR OF BACK and FOREHEAD are binary variables with targets, respectively, {olive,black} and {yellow,black}. Using them we can create the interactive variable COLOR OF BACK x FOREHEAD with target {(olive,yellow), (olive,black), (black,yellow), (black,black)}. In general if we have m variables with $k_1,...,k_m$ categories, i.e. a total of $k_1 + ... + k_m$ categories, then we can create an interactive variable with $k_1 \times ... \times k_m$ categories. We can also make interactive codings for all

pairs of variables, this gives us a total of $C(k_1,2) + ... + C(k_m,2)$ categories. Here C(k,r) is used for binomial coefficients. Thus there are many possibilities of coding a given set of variables.

The example above is quite straightforward, but it is not representative for a typical ecological data set. More representative examples are given, for example, in appendix A2 of Gittins (1985). The limestone grassland community example, discussed by Gittins in his chapter 7, defines eight estimates of species abundance and six soil variables on a random sample of 45 stands, each of 10 x 10 meter . Each stand was divided into 5000 units of 10 x 20 cm, and species abundance is defined as the percentage of these units in which the species occurred. It is clear that the most natural object in this experiment is the 10 x 20 cm unit, i.e. there are 45 x 5000 = 225000 such units. The eight species define binary variables on these units, with target {present,absent}. There is a variable called STAND, which takes 45 different values, and there are six soil variables, which have the property that units within the same stand get the same soil value on all six of them. We can also follow Gittins and use the stand as the fundamental unit. This process is called *aggregation*, because it involves aggregating the 5000 original units in a single stand. This aggregation process makes it possible to treat the abundancies as numerical variables, taking values between 0% and 100%. The example shows that the choice of unit is sometimes debatable.

The next example is also representative, but a bit more problematical. It is taken from Legendre and Legendre (1983, p. 191). Five ponds are characterized by the abundances of different species of zooplankton, given on a scale of relative abundance varying from 0 to 5. It is clear that this matrix is also based on aggregation, of the same sort as in the Gittins example. But we can also use it to illustrate *transposition*, or the choice between Q and R. In this example we can take the species as units, and the ponds as variables. Each pond maps the eights species into the target {0,1,2,3,4,5}. It is also possible to interpret the ponds as units and the species as variables, again with the same target {0,1,2,3,4,5}. We can also treat the example as bivariate. The grand-total of the data matrix is 52. These 52

'abundance credits' are used as the units, and the two variables are SPECIES and PONDS. Thus there are three credits with species-value 1 and pond-value 212, and four credits with species-value 5 and pond-value 214, and so on. The data matrix is, in this interpretation, the cross table of the two variables. And finally we can use the 40 ponds and species combinations as units, and interpret our results as measurements on a variable that maps these 40 combinations into {0,1,2,3,4,5}. Two other variables can be defined on these units. The first one is POND, with five variables in its target, and the second one is SPECIES, with eight values. In this last interpretation there are consequently 40 units, and three variables. There are no clear a priori reasons for preferring one interpretation over the other. The choice must be made by the investigator, in combination with the choice of the data analysis technique.

Species	Ponds				
	212	214	233	431	432
1	3	3	0	0	0
2	0	0	2	2	0
3	0	2	3	0	2
4	0	0	4	3	3
5	4	4	0	0	0
6	0	2	0	3	3
7	0	0	0	1	2
8	3	3	0	0	0

Table 2: zooplankton data of Legendre

Functions of Correlation Matrices

In this paper we shall mainly discuss multivariate techniques which compute statistics depending on the second order moments and product moments of the

variables, more specifically on their *correlation coefficients*. This implies, obviously, that the higher order moments of the distributions of the variables are irrelevant for the techniques we discuss. Thus the loglinear methods for frequency tables, for example, are not covered by the developments in this paper. On the other hand our techniques also do not depend on first order moments, i.e. on the means of the variables. This means that we can suppose, without loss of generality, that all variables we deal with are in deviations from the mean. We are not interested in the structure of the means, although our development of discriminant analysis and analysis of variance will show that in some cases means can be reintroduced by the use of dummy variables. Because our methods depend only on the correlation coefficients, this moreover means that they are *scale-free*. The unit of the variables and consequently their variances are irrelevant. All variables can be assumed to be standardized to unit variance. It is one of the purposes of this paper to show that this somewhat limited class of multivariate techniques still has many interesting special cases.

Now this description of the class of techniques we are interested in is somewhat problematical. We can compute correlations only between variables which are numerical, so either we must limit our attention to measured variables, or we must compute correlations between non-numerical variables which are coded numerically. And if we use coding of non-numerical variables, and then compute correlations, then it is clear that the correlations will depend on the particular coding or scaling that we have chosen. And, in fact, something similar is also true for measured variables. In stead of using abundance or yield, for instance, we could also use log-abundance or log-yield, which would give different correlations. We introduce some notation to describe this scaling or transformation of the variables.

Remember that we started with a multivariable $\Phi = \{\phi_t \mid t \in T\}$, where $\phi_t : \Omega \to \Gamma_t$. A *scaling* (or *quantification*, or *transformation*) of the targets of this multivariable is a system $\Psi = \{\psi_t \mid t \in T\}$, where $\psi_t : \Gamma_t \to \mathbb{R}$. The values of a scaling are often called the *category quantifications* of a variable (or the *transformed values*). A scaling of the targets *induces* a quantification Λ of the multivariable by

the simple rule $\Lambda = \{\lambda_t \mid t \in T\}$, where λ_t is the composite $\psi_t \circ \phi_t : \Omega \to \mathbb{R}$. This is illustrated in Figure 1. Write $R(\Lambda)$ for the correlation matrix induced by the scaling of the variables.

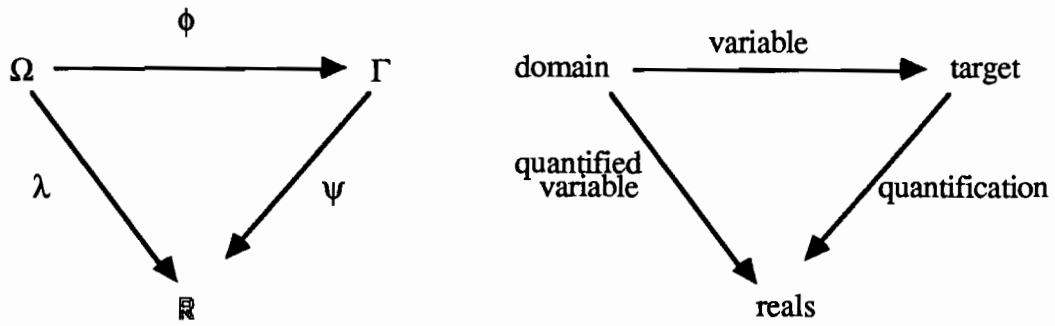

Figure 1:
Quantification diagram

Time to switch to an example. In the first three columns of Table 3 the zooplankton data of Legendre and Legendre are coded as 40 observations on the three variables SPECIES, POND, and ABUNDANCE. We use *integer coding*, or *category numbers*. Observe that SPECIES and POND are uncorrelated, because the design is *balanced*. Only the correlations of SPECIES and POND with ABUNDANCE depend on the scaling of the variables we have chosen. With integer coding the correlation between SPECIES and ABUNDANCE is -.01, and the correlation between POND and ABUNDANCE is -.06. Now suppose that we use a form of scaling which is sometimes called *criterion scaling*. This means that we use integer coding for ABUNDANCE, but both for SPECIES and for POND we choose the average ABUNDANCE values of a species or pond as the quantifications. The SPECIES - ABUNDANCE correlation increases to .29, and the POND - ABUNDANCE correlation to .16. The proportion of variance of ABUNDANCE 'explained' by SPECIES and POND is .1082.

