4

History of Nonlinear Principal Component Analysis

Jan de Leeuw

CONTENTS

In this chapter we discuss several forms of nonlinear principal components analysis (NLPCA) that have been proposed over the years. Our starting point is that ordinary or classical principal components analysis (PCA) is a well-established technique that has been used in multivariate data analysis for well over 100 years. But PCA is intended for numerical and complete data matrices, and cannot be used directly on data that contain missing, character, or logical values. At the very least, in such cases the interpretation of PCA results has to be adapted, but often we also require a modification of the loss functions or the algorithms.

Multiple correspondence analysis (MCA) is a method similar to PCA that can be and is routinely applied to categorical data. We discuss MCA as a

form of NLPCA, and explain the relationship with classical PCA. In addition, we discuss various other forms of NLPCA, such as linear PCA with optimal scaling, aspect analysis of correlations, Guttman's multidimensional scalogram analysis (MSA), logit and probit PCA of binary data, and logistic homogeneity analysis.

4.1 Linear PCA

Principal components analysis (PCA) is often attributed to Hotelling (1933), but that is surely incorrect. The equations for the principal axes of quadratic forms and surfaces, in various forms, were known from classical analytic geometry. There are some modest PCA beginnings in Galton (1889, pp. 100–102, and Appendix B), where the principal axes are connected for the first time with the 'correlation ellipsoid'.

There is a full-fledged (although tedious) discussion of the technique in Pearson (1901), and there is a complete application (seven physical traits of 3,000 criminals) by a Pearson coworker in MacDonell (1902). The early history of PCA in data analysis, with proper attributions, is reviewed in Burt (1949).

Hotelling's introduction of PCA follows the now familiar route of making successive orthogonal linear combinations of the variables with maximum variance. He does this by using Von Mises (power) iterations, discussed in Von Mises and Pollackzek-Geiringer (1929).

Pearson, following Galton, used the correlation ellipsoid throughout. He casts the problem in terms of finding low-dimensional subspaces (lines and planes) of best (least-squares) fit to a cloud of points, and connects the solution to the principal axes of the correlation ellipsoid.

The data for the problem are n points in $\mathbb{R}^m$, collected in an $n \times m$ matrix $\mathbf{Y}$. We want to find n points in a p-dimensional subspace of $\mathbb{R}^m$ that are close to the n data points. We measure closeness using squared Euclidean distances, which implies we want to minimize SSQ $(\mathbf{Y} - \mathbf{X}\mathbf{B}^T)$ over $n \times p$ matrices $\mathbf{X}$ and $m \times p$ matrices $\mathbf{B}$ with normalization conditions on $\mathbf{X}$ or $\mathbf{B}$. Throughout this chapter we use SSQ( ) for the sum of squares. For $p = 1$ we find the best line, for $p = 2$ the best plane, and so on.

4.2 Simple and Multiple Correspondence Analysis

The history of simple and multiple correspondence analysis is reviewed expertly in Chapter 3 by Lebart and Saporta in this book. We merely give

some additional references that serve to connect multiple correspondence analysis with nonlinear principal components analysis.

4.2.1 Correspondence Analysis

Simple correspondence analysis (CA) of a bivariate frequency table was first discussed, in a rather rudimentary form, by Pearson (1906), by looking at transformations linearizing regressions—see de Leeuw (1983). This was taken up by Hirschfeld (1935), where the technique was presented in a more complete form to maximize correlation and decompose contingency. This approach was later adopted by Gebelein (1941) and by Renyi (1959) and his students in their study of maximal correlation.

Fisher (1938) scores a categorical variable to maximize a ratio of variances (quadratic forms). This is not quite the same as CA, because it is presented in an (asymmetric) analysis of variance context. Both CA and the reciprocal averaging algorithm are discussed, however, in Fisher (1940, Section 3), and applied by his coworker Maung (1941a, 1941b).

Then in the early 1960s the chi-square distance-based form of CA, relating CA to metric multidimensional scaling (MDS), with an emphasis on geometry and plotting, was introduced by Benzécri, and published (with `FORTRAN` code) in the thesis of Cordier (1965).

4.2.2 Multiple Correspondence Analysis

Different weighting schemes to combine quantitative variables into an index that optimizes some variance-based discrimination or homogeneity criterion were proposed in the late 1930s by Horst (1936), Edgerton and Kolbe (1936), and Wilks (1938). Their proposals all lead to the equations for linear PCA.

The same idea of weighting (or quantifying) was applied to qualitative variables in a seminal paper by Guttman (1941), who was analysing qualitative data for the war department. He presented, for the first time, the equations defining multiple correspondence analysis (MCA). The equations were presented in the form of a row-eigen (scores), a column-eigen (weights), and a singular value (joint) problem. The paper introduced, without introducing explicit names, complete disjunctive coding (*codage disjonctif complet* in French), the Burt table (*tableau de Burt* in French), and pointed out the connections with the chi-square metric. There was no geometry, and the emphasis was on constructing a single scale. In fact, Guttman warned explicitly against extracting and using additional eigenpairs.

