NONLINEAR PRINCIPAL COMPONENT ANALYSIS AND RELATED TECHNIQUES

JAN DE LEEUW

1. Introduction

Principal Component Analysis (PCA from now on) is a multivariate data analysis technique used for many different purposes and in many different contexts. PCA is the basis for low rank least squares approximation of a data matrix, for finding linear combinations with maximum or minimum variance, for fitting bilinear biplot models, for computing factor analysis approximations, and for studying regression with errors in variables. It is closely related to simple correspondence analysis (CA) and multiple correspondence analysis (MCA), which are discussed in Chapters XX and YY of this book.

PCA is used wherever large and complicated multivariate data sets have to be reduced to a simpler form. We find PCA in microarray analysis, medical imaging, educational and psychological testing, survey analysis, large scale time series analysis, atmospheric sciences, high-energy physics, astronomy,

Date: July 8, 2005.

and so on. Jolliffe [2002] is a comprehensive overview of the theory and applications of classical PCA.

2. Linear PCA

Suppose we have measurement of n *objects* or *individuals* on m variables, collected in an $n \times m$ matrix $X = \{x_{ij}\}$. We want to have an approximate representation of this matrix in p-dimensional Euclidean space. There are many seemingly different, but mathematically equivalent, ways to define PCA. We shall not dwell on each and every one of them, but we consider the one most relevant for the nonlinear generalizations of PCA we want to discuss.

Our definition of PCA is based on approximating the elements of the data matrix X by the inner products of vectors in $\mathbb{R}^p$. We want to find n vectors a_i corresponding with the objects and m vectors b_j corresponding with the variables such that $x_{ij} \approx a_i'b_j$. The elements of the $n \times p$ matrix A are called *component scores*, while those of the $m \times p$ matrix B are *component loadings*.

We measure degree-of-approximation by using the least squares loss function

$$\sigma(A, B) = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - a_i'b_j)^2. \tag{1}$$

PCA is defined as finding the scores A and the loadings B that minimize this loss function. Another way of formulating the same problem is that we want to find p new unobserved variables, collected in the columns of A, such that the observed variables can be approximated well by linear combinations of these unobserved variables.

It is well known, since Householder and Young [1938], that the solution of this problem can be found by first computing the singular value decomposition $X = K\Lambda L'$, then truncating the singular value decomposition by only keeping the largest p singular values Λ_p and corresponding singular vectors K_p and L_p, and then by setting $\hat{A} = K_p\Lambda_p^{1/2}S$ and $\hat{B} = L_p\Lambda_p^{1/2}T$, where S and T are any two non-singular matrices of order p satisfying $ST' = I$. The minimum value of the loss function is equal to

$$\sigma(\hat{A}, \hat{B}) = \sum_{s=p+1}^{m} \lambda_s^2(X), \tag{2}$$

where the $\lambda_s(X)$ are the ordered singular values of X (so that λ_s^2 are the ordered eigenvalues of both $X'X$ and XX').

We illustrate this with an example, similar to the box problem in Thurstone [1947, Page 140]. We use 20 rectangles and describe them in terms of seven variables (the base, the height, the diagonal, the area, the circumference, the ratio of base to height, and the ratio of height to base). The data matrix, in which base and height are uncorrelated, is given in Table 1. The PCA model fits excellently in two dimensions (99.6% of the sum of squares is "explained"). A plot of the data and the fitted values is in Figure 1.

The representation in Figure 2 nicely reproduces the V-shape of the base-height plot. In this plot we have followed the biplot conventions from Gower and Hand [1996], in which loadings are plotted as directions on which we can project the scores. We see, for example, that the last ten rectangles have the same projection on the circumference direction, and that the base/height and height/base directions are very similar, because these two variables have a high negative correlation of -0.74.