We shall discuss other criteria and other solutions below, but first we have

VARS			INDICATOR CODINGS		
S	P	A	SPECIES	POND	ABUNDANCE
1	1	3	10000000	10000	000100
1	2	3	10000000	01000	000100
1	3	0	10000000	00100	100000
1	4	0	10000000	00010	100000
1	5	0	10000000	00001	100000
2	1	0	01000000	10000	100000
2	2	0	01000000	01000	100000
2	3	2	01000000	00100	001000
2	4	2	01000000	00010	001000
2	5	0	01000000	00001	100000
3	1	0	00100000	10000	100000
3	2	2	00100000	01000	001000
3	3	3	00100000	00100	000100
3	4	0	00100000	00010	100000
3	5	2	00100000	00001	001000
4	1	0	00010000	10000	100000
4	2	0	00010000	01000	100000
4	3	4	00010000	00100	000010
4	4	3	00010000	00010	000100
4	5	3	00010000	00001	000100
5	1	4	00001000	10000	000010
5	2	4	00001000	01000	000010
5	3	0	00001000	00100	100000
5	4	0	00001000	00010	100000
5	5	0	00001000	00001	100000
6	1	0	00000100	10000	100000
6	2	2	00000100	01000	001000
6	3	0	00000100	00100	100000
6	4	3	00000100	00010	000100
6	5	3	00000100	00001	000100
7	1	0	00000010	10000	100000
7	2	0	00000010	01000	100000
7	3	0	00000010	00100	100000
7	4	1	00000010	00010	010000
7	5	2	00000010	00001	001000
8	1	3	00000001	10000	000100
8	2	3	00000001	01000	000100
8	3	0	00000001	00100	100000
8	4	0	00000001	00010	100000
8	5	0	00000001	00001	100000

Table 3: Category numbers and indicators for Legendre zooplankton data

to develop some notation and terminology which make it possible to discuss the optimal scaling problem in general. The general approach and the notational system are due, in some specific cases, to Fisher (1941) and to Guttman (1941). A more comprehensive approach to nonlinear multivariate analysis along these lines originated with Guttman (1959) and De Leeuw (1973). The specific notational system and terminology we use in this paper are due to Gifi (1981), also compare De Leeuw (1984a).

Indicator matrices and quantification

Let us look at the second part of Table 3. This contains the same information as the first three columns, but coded differently. In the terminology of De Leeuw (1973) we call the codings of the variables *indicator matrices*, but in other contexts they are also called *dummies*. One interpretation is that SPECIES, for instance, is now coded as a set of eight different binary variables. The total number of variables, in this interpretation, is now equal to 19, which is the total number of categories of SPECIES, POND, and ABUNDANCE. The important property of indicator matrices, for our purposes, is that each possible quantification of the variables is a linear combination of the columns of the indicator matrix of that variable. Or, if there are n objects, we can say that the columns of the indicator matrix form a *basis* for the *subspace* of $\mathbb{R}^n$ defined by the quantifications of the variable. The columns *span* the space of possible quantifications.

Suppose G_t is the indicator matrix of variable t. Assume that there are n objects and that variable t has k_t categories. Then G_t has n rows and k_t columns. The matrix $D_t =_{df} G_t'G_t$ is diagonal, i.e. the columns of G_t are orthogonal (the categories of a variable are exclusive). And the rows of G_t sum to unity (the categories are exhaustive). A quantification ψ_t of the categories maps the k_t-element

set Γ_t into the reals, and is thus a k_t-element vector. Write it as y_t. Then λ_t, the quantified variable, is given by the product $q_t = G_t y_t$. Given vectors y_t of category quantifications we can construct quantified variables, and given quantified variables we can construct the correlation matrix $R(\Lambda)$. We limit our attention to normalized quantifications. If u is used for a vector with all elements equal to +1, the number of elements of u depending on the context, then we want $u'q_t = u'G_t y_t = u'D_t y_t = 0$ and $q_t'q_t = y_t'D_t y_t = n$. If s and t are two variables, with corresponding indicators and normalized quantifications, then the correlation between the quantified variables is given by $r_{st} = n^{-1}\, y_s'C_{st}y_t$, where $C_{st} =_{df} G_s'G_t$ is the *cross-table* of variables s and t. Observe that $D_t = C_{tt}$. Our formulation of the quantification problem in terms of vectors and matrices shows that the correlations r_{st} are functions of the bivariate frequencies, collected in the cross-tables C_{st}, and the category quantifications y_t. For a given problem, i.e. a given coding of a fixed data set, the C_{st} are constant and known, but varying the y_t will give varying correlation coefficients. The comparison of integer scaling and criterion scaling in the previous section was a first example of this.

Some common criteria for optimal scaling

We now make a next step. The correlations vary with the choice of the quantifications, and consequently all statistics depending on the correlations will also vary. Suppose $\kappa(R(\Lambda))$ is such a (real-valued) statistic, interpreted as a function of the scalings. We are interested in the variation of this statistic, and in many cases in the largest and/or smallest possible value, under choice of quantifications. It is possible, for instance, to look for the quantifications of the variables which maximize or minimize a specific correlation. Or, if we have a number of predictors

and a single variable which must be predicted, we can choose scalings for optimal prediction, i.e. with maximum multiple correlation coefficient. If the purpose of the multivariate technique is ordination or some other form of dimension reduction, then we can choose quantifications in such a way that a maximum amount of dimension reduction is possible. In a principal components context this could mean that we maximize the largest eigenvalue, or the sum of the p largest eigenvalues, of the correlation matrix $R(\Lambda)$. In fact we can look through the books on linear multivariate analysis and find many other criteria that are used to evaluate results of multivariate techniques. There are canonical correlations, likelihood ratio criteria in terms of determinants, largest root criteria, variance ratio's, and so on. For each of these criteria we can study their variation under choice of quantifications, and we can look for the quantifications that make them as large (or as small) as possible.

Before we give some examples, we briefly discuss the mathematical structure of such optimal scaling problems. If we restrict ourselves to the case of n units of observation, coded with indicator matrices, then the stationary equations for an extreme value of criterion κ over normalized quantifications are

$$\Sigma_{t\neq s}\,\pi_{st}C_{st}y_t = \mu_s D_s y_s,$$

where $\pi_{st} = \partial\kappa/\partial r_{st}$. These stationary equations suggest the algorithm

For s=1 to m:

A1: compute $q_s = \Sigma_{t\neq s}\,\pi_{st}G_t y_t$,

A2: compute $y_s = D_s^{-1}G_s' q_s$,

A3: compute y_s by normalizing y_s,

next s.