Guttman (1946) extended scale or index construction to paired comparisons and ranks. Then in Guttman (1950) it was extended to scalable binary items. In the 1950s and 1960s Hayashi introduced the quantification techniques of Guttman in Japan, where they were widely disseminated through the work of Nishisato. Various extensions and variations were added by the Japanese school; see Chapter 3 by Lebart and Saporta in this book for

references. Starting in 1968, MCA was studied as a simple form of metric MDS by de Leeuw (1968, 1973).

Although the equations defining MCA were basically the same as those defining PCA, the relationship between the two remained problematic. These problems were compounded by 'arches', or the Guttman effect (*l'effet Guttman* in French), i.e., by artificial curvilinear relationships between successive dimensions (eigenvectors).

4.3 Forms of Nonlinear Principal Component Analysis

There are various ways in which we can introduce nonlinearity into PCA to obtain what is abbreviated as NLPCA. First, we could seek indices that are nonlinear combinations of variables that discriminate maximally in some sense. For example, we could look for a multivariate polynomial P of the observed variables y_j, with some normalization constraints on the polynomial coefficients, such that $P(y_1, \cdots, y_m)$ has maximum variance. This generalizes the weighting approach of Hotelling. Second, we could find nonlinear combinations of unobserved components that are close to the observed variables. In a polynomial context this means we want to approximate the observed y_j by polynomial functions $P(x_1, \cdots, x_p)$ of the components. This generalizes the reduced rank approach of Pearson. Third, we could look for transformations of the variables that optimize the linear PCA fit. We approximate $T(y_j)$ by the linear combination $\mathbf{Xb}_j$, by fitting both transformations and low-rank approximation. This is known as the *optimal scaling* (OS) approach, a term of Bock (1960).

The first approach has not been studied much, although there may be some relations with item response theory. The second approach is currently popular in computer science, as nonlinear dimension reduction—see, for example, Lee and Verleysen (2007). There is no unified theory, and the papers are usually of the 'well, we could also do this' type familiar from cluster analysis. The third approach preserves many of the properties of linear PCA and can be connected with MCA as well. We shall follow the history of PCA-OS and discuss the main results.

4.4 NLPCA with Optimal Scaling

Guttman (1959) observed that if we require that the regressions between monotonically transformed variables are linear, then these transformations are uniquely defined. In general, however, we need approximations. The loss

function for PCA-OS is SSQ ($\mathbf{Y} - \mathbf{X}\mathbf{B}^{\intercal}$), as before, but now we minimize over components $\mathbf{X}$, loadings $\mathbf{B}$, and also over transformations $\mathbf{Y}$, thereby obtaining NLPCA-OS. Transformations are defined column-wise (over variables) and belong to some restricted class (monotone, step, polynomial, spline). Algorithms often are of the alternating least-squares (ALS) type, where optimal transformation and low-rank matrix approximation are alternated until convergence.

4.4.1 Software

PCA-OS only became interesting after it became computationally feasible. Consequently, the development and availability of software was critical for the acceptance of NLPCA. Shepard and Kruskal used the monotone regression machinery of the nonmetric breakthrough to construct the first PCA-OS programs around 1962. Their paper describing the technique was not published until much later (Kruskal and Shepard, 1974). Around 1970 versions of PCA-OS (sometimes based on Guttman's rank image principle) were developed by Lingoes and Roskam—see Roskam (1968), Lingoes and Guttman (1967), and Lingoes (1973). The rank image principle is an alternative to monotone regression, without clear optimality properties.

In 1973 de Leeuw, Young, and Takane started the alternating least squares with optimal scaling (ALSOS) project, which resulted in PRINCIPALS (Young et al., 1978) and PRINQUAL in SAS (SAS, 1992). In 1980 de Leeuw (with Heiser, Meulman, Van Rijckevorsel, and many others) started the Gifi project (Gifi, 1990), which resulted in PRINCALS (de Leeuw and Van Rijckevorsel, 1980), CATPCA in SPSS (SPSS, 1989), and the R package `homals` (de Leeuw and Mair, 2009a).

Winsberg and Ramsay (1983) published a PCA-OS version using monotone spline transformations. The loss function is the same, but the class of admissible transformations is different. The forms of PCA-OS we have discussed so far use polynomials or step functions, which may or may not be monotonic. Koyak (1987), using the ACE smoothing methodology of Breiman and Friedman (1985), introduced `mdrace`. Again, the class of transformations is different. Transformations in ACE do not necessarily decrease an overall loss function, but they approximate the conditional expectations by using smoothers. Indeed, ACE stands for alternating conditional expectation.

4.5 NLPCA in the Gifi Project

The Gifi project followed the ALSOS project. Its explicit goal was to introduce a system of multivariate analysis methods, and corresponding computer software, on the basis of minimizing a single loss function by ALS algorithms. The techniques that fit into the system were nonlinear regression, canonical analysis, and PCA.