3. Least Squares Nonlinear PCA

3.1. **Introduction.** When we talk about nonlinear PCA in this chapter, we have a specific form of nonlinearity in mind. PCA is a *linear* technique, in the sense that observed variables are approximated by linear combinations of principal components. It can also be a *bilinear* technique, in the sense that elements of the data matrix are approximated by inner products, which are bilinear functions of component scores and component loadings. The nonlinearities in the forms of PCA that we discuss are introduced as nonlinear transformations of the variables, and we still preserve the basic (bi)linearity of PCA. We do not discuss techniques in which the observed variables are approximated by nonlinear functions of the principal components.

Nonlinear PCA is used, for instance, if we do not have actual numerical values as our data but each variable merely ranks the objects. The prototypical example of data of this form are preference rank orders, in which the variables are actually individuals ranking a number of objects in order of preference. In other examples, similar to MCA, variables are categorical and partition the objects into a finite number of sets or categories. Binary variables (true/false, yes/no, agree/disagree, and so on) are a very common special case of both ordinal and categorical variables. And in yet other examples variables may have numerical values but we want to allow for the possibility of computing transformations to improve the fit of the bilinear model.

We have seen in the previous section that we evaluate fit of PCA in p dimensions by computing the sum of squares of the residual singular values X (or the sum of the residual eigenvectors of the product moment matrix $X'X$). This makes it natural to look for transformations or quantifications of the variables that minimize the same criterion. Thus we do not merely minimize loss over component scores A and component loadings B, but also over the *admissible transformations* of the columns of X. The loss function becomes

$$\sigma(A, B.X) = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - a_i'b_j)^2 \tag{3}$$

and we minimize, in addition, over $x_j \in \mathcal{X}_j$, where $\mathcal{X}_j \subseteq \mathbb{R}^n$ are the admissible transformations for variable j. By using (2), this is the same as finding

$$\min_{x_j \in \mathcal{X}_j} \sum_{s=p+1}^{m} \lambda_s(X). \tag{4}$$

This form of nonlinear PCA, in the special case of monotone transformations, has been proposed by, among others, Lingoes and Guttman [1967]; Roskam [1968]; Kruskal and Shepard [1974].

The notion of admissible transformation needs some additional discussion. We have already mentioned the class of monotone transformations as an important example. But other examples can also be covered. We could, for instance, allow low-order polynomial transformations for all or some of the variables. Or, combining the two ideas, monotone polynomials. We could also look for convex or concave transformations, increasing or not. Or for low-order splines on a given knot sequence, which again may or may not be restricted to be monotone. For categorical variables with a small number of categories we may simply allow the class of all possible transformations, which is also known as the class of *quantifications* in which category labels are replaced by real numbers. Nonlinear PCA has been extended to these wider classes of admissible transformations by Young et al. [1978]; Gifi [1990].

All the special cases of transformations so far are covered by the general restriction that the transformed variable must be in a convex cone $\mathcal{K}$ in $\mathbb{R}^n$. Convex cones are defined by the conditions that $x \in \mathcal{K}$ implies $\alpha x \in \mathcal{K}$ for all real $\alpha \geq 0$ and $x \in \mathcal{K}$ and $y \in \mathcal{K}$ implies $x + y \in \mathcal{K}$. It is easy to see that all classes of transformations discussed above are indeed convex cones. In fact some of them, such as the low-order polynomials and splines, are linear subspaces, which are special cones for which $x \in \mathcal{K}$ implies $\alpha x \in \mathcal{K}$ for all real α.

It is also clear that if a transformation x is in one of the cones mentioned above, then a positive linear function $\alpha x+\beta$ with $\alpha \geq 0$ is in the cone as well. As a consequence of this we need to normalize our transformations, both to identify them and to prevent the trivial solution in which all transformations are identically set to zero. Another way of saying this is that we redefine our cones to consist only of centered vectors, and we want all transformations x to be on the unit sphere $\mathcal{S} = \{x \in \mathbb{R}^n \mid x'x = 1\}$. Thus the sets of admissible transformations $\mathcal{X}_j$ are of the form $\mathcal{K}_j \cap \mathcal{S}$, where $\mathcal{K}_j$ is a convex cone of centered vectors.