Observe that the algorithm can be used for any criterion κ. The criterion influences

the algorithm only through the form of the partial derivatives π_{st}. It is not guaranteed that it works, i.e. converges, for all criteria. A detailed mathematical analysis is given by De Leeuw (1986), who shows that the algorithm does indeed work for some of the more usual criteria used in nonlinear multivariate analysis, such as the ones we have mentioned above.

Let us now look at an example. If we want to apply optimal scaling to the example of Mayr, in Table 1, then we get into trouble. Because all variables are binary, the possible scalings are completely determined by the normalization conditions. For binary variables, there is only one possible scaling, and in that sense they are are the same as numerical variables. We could create variables with more than two categories by using interactive coding, but the example is so small and delicate that this would probably not be worthwhile.

We thus apply the algorithm, with various different criteria, to the zooplankton example. The results are collected in Table 4. Row A contains the criterion scaling technique mentioned in the previous section. We use integer scaling for ABUNDANCE, and scale POND and SPECIES by maximizing the sum of the correlations between ABUNDANCE and POND and SPECIES. The quantifications are given in Table 4, for the correlations we find r(S,A) = .29 and r(P,A) = .16. In column B we maximize the correlation r(S,A) by scaling both SPECIES and ABUNDANCE. Of course this gives no quantification for POND. The optimal correlation is r(S,A) = .59. In column C the same is done for r(P,A), which can be increased to .36. Column D is more interesting. It optimizes r(S,A) + r(P,A) over all three quantifications. This gives r(S,A) = .58 and r(P,A) = .33. In this solution 44% of the variance in (scaled) ABUNDANCE is 'explained' by (scaled) SPECIES and POND.

We shall make no attempt to give an ecological interpretation of the scalings found by the techniques. The example is meant only for illustrative purposes. It seems, by comparing columns B, C, and D, that the optimal transformations are not very stable over choice of criterion, which is perhaps not surprising in such a small example. The optimal correlations are much more stable. So is the fact that the

ANALYSIS	A	B	C	D	E
SPECIES	-0.24	-0.20		+0.13	+1.24
	-1.18	-1.00		-0.71	-1.06
	+0.24	-0.90		-0.72	-0.45
	+1.65	+1.12		+1.01	-1.02
	+0.71	+1.84		+1.91	+1.24
	+0.71	-0.30		-0.29	-0.52
	-1.65	-1.16		-1.46	-0.98
	+0.24	+0.20		+0.14	+1.24
POND	-0.22		+0.70	+1.22	+1.38
	+1.94		+0.54	+0.03	+0.83
	-0.75		+0.74	+0.75	-0.94
	-0.75		-1.91	-1.54	-0.96
	-0.21		-0.07	-0.70	-0.84
ABUNDANCE	-0.88	+0.01	+0.12	+0.09	
	-0.20	-1.97	-5.37	-3.34	
	+0.48	-1.49	-0.11	-1.16	
	+1.16	+0.29	-0.23	+0.07	
	+1.84	+2.71	+1.85	+2.62	

Table 4: Various optimal scalings for the zooplankton data
A: criterion scaling: A integer, maximize r(S,A) + r(P,A).
B: maximize r(S,A).
C: maximize r(P,A).
D: maximize r(S,A) + r(P,A).
E: abundance credits solution

categories of ABUNDANCE are scaled in the correct order, except for the zero category which moves to the middle of the abundance scale.

Column E in Table 4 is quite different from the others. This is because it interprets the data as a single bivariate distribution, with 52 'abundance credits' as the units. If we now scale SPECIES and POND optimally, maximizing the correlation in the bivariate distribution, then we find the quantifications in column E, and the optimal correlation equal to .89. Again we give no interpretation, but we point out that the solution in column E can be used to reorder the rows and columns of Table 2 by using the order of the optimal quantifications. In this reordered version of the table the elements are nicely grouped along the diagonal. For more information about such optimal ordering aspects of nonlinear multivariate analysis techniques we refer to Heiser (1986).

In the book by Gifi (1981) special attention is paid to a particular class of criteria, that could be called *generalized canonical analysis* criteria. Also compare Van der Burg, De Leeuw, and Verdegaal (1984, 1986) for an extensive analysis of these criteria, plus a description of *alternating least squares methods* for optimizing them. In generalized canonical analysis the variables are partitioned into *sets of variables*. In ordinary canonical correlation analysis (Gittins, 1985) there are only two sets. In some of the special cases of ordinary canonical analysis, such as multiple regression analysis and discriminant analysis, the second set contains only a single variable. In principal component analysis the number of sets is equal to the number of variables, i.e. each set contains a single variable. The partitioning of the variables into sets induces a partitioning of the dispersion matrix of the variables into dispersion matrices *within sets* and dispersion matrices *between sets* . Suppose S is the dispersion matrix of all variables, and T is the direct sum of the within-set dispersions, i.e. T is a block-matrix with on the diagonal the within-set dispersions, and outside the diagonal blocks of zeroes. In ordinary canonical correlation analysis T consists of two blocks along the diagonal that are nonzero, and two zero blocks outside the diagonal. In principal component analysis T is the diagonal matrix of the variances of the variables. Van der Burg et al. (1985) define the generalized canonical correlations as the eigenvalues of $m^{-1}T^{-1}S$, where m is the number of

sets. In principal component analysis the generalized canonical correlations are the eigenvalues of the correlation matrix, in ordinary canonical analysis they are linearly related to the usual canonical correlation coefficients. Gifi (1981) concentrates on techniques that maximize the sum of the p largest generalized canonical correlation coefficients. These are, of course, functions of the correlation coefficients between the variables. This means that we are dealing with a special case of the previous set-up. But this special case is exceedingly important, because the usual linear multivariate analysis techniques are all forms of generalized canonical analysis.

Measurement level

In the examples we have discussed so far only two possible scalings of the variables were mentioned. Either the quantification of the categories is known, which is the case for measured or numerical variables, or the quantification is completely unknown, and must be found by optimizing the value of the criterion. Binary variables are special, because the quantification is unknown, but irrelevant. The two cases 'completely known' and 'completely unknown' are too extreme in many applications. We may be reasonably sure, for example, that the transformation we are looking for is *monotonic* with the original ordering of the target, which must be an ordered set in this case. Or we may decide that we are not really interested in nonmonotonic transformations, because they would involve a shift of meaning in the interpretation of the variable. If we predict optimally transformed yield, for instance, and the optimal transformation has a parabolic form, then we could say that we do not predict 'yield' but 'departure from average yield'. In such cases it may make sense to restrict the transformation to be *increasing*. The zooplankton example has shown that often monotonicities in the data appear even when we do not explicitly impose monotonicy restrictions.

It is one of the major advantages of our algorithm that it generalizes very easily to optimal scaling with ordinal or monotonic restrictions. It suffices to insert a *monotone regression* operator **MR**(.) in step A2. Thus

For s=1 to m:

A1: compute $\P_s = \Sigma_{t \neq s} \pi_{st} G_t y_t$,

A2: compute $y_s = \mathbf{MR}(D_s^{-1} G_s' \P_s)$,

A3: compute y_s by normalizing y_s,

next s.