The Gifi loss function is different from the previous PCA loss functions. The emphasis is not on transformation or quantifications of variables and low rank approximation, as in NLPCA. All variables are thought of as categorical, and the emphasis is on the reciprocal averaging and the corresponding geometry of centroids (*principe barycentrique* in French). The loss function is

$$\sigma(\mathbf{X}, \mathbf{Y}) = \sum_{j=1}^{m} \text{SSQ}(\mathbf{X} - \mathbf{Z}_j \mathbf{Y}_j),$$

which must be minimized over $n \times p$ scores $\mathbf{X}$ for the n objects satisfying $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$, and over $k_j \times p$ category quantifications $\mathbf{Y}_j$ of the m variables. The $\mathbf{Z}_j$ are the indicator matrices, coding category membership of the objects (*codage disjonctif complet*), where variable j has k_j categories—see Chapter 11 on MCA by Husson and Josse in this book. Thus, we require that the score of an object or individual in $\mathbf{X}$ is as close as possible, in squared Euclidean distance, to the quantifications of the categories $\mathbf{Y}_j$ that the object falls in.

In the context of generalized canonical analysis, the Gifi loss function is identical to the loss function proposed earlier by Carroll (1968). By using indicator matrices we make the technique identical to MCA, called homogeneity analysis by Gifi, while the various other techniques are special cases resulting from imposing restrictions on the quantifications $\mathbf{Y}_j$. One of the main contributions of Gifi is to show that the transformation approach of NLPCA can actually be fitted into the homogeneity analysis loss function. Interpreting the loss function in terms of Euclidean distance relates homogeneity analysis to nonmetric scaling and nonmetric unfolding.

The basic result that the unconstrained minimization of the Gifi loss function gives MCA follows from the fact that minimizing $\sigma(\mathbf{X}, \mathbf{Y})$ is equivalent to finding the largest eigenvalues of $\mathbf{P}^*$, where $\mathbf{P}^*$ is the average of the projectors $\mathbf{P}_j = \mathbf{Z}_j\mathbf{D}_j^{-1}\mathbf{Z}_j^\top$, with $\mathbf{D}_j = \mathbf{Z}_j'\mathbf{Z}_j$ the diagonal matrix of marginals. This is also equivalent to finding the largest eigenvalues of the generalized eigenvalue problem $\mathbf{BY} = m\mathbf{DY}\boldsymbol{\Lambda}$, where $\mathbf{B} = \mathbf{Z}^\top\mathbf{Z}$, with $\mathbf{Z} = (\mathbf{Z}_1\ \mathbf{Z}_2 \cdots \mathbf{Z}_m)$, is the Burt matrix of bivariate marginals and $\mathbf{D}$ is its diagonal.

NLPCA results by imposing the restriction that $\mathbf{Y}_j = \mathbf{z}_j\mathbf{a}_j^\top$; i.e., the category quantifications are of rank 1. To see what the effect of rank 1 restrictions is, substitute $\mathbf{Y}_j = \mathbf{z}_j^\top\mathbf{a}_j^\top$ into the Gifi loss function. Define $\mathbf{q}_j = \mathbf{Z}_j\mathbf{z}_j$ and normalize so that $\mathbf{q}_j^\top\mathbf{q}_j^\top = \mathbf{z}_j^\top\mathbf{D}_j\mathbf{z}_j^\top = 1$. Then $\sigma(\mathbf{X}, \mathbf{Y}) = m(p - 1) + \text{SSQ}\ (\mathbf{Q} - \mathbf{XA}^\top)$, which is the loss function for linear PCA applied to $\mathbf{Q}$. Thus, the ALS algorithm alternates optimal transformations of the variables in $\mathbf{Q}$ and doing a linear PCA. Further restrictions can require the single quantifications $\mathbf{z}_j$ to be either linear, polynomial, or monotonic functions of the original measurements. The monotonic case gives nonmetric PCA in the classical Kruskal-Shepard sense, and the linear case gives classical linear PCA. In the Gifi approach, and the `homals` program, we can combine different types of transformations for all variables, as well as multiple (unconstrained) and single (constrained) category quantifications.

The `homals` program makes it possible to impose additional so-called additivity constraints on the category quantifications. This allows us to write $\mathbf{Z}_j\mathbf{Y}_j = \mathbf{Z}_{j1}\mathbf{Y}_{j1} + \cdots + \mathbf{Z}_{jv}\mathbf{Y}_{jv}$, and we can interpret each j as a set of variables, not just a single variable. In this way we can incorporate regression, discriminant analysis, canonical analysis, and multiset canonical analysis in the Gifi loss function. Combined with the possibility of scaling each variable linearly, polynomially, or monotonically, and with the possibility of treating each variable as rank constrained or not, this gives a very general system of multivariate analysis methods.

The relation between MCA and NLPCA was further investigated in a series of papers by de Leeuw and his students (de Leeuw, 1982, 1988b, 2006b; Bekker and de Leeuw, 1988; de Leeuw et al., 1999). The research is centred on assuming simultaneous linearizability of the regressions. We assume that separate transformations of the variables exist that make all bivariate regressions linear. The transformations are not necessarily monotone or continuous, and simultaneous linearizability does not say anything about the higher-order multivariate regressions.