The normalizations we use imply that the product moment matrix $X'X$ is actually the *correlation matrix* of the variables. Thus the optimization problem for nonlinear PCA in p dimensions is to find admissible transformations of the variables in such a way that the sum of the $n - p$ smallest eigenvalues of the correlation matrix is minimized, or, equivalently, such that the sum of the p largest eigenvalues is maximized. We write our nonlinear PCA problem in the final form as

$$\max_{x_j \in \mathcal{K}_j \cap \mathcal{S}} \sum_{s=1}^{p} \lambda_s(R(X)) \tag{5}$$

where the real-valued function ϕ is defined as the sum of the p largest eigenvalues of the correlation matrix $R(X)$.

This seems a natural and straightforward way to generalize PCA. Allowing for nonlinear transformations of the variables makes it possible to concentrate more variation in the first few principal components. Instead of looking at high-dimensional projections we can look at low-dimensional projections together with plots of the non-linear transformations that we compute [De Leeuw and Meulman, 1986].

3.2. **Aspects.** Instead of tackling the problem (5) directly, as in done in most earlier publications, we embed it in a much larger family of problems for which we then construct a general algorithm. Let us look at problem (5) in which we maximize any convex function ϕ of the correlation matrix. Not just the sum of the p largest eigenvalues, but any convex function. We call any convex real valued function defined on the space of correlation matrices an *aspect* of the correlation matrix [De Leeuw, 1988, 1990].

Of course we first have to show that, indeed, the sum of the p largest eigenvalues is a convex function of the correlation matrix. For this we use the very useful lemma that if $f(x, y)$ is convex in x for every y, then $g(x) = \max_y f(x, y)$ is also convex in x. The sum of the p largest eigenvalues of a matrix R is the maximum of $\mathbf{tr}\ K'RK$ over all $n \times p$ matrices K with $K'K = I$. Thus the aspect is the pointwise maximum of a family of functions linear, and thus convex, in R, and the lemma applies.

We take the opportunity to give some additional examples of convex aspects that illustrate the great generality of our approach. A very simple aspect is the sum of the correlation coefficients. It doesn't use eigenvalues to measure how closely variables are related, but it does measure the strength of the overall relationships. Related aspects are the sum of even powers of the correlation coefficients, or the sum of odd powers of the absolute values of the correlation coefficients. Observe that the sum of squares of the correlations coefficients is actually equal to the sum of squares of the eigenvalues of the correlation matrix. Because the sum of the eigenvalues is a constant, maximizing the sum of squares is the same as maximizing the variance of the eigenvalues. This aspect gives another way to concentrate as much of the variation as possible in the first few principal components.

3.3. **Algorithm.** The algorithm we propose is based on the general *principle of majorization*, which we explain in Appendix A. Using convexity of the aspect ϕ, and the fact that a convex function is always above its tangents, gives the inequality

$$\phi(R(X)) \geq \phi(R(Y)) + \sum_{1\leq i\neq j\leq n}\sum \left.\frac{\partial\phi}{\partial r_{ij}}\right|_{R=R(Y)} (x_i'x_j - y_i'y_j) \tag{6}$$

for all matrices X and Y of normalized admissible transformations. The normalization ensures that the diagonal terms in the double sum on the right disappear.

Each step in the majorization algorithm requires us to maximize the right-hand side of (6). We do this by *block relaxation*, that is by maximizing over one transformation at the time, keeping the other transformations fixed at their current values [De Leeuw, 1994]. Thus in each iteration we solve m of these *optimal scaling* problems, transforming or quantifying each of the variables in turn.