We do not explain monotone regression here, but we refer to Kruskal (1964) or Gifi (1981) for details. The basic property we need is that monotone regression does indeed give monotone quantifications, and that it gives the optimum from the set of all such quantifications in each stage.

By this modification of the algorithm we can now analyze at least three types of variables. If we use the **MR**(.)-operator in A2 we impose monotonicity restrictions, and consequently analyze *ordinal* variables. If we use the **LR**(.)-operator, which performs a linear regression of the original values, then we analyze *numerical* variables. And if we use **IR**(.), the identity operator, then we analyze *nominal* variables. In the Legendre and Legendre scheme, discussed earlier, this corresponds with (multi-state) ordered and nonordered variables, while the numerical variables are called quantitative. It is now relative easy to think of other operators which can be used in A2. A very familiar one is **PR**(.), or *polynomial regression*, which fits the optimal polynomial of specified degree. Another one, which is somewhat less familiar, but definitely more useful is **SR**(.), *spline regression*. Splines will be discussed briefly below. As a final example we mention **SM**(.), the linear *smoother* used by Breiman and Friedman (1985) in their ACE-method. The ACE-methods are nonlinear multivariate analysis methods which show great promise, but we do not have enough experience with them to discuss them in any detail. We can also combine monotonicity with the spline or polynomial constraints, and look for the optimal monotone spline or polynomial.

In order to illustrate these new concepts it is, perhaps, time to analyze a somewhat larger example. We have chosen the nitrogen nutrition example from

Gittins (1985, chapter 11). Eight species of grass were given nitrogen treatments of 1, 9, 27, 81, and 243 ppm N by varying the amounts of $NaNO_3$ in a culture solution. Individuals of each species were grown separately in pots under sand culture in an unheated greenhouse using a split-plot experimental design. There were 5 blocks of replications of the complete experiment, and consequently 5 x 5 x 8 = 200 individual pots, which are the natural units in this case. The logarithm of the dry weight yield after a growth period of two months is the outcome variable for this experiment. We do not repeat the data here, but we refer the interested reader to Gittins (1985, appendix A2).

From the point of view of data analysis the most interesting problem seems to be to predict the yield from the knowledge of the species and the nitrogen treatment. The situation is in some respects quite similar to the zooplankton example, because there we also has two orthogonal variables SPECIES and POND that were used to predict ABUNDANCE. The nature of the variables is quite different, however, in this larger example. SPECIES is a *nominal* (or *multi-state unordered*) variable, and NITRO, the amount of nitrogen, is a *numerical* (or *measured*) variable. But NITRO takes on only the five discrete values 1, 9, 27, 81, and 243, and in this respects it differs from the numerical variable YIELD, which can in principle take on a continuum of possible values. In the Legendre and Legendre classification NITRO is discontinuous quantitative, while YIELD is continuous quantitative. This implies that the indicator matrix for YIELD is not very useful. Because of the continuity of the variable each value will occur only once, and the indicator matrix will be a *permutation matrix*, with the number of categories equal to the number of observations. This will make it possible to predict any quantification of YIELD exactly and trivially, and thus the result of our optimal scaling will be arbitrary and not informative. If we want to apply indicator matrices to continuous variables, then we have to group their values into intervals, that is we have to *discreticize* them.

Discreticizing can be done in many different ways, and consequently has some degree of arbitrariness associated with it. Moreover if we plot the orginal variable against the optimal quantified variable, then we always find a step function,

because by definition data values in the same interval of the discretization get the same quantified value. Step functions are not very nice representations of continuous functions. It is very difficult to recognize the *shape* of a function from its step function approximation. On the other hand *polynomials* are far too rigid for satisfactory approximation. This is the main resons for using *splines* in nonlinear multivariate analysis. In order to define a spline we must first choose a number of *knots* on the real line, which have a similar function as the discretization points for step functions. We then fix the *degree* p of the spline. Given the knots and the degree a spline is any function which is a polynomial of degree p between knots, and which has continuous derivatives of degree p - 1 at the knots. Thus a spline can be a different polynomial in each interval, but not arbitrarily different because of the smoothness constraints at the knots, i.e. the endpoints of the intervals. For p = 0 this means that the splines are identical with the step functions, that have steps at each of the knots. For p = 1 splines are piecewise linear, and the pieces are joined continuously at the knots. For p = 2 splines are piecewise quadratic, and continuously differentiable at the knots, and so on. Thus step functions are special splines. If we choose the knots in such a way that all data values are in one interval, then we see that polynomials are also special cases. Thus **SR**(.) has step functions and polynomials as a special case, and **MSR**(.), which is monotone spline regression, includes ordinary monotone regression and monotone polynomials.

We now apply spline regression to the nitrogen example. The transformation for YIELD is restricted to be a piecewise linear spline, with knots at 0, .25, ... , 2.25. Transformations for SPECIES and NITRO are not restricted. If we use integer coding for SPECIES, the values 1, 9, 27, 81, 243 for NITRO, and the original data values for YIELD, we find r(S,Y) = -.47 and r(N,Y) = .42. The squared multiple correlation (SMC) is .3960. With optimal transformation, as specified above, we find and SMC of .7816. The optimal transformation of SPECIES is

(-1.55 -1.31 -0.82 +0.91 +0.78 +0.71 +1.10 +0.18),

and that of NITRO is

(+1.88 -0.06 -0.14 -0.77 -0.92).

Observe that the NITRO scaling is monotonic, but not at all linear. The transformation for YIELD is plotted in Figure 2a. We see that it is roughly monotonic, except for eight pots with small values of yield (less than .50). In fact it is close to linear: the correlation between original and transformed values is -.9694. An inspection of the data, and of the analysis of Gittins in his chapter 11, shows that it is perhaps not entirely reasonable to use the same NITRO transformation for each species. Species 1, 2, and 3 have very similar behaviour, and average YIELD values are nicely monotonic with NITRO, but the other species react much less clearly to the nitrogen treatments. For this reason we have repeated the analysis with two variables. The first one is an interactive combination of SPECIES and NITRO, with 40 categories, and the second one is YIELD. Quantifications of SPECIES x NITRO are derived from the indicator matrix, with 40 columns, and quantifications of YIELD by using the same piecewise linear splines as before. The transformed YIELD is in Figure 2b. It is still almost monotonic, but less linear than the previous transformation. The correlation between observed and transformed values is down to -.9094, the SMC is up to .9339. Figure 3 shows the quantification of SPECIES x NITRO, plotted as eight separate transformations, one for each species. We clearly see the difference between the first three species and the other ones, presumably a difference in sensitivity to the nitrogen content. A clustering of species that suggests itself is [{1,2,3},{4,5,6},{7,8}].