Simultaneous linearizability generalizes the result of Pearson (1906) to $m > 2$. It also generalizes the notion (Yule, 1912) of a strained multivariate normal, i.e., a multivariate distribution obtained by applying monotone and invertible transformations to each of the variables in a multivariate normal.

If simultaneous linearizability is satisfied (as it is in the case of two variables, in the case of m binary variables, and in the case of a strained multivariate normal distribution), then MCA can be interpreted as performing a sequence of NLPCAs on a sequence of related correlation matrices. All solutions to the MCA equations are also solutions to the NLPCA equations. This also elucidates the arch or Guttman effect and the role of rank 1 constraints.

4.6 Example

We use the YouthGratitude data from the R package `psychotools` (Zeileis et al., 2012). They are described in detail in the article by Froh et al. (2011). The six 7-point Likert scale variables of the GQ-6 scale are used, with responses from 1,405 students aged 10–19 years. Froh et al. (2011) indicate that classical linear factor analysis shows a one-dimensional structure for the first five items, while they removed the sixth item from further analysis because of a very low factor loading. We keep all six items.

We use the `homals` package for the first analysis. Allowing multiple nominal quantifications for all six variables, i.e., doing an MCA, gives the object scores in Figure 4.1.

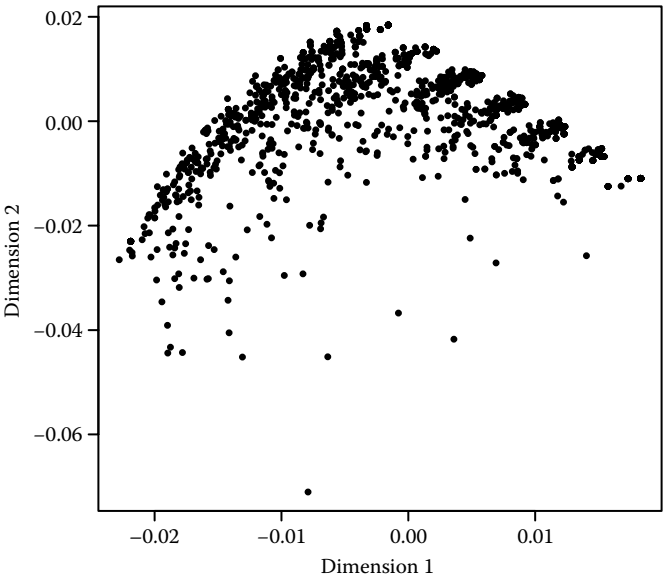


FIGURE 4.1
Object scores multiple nominal.

The figure shows the shape of the point cloud, which is an arch with various subclouds corresponding to closely related profiles.

Figure 4.2 shows the category quantifications, centroids of the corresponding object scores, for variables 1 and 6. The poor discriminatory power of variable 6 is obvious.

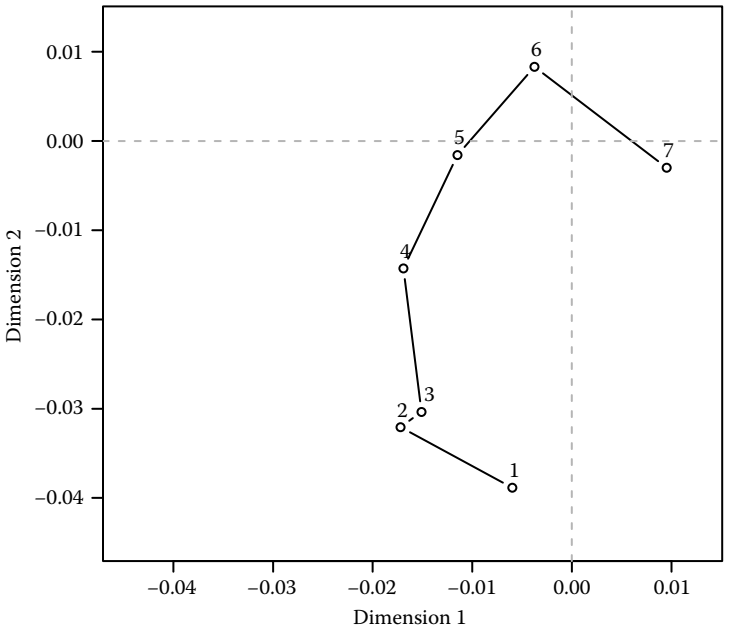
In the next analysis we use single numerical quantifications; i.e., we do a linear PCA. The object scores are in Figure 4.3. The arch has disappeared and the two dimensions appear much less related.

Figure 4.4 gives projection plots for variables 1 and 6. These are joint plots in which component loadings define a direction in space, with category quantifications being lines perpendicular to the loading direction. Object scores are projected on the category they belong to for the variable. The total sum of squares of the projection distances is the loss for this solution.