By separating out the part of (6) that depends only on x_j, we find that each optimal scaling problem amounts to solving a least squares problem of the form

$$(7) \qquad \min_{x_j \in \mathcal{K}_{|} \cap \mathcal{S}} (x_j - \tilde{x}_j^{(k)})'(x_j - \tilde{x}_j^{(k)}).$$

Here $\tilde{x}_j^{(k)}$ is the current *target*, defined by

$$\tilde{x}_j^{(k)} = \sum_{\ell < j} g_{j\ell}^{(k,j)} x_\ell^{(k+1)} + \sum_{\ell > j} g_{j\ell}^{(k,j)} x_\ell^{(k)},$$

and the matrices $G^{(k,j)}$ are the partial derivatives, evaluated while updating variable j in iteration k. Thus

$$g_{j\ell}^{(k,j)} = \left.\frac{\partial \phi}{\partial r_{j\ell}}\right|_{R=R(x_1^{(k+1)},\cdots,x_{j-1}^{(k+1)},x_{j+1}^{(k)}\cdots,x_m^{(k)})}.$$

The formula looks complicated, but the only thing it does is keep track of the iteration indices. If we have an expression for the partial derivatives, and a way to solve the least squares problem in (7), then we have a simple and general way to maximize the corresponding aspect. From the software point of view, we can write a high level algorithm that uses subroutines to compute aspects and their partial derivatives as arguments.

If the aspect we use is the sum of the correlation coefficients, then all elements of $G^{(k,j)}$ are equal to $+1$, and thus the target is just the sum of all variables (except for the one we are updating). If the aspect is a single correlation coefficient in the matrix, say $r_{j\ell}$, then the target when updating x_j will be x_ℓ, and vice versa. In the general case we have to recompute the correlations and the partials after updating each variable. This may be expensive computationally. If our aspect is the classical nonlinear PCA sum of the p largest eigenvalues, for instance, then

$$\frac{\partial\phi}{\partial R} = KK',$$

with K the normalized eigenvectors corresponding with the p largest eigenvalues of R. Computing the partials means solving an eigenvalue problem. De Leeuw [1990] discusses some (minor) variations of the algorithm which allow for updating all variables before recomputing the correlations and the partial derivatives.

It is also shown in De Leeuw [1990] that (7) can be minimized by first projecting on the cone, thus ignoring the normalization constraint, and then normalizing afterwards. Generally, such cone projection problems are simple to solve. In the categorical case, for instance, we merely have to compute category averages. In the monotone case, we must perform a monotone regression to project the target on the cone [De Leeuw, 2005]. In the polynomial case, we must solve a polynomial regression problem.

3.4. **Relation with Multiple Correspondence Analysis.** MCA is a special case of our general aspect approach. It corresponds with maximizing the largest eigenvalue of the correlation matrix (and with the case in which all variables are categorical). As shown in Chapter XX, MCA solves the generalized eigen-problem for the Burt matrix. This corresponds with finding the stationary values of the Rayleigh quotient

$$\lambda(a) = \frac{\sum_{j=1}^{m}\sum_{\ell=1}^{m} a_j'C_{j\ell}a_\ell}{m\sum_{j=1}^{m} a_j'C_{jj}a_j}$$

Change variables by letting $a_j = \alpha_j y_j$, where $y_j'C_{jj}y_j = 1$. Then

$$\lambda(\alpha, y) = \frac{\alpha' R(y)\alpha}{m\alpha'\alpha},$$

where $R(y)$ is the correlation matrix *induced* by the quantifications in a. Clearly

$$\max_y \max_\alpha \lambda(\alpha, y) = \max_y \lambda_{max}(R(y)),$$

which is what we wanted to show.

Thus the dominant MCA solution gives us the quantifications maximizing the largest eigenvalue aspect. And the largest eigenvalue of the induced correlation matrix is the largest eigenvalue of the MCA problem. But what about the remaining MCA solutions ? They provide additional solutions of the stationary equations for maximizing the largest eigenvalue aspect, corresponding with other non-global minima, local maxima, and saddle points. As was pointed out very early on by Guttman [1941], the first MCA solution should be distinguished clearly from the others, because the others correspond with suboptimal solutions of the stationary equations. In fact, each MCA eigenvector a has its own associated induced correlation matrix. And each MCA eigenvalue is an eigenvalue (and not necessarily the largest one) of the correlation matrix induced by the corresponding MCA eigenvector.