The use of copies

By combining the various criteria with the various options for measurement levels we get a very large number of multivariate analysis techniques. Nevertheless there are some very common techniques, which are still not covered by our developments. The major example is *multiple correspondence analysis* (also known as *homogeneity analysis*, or *Guttman's principal components of scale analysis*). For the details and history of this technique we refer to Nishisato (1980, chapter 5), Gifi (1981, chapter 3), Lebart, Morineaux, and Warwick (1984, chapter 4), and

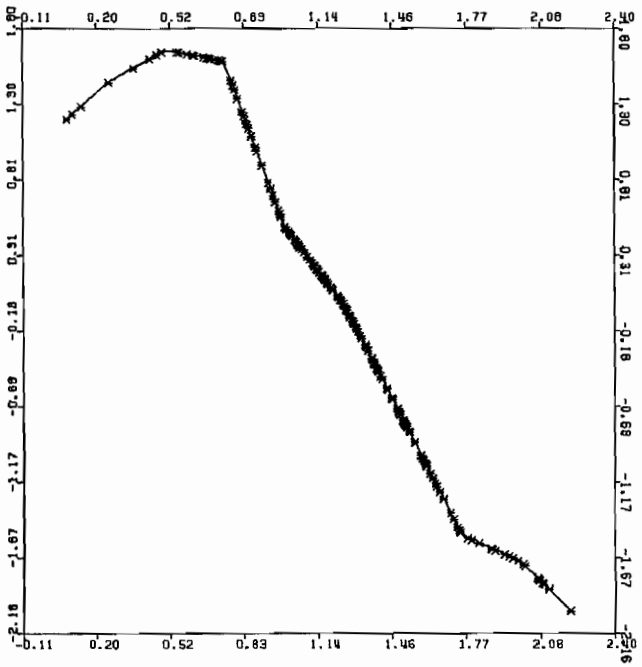

Figure 2a:
YIELD transform
Additive model

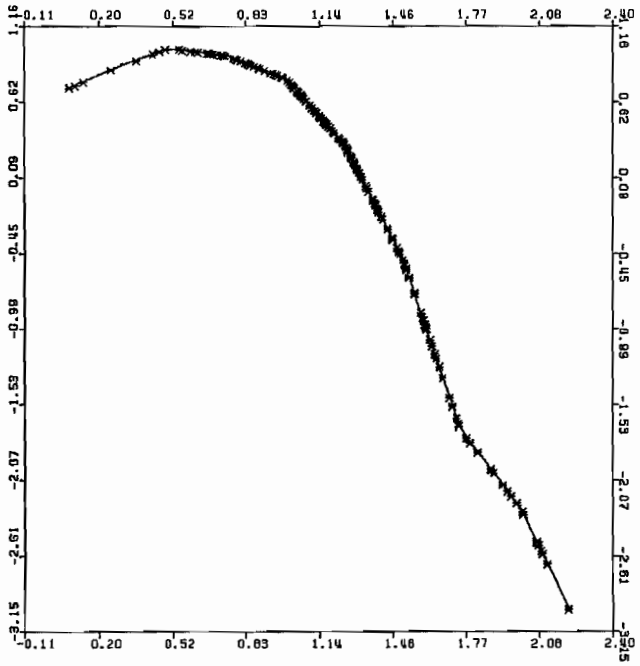

Figure 2b:
Yield transform
Interactive model

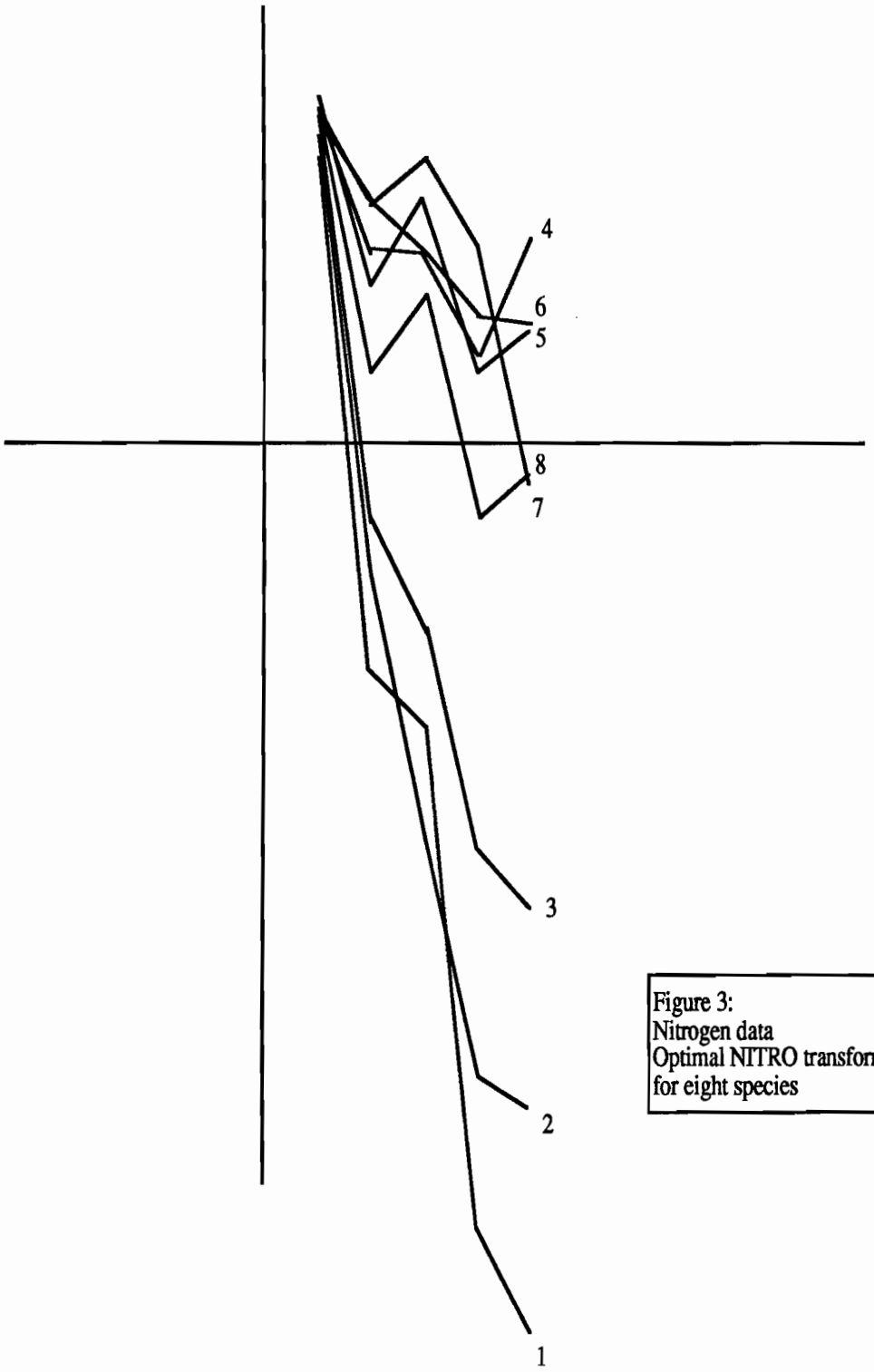

Figure 3:
Nitrogen data
Optimal NITRO transformations
for eight species

Greenacre (1984, chapter 5). In ecology multiple correspondence analysis was already discussed by Hill (1973, 1974), and it is closely related to the popular ordination method called *reciprocal averaging*. We derive the technique here as a form of generalized canonical analysis.