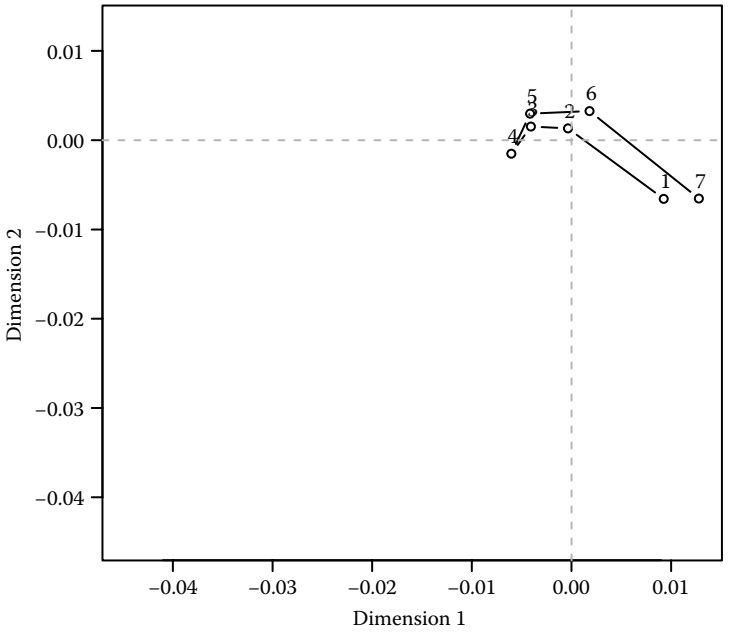
In this case the seven category quantifications are equally spaced perpendicular lines.

In the single ordinal case we allow for monotonic transformations of the category quantifications. This leads to object score plot and projection plots in Figures 4.5 and 4.6.

We see that variable 6 now defines the second dimension, with the loadings almost perpendicular to the loading of the first variable (in fact, the first five variables). Monotone regression creates ties, so we do not see seven category quantification lines any more, and they certainly are not equally spaced.



(a)



(b)

FIGURE 4.2
Category quantifications multiple nominal. (a) Variable 1. (b) Variable 6.

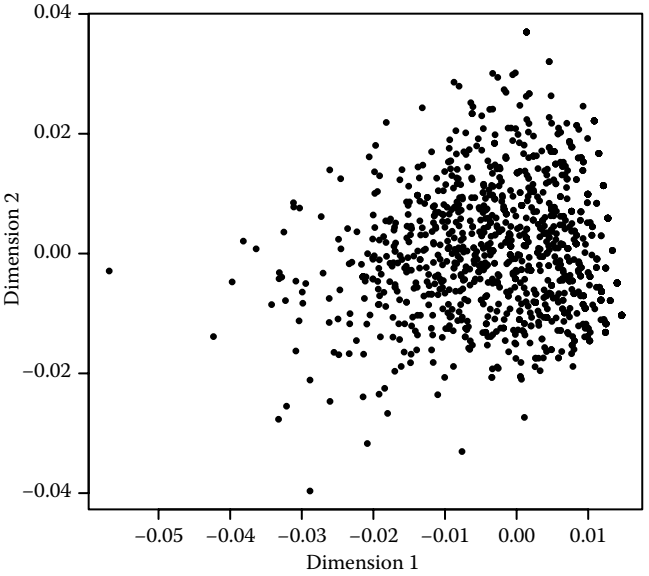


FIGURE 4.3
Object scores single numerical.

4.7 NLPCA Using Pavings

The geometrical approach to NLPCA and MCA has given rise to other related techniques. Suppose we map n objects into low-dimensional Euclidean space, and use a categorical variable to label the points. Each variable defines a partitioning of the points into subsets, which we call a paving. In NLPCA we want these category subsets to be either small (relative to the whole set) or separated well from each other. And we want this for all variables simultaneously. The two objectives are not the same, although they are obviously related.

In MCA we want small subsets, where smallness is defined in terms of total squared Euclidean distance from the category centroids. In PCA-OS we want separation of the subsets by parallel hyperplanes, and loss is defined as squared Euclidean distance to approximate the separating hyperplanes. Total loss measures how well our smallness or separation criteria are satisfied over all variables.

There have been various experiments with different loss functions, based on different ways to measure pavings. For example, Michailidis and de Leeuw (2005) used multivariate medians and sums of absolute deviations to measure homogeneity. Alternatively, de Leeuw (2003) used the length of the