It goes without saying that simple correspondence analysis or CA is the special case in which we only have two variables, and both are categorical. The correlation matrix has only one non-constant element, and all reasonable aspects will be monotone functions of that single correlation coefficient. Maximizing the aspect will give us the maximum correlation coefficient, and the CA solutions will be the transformations solving the stationary equations of the maximum correlation problem.

3.5. **Relation with Multiple Regression.** Multiple regression and PCA are quite different techniques, but nevertheless there are some important relationships. Consider the PCA problem of maximizing the sum of the $m-1$ largest eigenvalues of the correlation matrix. This is the same, of course, as minimizing the smallest eigenvalue, and thus it can be interpreted as looking for a singularity in the transformed data matrix. This is generally known as *principal component regression*, and its dates back to Pearson [1901]. It is a form of regression analysis, except that in the usual regression analysis we single out one variable as the criterion and define the rest as the predictors and we measure singularity by finding out if the criterion is in the space spanned by the predictors.

More precisely, the squared multiple correlation coefficient of variable j with the remaining $m-1$ variables can be written as

$$\phi(R(X)) = \max_{\beta} 1 - \beta' R\beta$$

where the vector β is restricted to have $\beta_j = 1$. By the lemma we used before, this is a convex function of R, which can be maximized by our majorization algorithm. The partials are simply

$$\frac{\partial\phi}{\partial R} = -\beta\beta'.$$

This can be easily extended to the sum of all m squared multiple correlation coefficients of each variable with all others, which has been discussed in the context of factor analysis by Guttman [1953] and others.

3.6. **Relation with Structural Equation Modeling.** So far, we have written down our theory for the case in which we are maximizing a convex aspect. Of course exactly the same results apply for minimizing a concave aspect. Some aspects are more naturally discussed in this form.

Consider, for example, the determinant of the correlation matrix. Minimizing the determinant can also be thought of as looking for a singularity, i.e. as yet another way of approaching regression. The representation

$$\log \|R\| = \min_{\Gamma \gtrsim 0} \log \|\Gamma\| + \mathbf{tr}\, \Gamma^{-1} R - m,$$

where $\Gamma \gtrsim 0$ means we require Γ to be positive semi-definite, shows that the logarithm of the determinant is a concave function of the correlation matrix. Also

$$\frac{\partial\phi}{\partial R} = R^{-1},$$

which means that the target for updating a variable is its *image*, in the sense of Guttman [1953], the least squares prediction of the variable from all others. Minimizing the determinant can be done by sequentially projecting images on cones of admissible transformations.

In the same way, the aspect

$$\phi(R(X)) = \min_{\theta} \log \|\Gamma(\theta)\| + \mathbf{tr}\ \Gamma^{-1}(\theta)R(X)$$

is the maximized multinormal log-likelihood of a parametric model for the correlations. Such aspects are commonly used in structural equation modeling (SEM) of correlation matrices. The aspect is concave in R, with partials $\Gamma^{-1}(\hat{\theta})$, which means our algorithm applies by solving a parametric maximum likelihood problem (using a SEM program such as LISREL, EQS, AMOS, or CALIS) in each step. We then transform the variables, and reapply maximum likelihood. Exploratory and confirmatory factor analysis are special cases of this general setup.

3.7. **Bilinearizability.** There is more that can be said about the relationship between MCA and the correlation aspects of nonlinear PCA. Most of the theory is taken from De Leeuw [1982]; Bekker and De Leeuw [1988]; De Leeuw [1988]; De Leeuw et al. [1999].