First suppose that we want to find quantifications or transformations of the variables in such a way that the largest eigenvalue of the correlation matrix (i.e. the percentage of variance 'explained' by the first dimension) is maximized. We illustrate this with the zooplankton example, using the ponds as variables ordering the eight species. As indicated by Hill (1974) this amounts to solving the eigenvalue problem

$$Cx = m\mu Dx.$$

Here C is the supermatrix containing all cross tables C_{st}. This optimal scaling problem was originally formulated and solved by Guttman (1941). Matrix C is called the *Burt table* in the French correspondence analysis literature. Matrix D is the diagonal of C, and m is the number of variables. The category quantifications y_t are found by normalizing the m subvectors of the eigenvector x corresponding with the dominant nontrivial eigenvalue. In the zooplankton example C is of order 25, because there are five variables with 5 categories each. The largest eigenvalue, which was 3.41 with integer scaling, goes up to 3.70 with optimal scaling. The percentage variance 'explained' goes from 68% to 74%. Table 5a gives the optimal quantifications for the five variables. They are quite regular and close to monotonic, but distinctly nonlinear.

There are now at least three ways in which the problem can be made multidimensional. In the first place we can compute the induced correlation matrix R, and find its subsequent eigenvalues and eigenvectors as in ordinary metric component analysis. This is straightforward. In the second place we can change the criterion to a multidimensional one. Thus we can maximize the sum of the first two, or the sum of the first three eigenvalues of the correlation matrix. In general this will give different correlation matrices, and different eigenvalue distributions. We

illustrate this for the sum of the first two eigenvalues in the zooplankton example. In the previous solution, which optimized the largest eigenvalue, the first two eigenvalues 'explained' 74% and 14%. If we optimize the sum of the two largest eigenvalues we find 'explained' variances of 56% and 44%. The optimal quantifications in Table 5b make the transformed data matrix exactly of rank two. In order to obtain this perfect fit, the technique transforms variables 3 and 4 in a somewhat peculiar way.

The third way of finding a multidimensional solution is quite different. It simply computes additional eigenvalues and eigenvectors of the pair (C,mD). This defines multiple correspondence analysis. The technique was introduced in psychometrics by Guttman and Burt (Guttman, 1941, 1950, 1953, Burt, 1950, 1953). Each eigenvector now defines a vector of category quantifications, which induces a correlation matrix. In Table 5c, for example, we give the quantifications corresponding with the second eigenvalue of (C,mD), which is 2.55. The correlation matrix that goes with these quantifications has a dominant eigenvalue 'explaining' 51% of the variance, and a subdominant one 'explaining' 35%. The quantifications in Table 5c look peculiar. We could go on, of course, by using additional eigenvalues of (C,mD).

If one thinks about this a little bit, then it is somewhat disconcerting. The multiple correspondence problem in general has $\Sigma (k_t - 1)$ nontrivial eigensolutions, which give an equal number of induced correlation matrices. Applying ordinary metric principal component analysis to each of these correlation matrices gives m times $\Sigma (k_t - 1)$ dimensions. In the zooplankton example there are thus $5 \times (4 + 4 + 4 + 4 + 4) = 100$ dimensions. This is a bit much. Gifi (1981) calls this *data production*, to contrast it with the more common and more desirable concept *data reduction*. Careful mathematical analysis (Gifi, 1981, chapter 11, De Leeuw, 1982, Bekker, 1986) shows that in many cases there are mathematical relationships between the different dimensions, so that they are not independent. This is probably familiar to most ecologists as the *horseshoe* or *Guttman effect*, which makes the second ordination dimension a curved function of the first one. Remember that

category	1	2	3	4	5
variable 1	.77	.00	.00	-1.29	-1.29
variable 2	.85	.00	.65	-1.29	-1.29
variable 3	-.69	.00	.67	.58	2.20
variable 4	-.96	.69	.51	1.33	.00
variable 5	-.96	.00	.57	1.35	.00
variable 1	-.77	.00	.00	1.29	1.29
variable 2	-.77	.00	-.77	1.29	1.29
variable 3	.38	.00	-2.64	.38	.38
variable 4	-.38	-.38	2.64	-.38	.00
variable 5	1.00	.00	-1.00	-1.00	.00
variable 1	.77	.00	.00	-1.29	-1.29
variable 2	1.07	.00	-1.51	-.07	-.07
variable 3	-.21	.00	2.41	-1.40	.03
variable 4	-.42	-.10	2.63	-.42	.00
variable 5	.98	.00	-1.29	-.60	.00

Table 5: Nonlinear Principal Component Analysis
5a: Quantifications maximizing the largest eigenvalue.
5b: Quantifications maximizing the sum of the two largest eigenvalues.
5c: Second dimension multiple correspondence analysis.

Noy-Meir and Whittaker (1978) already mentioned the curving of the dimensions as an important problem for multivariate ordination, and that Hill and Gauch (1980) consider this curvature problem the main shortcoming or correspondence analysis as an ordination technique.

From the principal component point of view multiple correspondence analysis does not solve an optimal scaling problem in the same sense as the other techniques we have discussed. The eigen-equations for (C,mD) are the stationary equations for finding the quantifications optimizing the largest eigenvalue, but additional solutions of these stationary equations only define suboptimal stationary values for this problem. The natural multidimensional generalization of nonlinear principal component analysis is finding a single set of quantifications that maximizes the sum of the first p eigenvalues, and for this problem there are no horseshoe-like complications. On the other hand it is possible to interprete multiple correpondence analysis as a form of generalized canonical analysis. If we think of each category as a binary variable, while the original variables define sets of these binary variables, then a generalized canonical analysis of these m sets is identical to multiple correspondence analysis. With binary variables there is nothing to transform or quantify, and thus we have an essentially linear technique applied to indicator matrices.

A somewhat more satisfactory description is possible by introducing the notion of *copies* (De Leeuw, 1984a). This also means that we define sets of variables using the original m variables, but now a variable is not split up into categories. If we are interested in a two-dimensional solution, for instance, we take two copies of each variable in each of the m sets. We then optimize the sum of the first two generalized canonical correlations over quantifications. Thus a set consists of two identical variables, identical in the sense that the functions ϕ_1 and ϕ_2, mapping Ω into $\Gamma_1 = \Gamma_2$, are the same. Of course the quantifications ψ_1 and ψ_2 can be different, and because the variables are in the same set they will generally be different at the optimum of the criterium. In fact the two quantifications can without loss of generality be chosen to be orthogonal, i.e. we can require $y_1'Dy_2 = 0$. Using

p copies of a variable to define m sets of p variables in this way defines multiple correspondence analysis as a special case of generalized canonical correlation analysis.

But this way of looking at things immediately suggests several useful generalizations. In the first place we can use a different number of copies for different variables. It is reasonable, in many cases, to use copies for unordered multi-state nominal variables only, and to use a single copy for ordinal variables. In the second place the notion of copies can be combined with the various measurement levels we have discussed above. Thus we can require copies to be monotonic (in that case they can not also required to be orthogonal), or we can require that some copies are monotonic, while others are free. If there are two copies of a variable in a set, we can require the first one to be linear, and the second one to be free. And so on. This is again a decision about the coding of a variable. For each variable we have to decide what measurement level we impose, and we also have to decide how many copies of the variable we use. We do not illustrate the use of copies with our zooplankton example, because the solution using the first multiple correspondence analysis dimension (which optimizes the largest eigenvalue of the correlation matrix) is already monotonic, and quite satisfactory. Using rather complicated procedures on such a small example is bound to produce trivial and uninteresting solutions, as the technique that maximizes the sum of the two largest eigenvalues already shows.