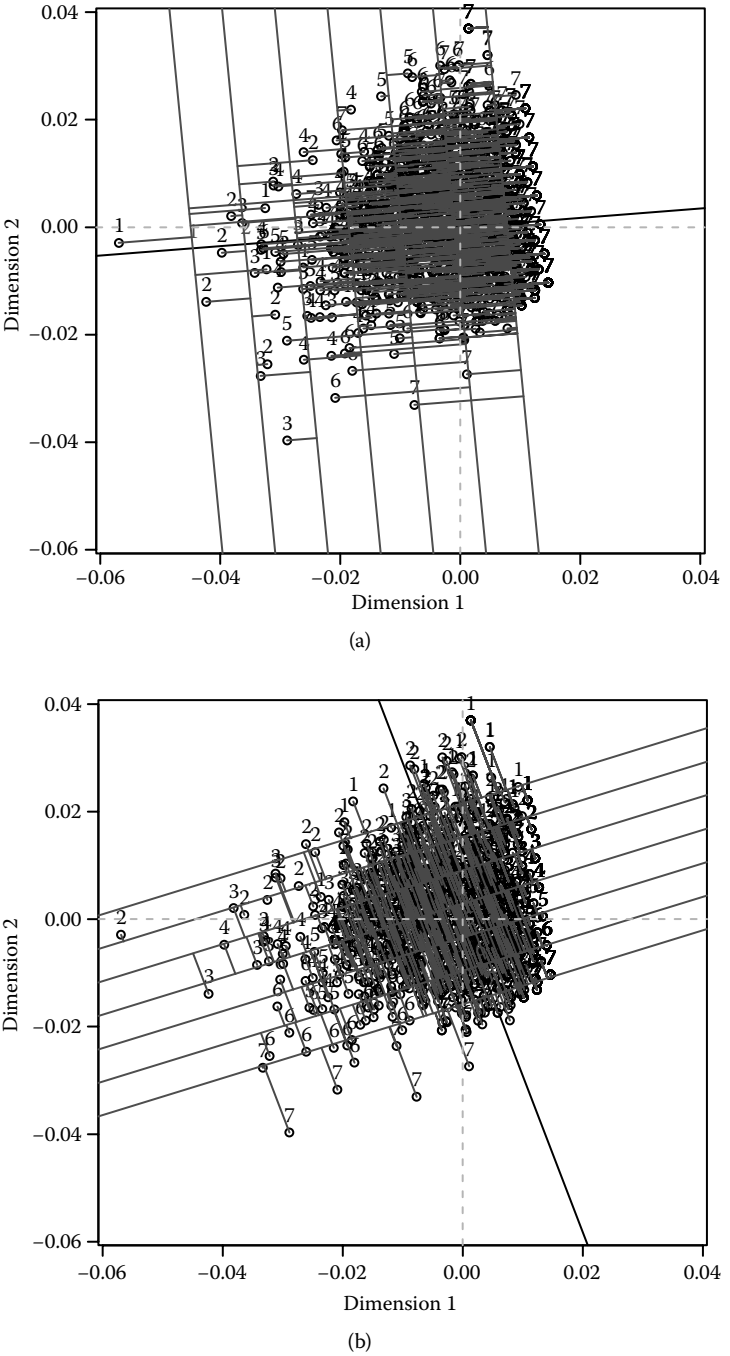


FIGURE 4.4
Projection plots single numerical. (a) Variable 1. (b) Variable 6.

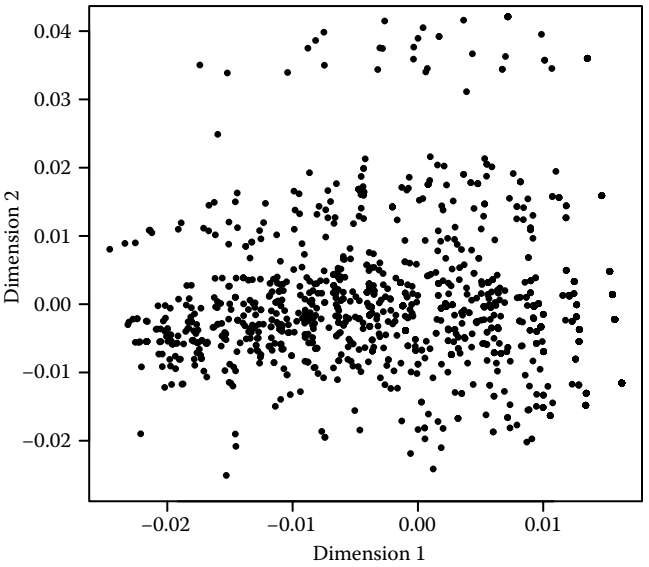


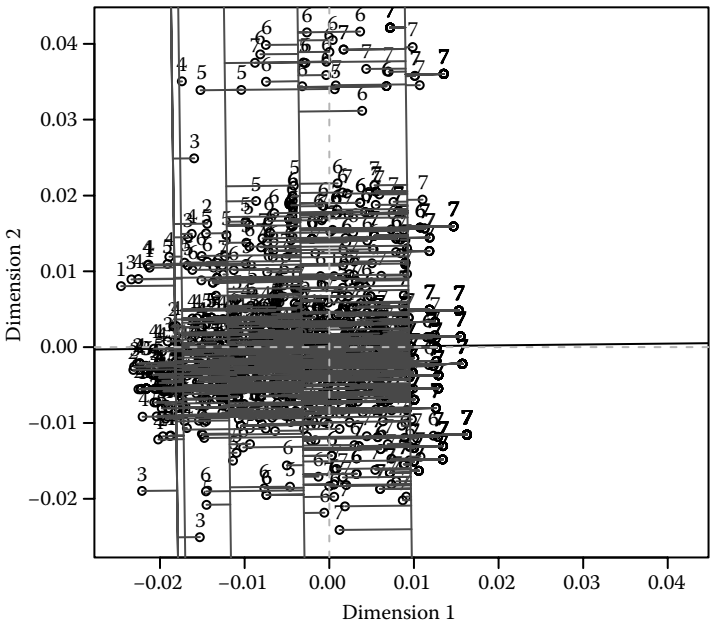
FIGURE 4.5
Object scores single ordinal.