Let us start by looking at the condition of *bilinearizability* of regressions. This means that we can find transformation of the variables (in our class of admissible transformations) such that all bivariate regressions are exactly linear. In the case of m categorical variables with Burt table C this means that the system of bilinearizability equations

$$C_{j\ell}y_\ell = r_{j\ell}C_{jj}y_j \tag{8}$$

has a solution, normalized by $y_j'C_{jj}y_j = 1$ for all j. The corresponding induced correlation matrix $R(y)$ has m eigenvalues λ_s and m corresponding normalized eigenvectors α_s. We can now define the m vectors $a_{js} = \alpha_{js}y_j$, and we find $\sum_{l=1}^m C_{j\ell}a_{\ell s} = \sum_{l=1}^m C_{j\ell}y_\ell\alpha_{\ell s} = C_{jj}y_j\sum_{l=1}^m r_{j\ell}\alpha_{\ell s} = \lambda_s\alpha_{js}C_{jj}y_j = \lambda_s C_{jj}a_{js}$. In other words, for each s the vector a_s defines a solution to the MCA problem, with eigenvalue λ_s, and each of these m solutions induces the same correlation matrix.

Bilinearizability has some other important consequences. A system of transformations that linearizes all regressions solves the stationary equations for any aspect of the correlation matrix. Thus in a multivariate data matrix with bilinearizability, it does matter which aspect we choose, because they will all give the same transformations. Another important consequence of bilinearizability is that the standard error of the correlation coefficients computed by maximizing an aspect have the same standard errors as the correlation coefficients computed from known scores. This means we can apply the asymptotically distribution free methods of SEM programs to optimized correlation matrices, and they will still compute the correct tests and standard errors if the data are bilinearizable (or a sample from a bilinearizable distribution).

3.8. **Complete Bilinearizability.** It may be the case that there is a second set of transformations $\overline{y}_j$ that satisfy the equations (8). Again, such set generates m additional MCA solutions, all inducing the same correlation matrix. Moreover $y_j'C_{jj}\overline{y}_j = 0$ for all j so the second set is orthogonal to the first for each variable separately. And there may even be more sets. If bilinearizability continues to apply, we can build up all MCA solutions from the solutions to (8) and the eigenvectors of the induced correlation matrices. Another way of thinking about this is that we solve $\binom{m}{2}$ simple CA problems for each of the subtables of the Burt matrix. Equation (8) then says that if we have complete bilinearizability we can patch these CA solutions together to form the MCA solution.

More precisely, suppose C is a Burt matrix and D is its diagonal. We have complete bilinearizability if there are matrices K_j such that $K_j'C_{jj}K_j = I$ for each j and $K_j'C_{j\ell}K_\ell$ is diagonal for each j and ℓ. Remember that the direct sum of matrices stacks the matrices in the diagonal submatrices of a large matrix, which has all its non-diagonal submatrices equal to zero. If K is the direct sum of the K_j, then $K'DK = I$ while $E = K'CK$ has the same structure as the Burt matrix, but all submatrices $E_{j\ell}$ are now diagonal. This means there is a permutation matrix P such that $P'K'CKP$ is the direct sum of correlation matrices. The first correlation matrix contains all $(1,1)$ elements of the $E_{j\ell}$, the second correlation matrix contains all $(2,2)$ elements, and so on. By making L the direct sum of the matrices of eigenvectors of these correlation matrices, we see that $L'P'K'CKPL$ is diagonal, while $L'P'K'DKPL = I$. Thus the matrix KPL contains all the MCA solutions and gives a complete eigen decomposition of the Burt matrix.

This may be somewhat abstract, so let's give a very important example. Suppose we perform an MCA of a standard multivariate normal, with correlation matrix Γ. Because all bivariate regressions are linear, linear transformations of the variables are a bilinearizability system, with correlation matrix Γ. But the quadratic Hermite-Chebyshev polynomials are another bilinearizability system, with correlation matrix $\Gamma^{(2)}$, the squares of the correlation coefficients, and so on. Thus we see that applying MCA to a multivariate normal will give m solutions consisting of polynomials of degree d, where the eigenvalues are those of $\Gamma^{(d)}$, for all $d = 1, 2, \cdots$.