The notion of copies is not limited to principal component analysis, i.e. to a generalized canonical correlation problem with only one variable in each set. In other forms of canonical analysis we can use copies as well. In fact we can even decide to include copies of a variable in different sets. If we include a copy in each set, then the largest generalized canonical correlation will be unity, and it will be defined completely by this (quantified) variable. The remaining canonical variables will be orthogonal to the first, i.e. to this quantified variable. Thus using a copy of a variable in each set amounts to performing a partial canonical correlation analysis, with the variables of which copies are used in the sets partialed out. Combining

partitioning into sets with the various measurement levels, and with the notion of copies, gives an even richer class of techniques (De Leeuw, 1984b).

Some computer programs

It is nice to have a number of principles and technical tools that can be used to create very general nonlinear multivariate analysis techniques. But it is perhaps even nicer to know that some of the possible options have already been combined into various series of computer programs, and that these programs are readily available. The ALSOS series of programs comprises programs for analysis of variance, multiple regression, principal component analysis, factor analysis, and multidimensional scaling. An overview is given by Young (1981). The GIFI series has programs for correspondence analysis, multiple correspondence analysis, principal component analysis, canonical correlation analysis, path analysis, and multiple-set canonical analysis. Gifi (1981) has the necessary references. A relative newcomer is the ACE series, discussed in Breiman and Friedman (1985). There are programs for multiple regression, discriminant analysis, time series analysis, and principal component analysis.

The three series of nonlinear multivariate analysis programs differ in many respects, even if they really implement the same technique. The various possibilities of choosing the regression operators differ, the algorithms differ, and the input and output can also be quite different. But it is of course much more important to emphasize what they have in common. All three series generalize existing linear multivariate analysis techniques by combining them with the notion of optimal scaling or transformation. Thus they make them more nonparametric and less model-based, more exploratory and less confirmatory, more data analytic and less inferential.

Discussion and conclusion

We have introduced our nonlinear multivariate analysis techniques without referring to any statistical model. As we briefly indicated in an earlier section our derivations and ideas also apply directly to correlations defined in the *population*, i.e. to the transformation or quantification of random variables. In the book by Gifi (1981) many population models are discussed, and the behaviour of our techniques when they are applied to random samples from such models is also analyzed. For the population models we also refer to Breiman and Friedman (1985) and their discussants, to De Leeuw (1982), and to Schriever (1985). The statistical stability of our techniques can be studied by using asymptotic techniques such as the delta method, and the modern resampling techniques such as the Jackknife and Bootstrap. Gifi (1981) gives examples. Also compare De Leeuw (1984c). Observe that stability is an important consideration here, because we fit many parameters. We must guard against *chance capitalization,* i.e. against the possibility that our results and our interpretations are based on haphazard properties of the sample. Techniques of testing the stability (or significance) of generalized canonical correlations have been discussed by De Leeuw and Van der Burg (1986). Although these techniques for analyzing stability are often expensive computationally, we think that in almost all cases the extra computations are quite worthwhile. A confidence band around a nonlinear transformation, or a confidence ellipsoid around a plane projection give useful additional information, even if the random sampling assumptions do not seem to apply.

Books such as Legendre and Legendre (1983), Gauch (1982), and Gittins (1985) have already shown to ecologists that linear multivariate analysis techniques, if applied carefully, and by somebody having expert knowledge of the subject area in question, can be extremely helpful and powerful tools. It seems to us that combining multivariate exploration with automatic reexpression of variables is an even more powerful tool, which has already produced interesting results in many different scientific disciplines. We think that they show great promise for ecology

too, but we must emphasize that perhaps even more care, and an even more expert knowledge of the ecological problems, is required. Attacking very simple problems with very powerful tools is usually unwise and sometimes dangerous. One does not rent a truck to move a box of matches, and one does not use a chain saw to sharpen a pencil. The techniques we have discussed in this paper are most useful in dealing with large, relatively unstructured, data sets, in which there is not too much prior information about physical or causal mechanisms. In other cases, often better techniques are available. But these other cases occur far less frequently than the standard mathematical statistics or multivariate analysis texts suggest.

References and bibliography

Agresti, A. (1983). **Analysis of Ordinal Categorical Data.** New York: Wiley.

Anderson, T.W. (1958). **An Introduction to Multivariate Statistical Analysis.** New York: Wiley.

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

Benzécri, J.P. et al. (1980). **Pratique de l'Analyse des Données.** (3 vols). Paris: Dunod.

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. (1975). **Discrete Multivariate Analysis: Theory and Practice.** Cambridge, Ma.: MIT-Press.

Breiman, L, Friedman, J.H. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. **J. Am. Statist. Assoc.**, 80., 580-619.

Burt, C. (1950). The Factorial Analysis of Qualitative Data. **British J. Psychol. (Statist. Section)**, 3, 166-185.

Burt, C. (1953). Scale Analysis and Factor Analysis. Comments on Dr. Guttman's Paper. **British J. Statist. Psychol.**, 6, 5-23.

Calliez, F., Pagès, J.-P. (1976). **Introduction à l'Analyse des Données.** Paris: SMASH.

Carroll, J.D., Arabie, P. (1980). Multidimensional Scaling. **Ann. Rev.**

Psychol., 31, 607-649.

De Leeuw, J. (1973). **Canonical Analysis of Categorical Data.** Unpublished dissertation. Reissued DSWO-Press, Leiden, 1984.

De Leeuw, J. (1982). Nonlinear Principal Component Analysis. In H. Caussinus et al. (eds), **COMPSTAT 82.** Wien: Physika Verlag.

De Leeuw, J. (1984a). The Gifi System of Nonlinear Multivariate Analysis. In E. Diday et al. (eds), **Data Analysis and Informatics IV.** Amsterdam: North Holland Publishing Co.

De Leeuw, J. (1984b). **Beyond Homogeneity Analysis.** Report RR-84-08, Department of Data Theory, University of Leiden.

De Leeuw, J. (1984c). **Statistical Properties of Multiple Correspondence Analysis.** Report RR-84-07, Department of Data Theory, University of Leiden.

De Leeuw, J. (1986). **Multivariate Analysis with Optimal Scaling.** Report RR-86-01, Department of Data Theory, University of Leiden.

De Leeuw, J., Heiser, W.J. (1982). Theory of Multidimensional Scaling. In P.R. Krishnaiah and L. Kanal (eds.), **Handbook of Statistics II.** Amsterdam: North Holland Publishing Co.

De Leeuw, J., Meulman, J. (1986). Principal Component Analysis and Restricted Multidimensional Scaling. In W. Gaul, M. Schader (eds.), **Classification as a Tool of Research.** Amsterdam: North Holland Publishing Co.