minimum spanning tree defined by each category. Generally these attempts to try alternative loss functions and alternative definitions of homogeneity have not worked out well, seemingly because the sparse indicators do not provide enough constraints to find nontrivial solutions.

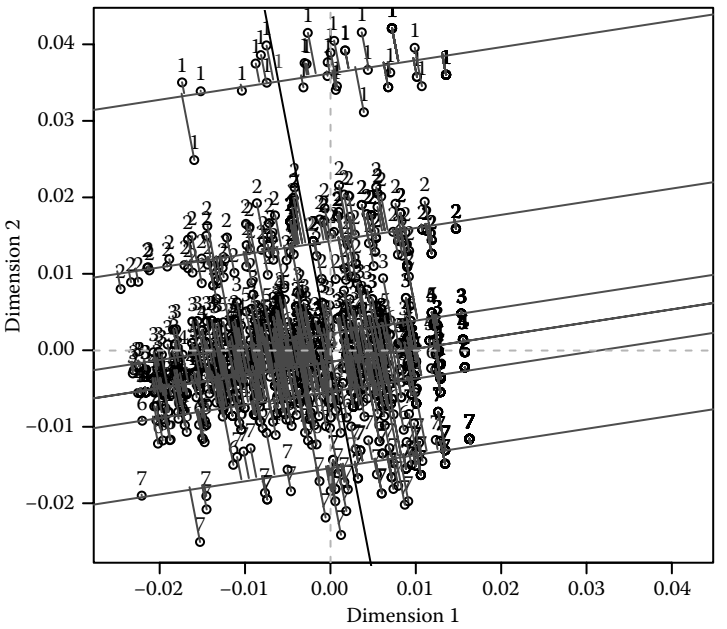
This points to other ways to define separation and homogeneity, which have been explored mostly by Guttman, in connection with his facet theory (cf. Lingoes (1968); see also Chapter 7 by Groenen and Borg in this book). In particular, Guttman's MSA-I can be thought of as a form of NLPCA that has a way of measuring separation by using a pseudotopological definition of inner and outer points of the subsets defined by the categories. There is an illustration in Guttman (1985).

4.8 NLPCA Using Aspects

In Mair and de Leeuw (2010) the R package `aspect` is described. This implements theory from de Leeuw (1988a, 1988b), and gives yet another way to arrive at NLPCA. An aspect is defined as any real-valued function of the matrix of correlation coefficients of the variables. The correlation matrix itself is a function of the quantifications or transformations of the variables.



(a)



(b)

FIGURE 4.6
Projection plots single ordinal. (a) Variable 1. (b) Variable 6.

Many different aspects can be handled. We mention the determinant of the correlation matrix, the squared multiple correlation of one variable with the others, the maximum of the multinormal log-likelihood over a parametric correlation structure, and so on. The `aspect` software maximizes the aspect over transformations, by using majorization methods (de Leeuw, 1994), which are guaranteed to converge if the aspect is a convex function of the correlation matrix.

We briefly indicate why this works. As in other forms of optimal scaling, the correlation matrix $\mathbf{R}(\mathbf{Y}) = \mathbf{Y}^T\mathbf{Y}$ is a function of the standardized transformed or quantified variables. If f is a differentiable convex aspect, we write $\mathbf{D}f(\mathbf{R}(\mathbf{Y}))$ for the $m \times m$ matrix of partial derivatives of f with respect to $\mathbf{R}$, evaluated at $\mathbf{R}(\mathbf{Y})$. Then $f(\mathbf{R}(\mathbf{Y})) \geq f(\mathbf{R}(\tilde{\mathbf{Y}})) + \text{trace}[\mathbf{D}f(\mathbf{R}(\tilde{\mathbf{Y}}))(\mathbf{R}(\mathbf{Y}) - \mathbf{R}(\tilde{\mathbf{Y}}))]$, because a convex function is above its tangents. This means that if $\tilde{Y}$ is the current solution, and $\hat{\mathbf{Y}}$ maximizes the quadratic $\text{trace}[\mathbf{Y}\mathbf{D}f(\mathbf{R}(\tilde{\mathbf{Y}}))\mathbf{Y}^T]$ over normalized transformations $\mathbf{Y}$, then $f(\mathbf{R}(\hat{\mathbf{Y}})) \geq f(\mathbf{R}(\tilde{\mathbf{Y}}))$, and we have increased the aspect. Iterate these steps and we have a convergent algorithm.