In standard MCA we usually order the eigenvalues, and look at the largest ones, often the two largest ones. The largest eigenvalue for the multivariate normal is always the largest eigenvalue of Γ, but the second largest eigenvalue can be either the second largest eigenvalue of Γ or the largest eigenvalue of $\Gamma^{(2)}$. If the second largest eigenvalue in the MCA is the largest eigenvalue of $\Gamma^{(2)}$, then for each variable the first transformation will be linear and the second will be quadratic, which means we will find horseshoes [Van Rijckevorsel, 1987] in our scatterplots. There is an example in Gifi [1990, page 382–384], where two-dimensional MCA takes both its transformations from Γ, which means it finds the usual nonlinear PCA solution.

Our analysis shows clearly what the relationships are between MCA and nonlinear PCA. In PCA we find a single set of transformations, and a corresponding induced correlation matrix which is optimal in terms of an aspect. In MCA we find multiple transformations, each with its own corresponding induced correlation matrix. Only in the case of complete bilinearizability (such as obtains in the multivariate normal) can we relate the two solutions because they basically the same solution. MCA, however, presents the solution in a redundant and confusing manner. This gives a more precise meaning to the warning by Guttman [1941] that the additional dimensions beyond the first one in MCA should be interpreted with caution.

3.9. **Examples of Nonlinear PCA.** The first example of a nonlinear PCA is from Roskam [1968, p. 152]. The Department of Psychology at the University of Nijmegen has, or had, 9 different areas of research and teaching. Each of the 39 psychologists working in the department ranked all 9 areas in order of relevance for their work. The areas are given in Table 3, and the data in Table 2. Think of this example as 9 observations on 39 variables. We first perform a linear PCA on the rank numbers, which is sometimes known as Tucker's Preference Analysis [Tucker, 1960]. The first two eigenvalues of R/m are 0.374 and 0.176, which means the first two principal components capture 55% of the variation in the rank numbers. We now optimize the sum of the first two eigenvalues over all monotone transformations of the 39 variables. The eigenvalues increase to 0.468 and 0.297, and thus the two principal components capture 76.6% of the transformed rank numbers. For completeness we also note that maximizing the largest eigenvalue gives 0.492 and maximizing the sum of the first three eigenvalues bring the percentage of captured variance up to 87.2%.

If we look at the plots of eigenvectors (scaled by the square roots of the eigenvalues) for the two-dimensional solution in Figures 3 and 4 we see that the linear PCA produces groupings which are somewhat counter-intuitive, mostly because there is so much variation left in the third and higher dimensions. The grouping in the nonlinear PCA is much clearer. Psychologists in the same area are generally close together, and there is a relatively clear distinction between qualitative and quantitative areas.

A second data set are the GALO data, taken from Peschar [1975]. The objects (individuals) are 1290 school children in the sixth grade of elementary school in the city of Groningen (Netherlands) in 1959. The four variables are Gender, IQ, Advice, and SES. IQ has been categorized into 9 ordered categories. In “Advice” the sixth-grade teachers categorizes the children into seven possible forms of secondary education. In “SES” the parent’s profession is categorized in six categories.

We use these data to maximize a large number of different aspects of the correlation matrix. All variables are categorical, and no monotonicity or smoothness constraints are imposed. Results are in Table 4 in which we give the four eigenvalues of the correlation matrix, and in the final column the induced correlation between IQ and Advice. The largest possible eigenvalue is 2.157 and the smallest possible one is 0.196. The regression type solutions, seeking singularities, tend to give a small value for the smallest eigenvalue. In general the pattern of eigenvalues is very similar for the different aspects, suggesting approximate bilinearizability. We give the transformations for the aspect that maximizes the largest eigenvalue in Figure 5.