De Leeuw, J., Van der Burg, E. (1986). The Permutation Distribution of Generalized Canonical Correlations. In E. Diday et al. (eds.), **Data Analysis and Informatics V.** Amsterdam: North Holland Publishing Co.

De Leeuw, J., Van der Heijden, P., Kreft, I. (1984). Homogeneity Analysis of Event-history data. **Methods of Operations Research**,50, 299-316.

De Leeuw, J., Van Rijckevorsel, J. (1980). HOMALS and PRINCALS: some Generalizations of Principal Component Analysis. In E. Diday et al. (eds), **Data Analysis and Informatics.** Amsterdam: North Holland Publishing Co.

De Leeuw, J., Van Rijckevorsel, J, Van der Wouden, H. (1981). Nonlinear Principal Component Analysis using B-splines. **Methods of Operations**

Research, 23, 211-234.

De Leeuw, J., Young, F.W., Takane, Y. (1976). Additive Structure in Qualitative Data. **Psychometrika**, 41, 471-503.

Fisher, R.A. (1941). **Statistical Methods for Research Workers.** (8th edition). Edinburgh: Oliver and Boyd.

Gauch, H.G. (1982). **Multivariate Analysis in Community Ecology.** Cambridge, G.B.: Cambridge University Press.

Gifi, A. (1981). **Nonlinear Multivariate Analysis.** Department of Data theory FSW, University of Leiden. To be reissued by DSWO-Press, 1987.

Gittins, R. (1985). **Canonical Analysis. A Review with Applications in Ecology.** Berlin: Springer.

Gnanadesikan, R., Kettenring, J.R. (1984). A Pragmatic Review of Multivariate Methods in Applications. In H.A. David and H.T. David (eds.), **Statistics: an Appraisal.** Ames: Iowa State University Press.

Greenacre, M.J. (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.

Guttman, L. (1950). The Principal Components of Scale Analysis. In S.A. Stouffer et al. (eds.), **Measurement and Prediction**, Princeton, Princeton University Press.

Guttman, L. (1953). A Note on Sir Cyril Burt's "Factorial Analysis of Qualitative Data". **British J. Statist. Psychol.**, 6, 1-4.

Guttman, L. (1959). Intruduction to Facet Design and Analysis. **Proc. 15th Int. Congress Psychol.** Amsterdam: North Holland Publishing Co.

Haberman, S.J. (1979). **Analysis of Qualitative Data.** (2 vols.). New York: Academic Press.

Heiser, W.J. (1981). **Unfolding Analysis of Proximity Data.** Department of Data Theory, University of Leiden.

Heiser, W.J. (1986). **Shifted Single-peakedness, Unfolding,**

Correspondence Analysis, and Horseshoes. These Proceedings.

Hill, M.O. (1973). Reciprocal Averaging: an Eigenvector Method of Ordination. **J. Ecology**, 61, 237-251.

Hill, M.O. (1974). Correspondence Analysis: a Neglected Multivariate Method. **Appl. Statist.**, 3, 340-354.

Hill, M.O., Gauch, H.G. (1980). Detrended Correspondence Analysis, an Improved Ordination Technique. **Vegetatio**, 42, 47-58.

Hoffman, D.L., Young, F.W. (1983). Quantitative Analysis of Qualitative Data: Applications in Food Research. In H. Martens and H. Russwurm Jr. (eds), **Food Research and Data Analysis.** London: Applied Science Publishers.

Koyak, R. (1985). **Nonlinear Dimensionality Reduction.** Unpublished Ph.D. Thesis. Department of Statistics, University of California, Berkeley.

Kruskal, J.B. (1964). Multidimensional Scaling by Optimizing Goodness-of-Fit to a Nonmetric Hypothesis.**Psychometrika**, 29, 1-28.

Lebart, L., Morineau, A., Warwick, K.M. (1984). **Multivariate Descriptive Statistical Analysis.** New York: Wiley.

Legendre, L., Legendre, P. (1983). **Numerical Ecology.** Amsterdam: Elsevier Scientific Publishing Co.

Mayr, E. (1932). Birds Collected during the Whitney South Sea Expedition. **Amer. Museum. Novitates**, 20, 1-22.

Meulman, J. (1982). **Homogeneity Analysis of Incomplete Data.** Leiden: DSWO-Press.

Muirhead. R.M. (1983). **Aspects of Multivariate Statistical Theory.** New York: Wiley.

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

Noy-Meir, I.,Whittaker, R.H. (1978). Recent Developments in Continuous Multivariate Techniques. In R.H. Whittaker (ed.), **Ordination of Plant Communities.** The Hague: Junk.

Perreault Jr., W.D., Young, F.W. (1980). Alternating Least Squares Optimal Scaling: Analysis of Nonmetric Data in Marketing Research. **J. Marketing**

Research, 17, 1-13.

Schriever, B.F. (1985). **Order Dependence.** Amsterdam: Mathematical Centre.

Takane, Y., Young, F.W., De Leeuw, J. (1979). Nonmetric Common Factor Analysis. **Behaviormetrika**, 6, 45-56.

Takane, Y., Young, F.W., De Leeuw, J. (1980). An Individual Differences Additive Model. **Psychometrika**, 45, 183-209.

Van der Burg, E. (1984). Homals Classification of Whales, Porpoises and Dolphins. In J. Janssen et al. (eds), **New Trends in Data Analysis and Applications**. Amsterdam: North Holland.

Van der Burg, E., De Leeuw, J. (1983). Nonlinear Canonical Correlation. **British Journal of Mathematical and Statistical Psychology**, 36, 54-80.

Van der Burg, E., De Leeuw, J., Verdegaal, R. (1984). **Non-linear Canonical Correlation with M Sets of Variables.** Report RR-84-12, Department of Data Theory, University of Leiden.

Van der Burg, E., De Leeuw, J., Verdegaal, R. (1986). **Homogeneity Analysis with k Sets of Variables.** Submitted for Publication.

Van Rijckevorsel, J. (1982). Canonical Analysis with B-splines. In H. Caussinus et al. (eds), **COMPSTAT 82**. Wien: Physika Verlag.

Van Rijckevorsel, J., Van Kooten, G. (1985). Smooth PCA of Economic Data. **Computational Statistics Quarterly**, 2, 143-172.

Van Rijckevorsel, J., Walter, J. (1983). An Application of two Generalizations of Nonlinear Principal Components Analysis. In J. Janssen et al. (eds), **New Trends in Data Analysis and Applications**. Amsterdam: North Holland.

Young, F.W. (1981). Quantitative Analysis of Qualitative Data. **Psychometrika**, 46, 347-388.

Young, F.W., De Leeuw, J., Takane, Y. (1976). Regression with Qualitative and Quantitative Variables. **Psychometrika**, 41, 505-529.

Young, F.W., De Leeuw, J. , Takane, Y. (1980). Quantifying Qualitative Data. In E.D. Lantermann and H. Feger (eds), **Similarity and Choice**. Bern: Huber.

Young, F.W., Takane, Y., De Leeuw, J. (1978). The Principal Components of Mixed Measurement Level Multivariate Data. **Psychometrika**, 43, 279-281.