In MCA the aspect is the largest eigenvalue, and $\mathbf{D}f(\mathbf{R}(\mathbf{Y})) = \mathbf{v}\mathbf{v}$, with $\mathbf{v}$ the corresponding normalized eigenvector. Each MCA dimension provides a stationary value of the aspect. In PCA-OS the aspect is the sum of the largest p eigenvalues, and $\mathbf{D}f(\mathbf{R}(\mathbf{Y})) = \mathbf{V}\mathbf{V}^T$, with the columns of $\mathbf{V}$ containing the eigenvectors corresponding to the p largest eigenvalues. We can also easily define regression and canonical analysis in terms of the aspects they optimize.

For our example of the six gratitude variables we applied four different aspects: the largest eigenvalue, the sum of the two largest eigenvalues, the determinant, and the sum of the correlations. We give the eigenvalues of the correlation matrices corresponding to these aspects in Table 4.1.

The differences between the four solutions are obviously very small, which gives us confidence in the optimal scaling of the categories that are computed.

In addition, the R package also has a loss function defined as the sum of the differences between the ½m(m–1) correlation ratios and squared correlation coefficients. Minimizing this loss function quantifies the variables to optimally linearize all bivariate regressions, close to the original objective of Pearson (1906) and Guttman (1959).

TABLE 4.1

Eigenvalues for Different Aspects

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
Largest eigenvalue	3.39	0.83	0.60	0.48	0.43	0.27
Largest two eigenvalues	3.19	1.07	0.56	0.48	0.43	0.26
Determinant	3.38	0.84	0.61	0.48	0.43	0.26
Sum of correlations	3.39	0.83	0.60	0.48	0.43	0.27

4.9 Logit and Probit PCA of Binary Data: Gifi Goes Logistic

The idea of using separation as a basis for developing NLPCA has been popular in social science. Let's consider binary data first, using some old ideas of Coombs and Kao (1955). Think of politicians voting on a number of issues. We want to map the politicians as points in low-dimensional space in such a way that, for all issues, those voting in favour can be linearly separated by those voting against. Techniques based on this idea have been developed by political scientists such as Poole and Rosenthal (1985) and Clinton et al. (2004).

A general class of NLPCA techniques for binary data, using logit or probit likelihood functions, in combination with majorization algorithms was initiated by de Leeuw (2006a). The basic idea for defining the loss function is simple. Again, we use the idea of an indicator matrix. Suppose variable j has an $n \times k_j$ indicator matrix $\mathbf{Z}_j$. Let us assume the probability that individual i chooses alternative ℓ for variable j is proportional to $\beta_{j\ell}\exp\{\phi(\mathbf{x}_j, \mathbf{y}_{j\ell})\}$, where ϕ is either the inner product, the negative Euclidean distance, or the negative squared Euclidean distance between vectors $\mathbf{x}_i$ and $\mathbf{y}_{j\ell}$. The $\beta_{j\ell}$ are bias parameters, corresponding to the basic choice probabilities in the Luce or Rasch models. Assuming independent residuals, we can now write down the negative log-likelihood and minimize it over object scores and category quantifications. The negative log-likelihood is

$$\mathcal{L} = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell=1}^{k_j} z_{ij\ell}\log\left\{\frac{\beta_{j\ell}\exp\{\phi(\mathbf{x}_i, \mathbf{y}_{j\ell})\}}{\sum_{v=1}^{k_j}\beta_{jv}\exp\{\phi(\mathbf{x}_i, \mathbf{y}_{jv})\}}\right\}.$$

Majorization (see Chapter 7 by Groenen and Borg in this book) allows us to reduce each step to a principal component (if ϕ is the inner product) or multidimensional scaling (if ϕ is negative distance or squared distance) problem.

This formulation allows for all the restrictions on the category quantifications used in the Gifi project, replacing least squares by maximum likelihood and ALS by majorization (de Leeuw, 2005). Thus, we can have multiple and single quantifications, polynomial and monotone constraints, as well as additive constraints for sets of variables. This class of techniques unifies and extends ideas from ideal point discriminant analysis, maximum likelihood correspondence analysis, choice models, item response theory, social network models, mobility tables, and many other data analysis areas.

4.10 Conclusion

NLPCA can be defined in various ways, but we have chosen to stay close to MCA, mostly by using the rank constraints on the category quantifications in the Gifi framework. The transformation approach to NLPCA, which was developed in the nonmetric scaling revolution, generalizes naturally to the aspect approach. The MDS approach to scaling categorical variables, which inspired the Gifi loss function, can be generalized to various geometric definitions of homogeneity and separation, implemented in both the pavings approach and the logit approach.

The logit and probit approach to categorical data analysis is a promising new development. It can incorporate the various constraints of the Gifi framework, but it also allows us to unify many previous approaches to categorical data proposed in statistics and the social sciences. Both the aspect framework and the logit and probit framework show the power of majorization algorithms for minimizing loss functions that cannot be tackled directly by the alternating least squares methods of Gifi.