We can also use this example to illustrate the difference between MCA and nonlinear PCA. Figure 6 has the two principal components from an MCA solution. The components come from different correlation matrices, one corresponding with linear transformations and one corresponding with quadratic ones. Thus the component scores form a horseshoe. The nonlinear PCA solution for the same data is shown in Figure 7. Both components come from the correlation matrix induced by the transformations in Figure 5. We see a completely different plot, without horseshoe, in which the discrete parallel strips of points come about because the dominant variables IQ and Advice only have a small finite number of values.

4. Logistic Nonlinear PCA

In the remainder of this chapter we discuss an entirely different way to define and fit nonlinear PCA. It does not use least squares, at least not to define the loss function. The notion of correlation between variables is not used in this approach, because we do not construct numerical quantified or transformed variables.

Suppose the data are categorical, as in MCA, and coded as indicator matrices. The indicator matrix Z_j for variable j has n rows and k_j columns. Remember that $\sum_{\ell=1}^{k_j} z_{ij\ell} = 1$ for all i and j. As in MCA we represent both the n objects and the k_j categories of variable j as points a_i and b_{jl} in low-dimensional Euclidean space.

We measure loss by using the *deviance*, or the negative log-likelihood,

$$\Delta(A, B) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell=1}^{k_j} z_{ij\ell} \log \pi_{ij\ell}(A, B),$$

where

$$\pi_{ijl}(A, B) = \frac{\exp(\eta(x_i, y_{j\ell}))}{\sum_{\nu=1}^{k_j} \exp(\eta(x_i, y_{j\nu})}.$$

For the time being we do not specify the *combination rule* η and we develop our results for a prefectly general combination rule. But to make matters less abstract, we can think of the inner product $\eta(a_i, b_{j\ell}) = a_i'b_{j\ell}$, or the negative distance $\eta(a_i, b_{j\ell}) = -\|a_i - b_{j\ell}\|$.

4.1. **Algorithm.** To minimize the loss function we use quadratic majorization [Böhning and Lindsay, 1988; De Leeuw, in press]. We need the first and the second derivatives of the deviance with respect to the $\eta_{ij\ell}$. Here $\eta_{ij\ell} = \eta_{ij\ell}(A, B)$ is used interchangeably with $\eta(a_i, b_{j\ell})$. Simple computation gives

$$\frac{\partial\Delta}{\partial\eta_{ij\ell}} = \pi_{ij\ell} - z_{ij\ell}$$

and

$$\frac{\partial^2\Delta}{\partial\eta_{ij\ell}\partial\eta_{ij\nu}} = \pi_{ij\ell}\delta^{\ell\nu} - \pi_{ij\ell}\pi_{ij\nu}.$$

It follows that for each (i, j) the matrix of second derivatives, which is of order k_j, has a largest eigenvalue less than or equal to $\frac{1}{2}$. Thus, from a Taylor expansion at $\eta(\tilde{A}, \tilde{B})$,

$$\begin{aligned}(9)\quad \Delta(A,B) \leq \Delta(\tilde{A},\tilde{B}) + \\ + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell=1}^{k_j}(\pi_{ij\ell}(\tilde{A},\tilde{B}) - z_{ij\ell})(\eta(a_i,b_{j\ell}) - \eta(\tilde{a}_i,\tilde{b}_{j\ell})) + \\ + \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell=1}^{k_j}(\eta(a_i,b_{j\ell}) - \eta(\tilde{a}_i,\tilde{b}_{j\ell})^2.\end{aligned}$$

By general majorization theory explained in Appendix A, it suffices to minimize the right hand side of (9) in each iteration. This is equivalent, by completing the square, to minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell=1}^{k_j}[\eta_{ij\ell}(A,B) - \tau_{ij\ell}(\tilde{A},\tilde{B})]^2 \tag{10a}$$

where the current *target* is defined by

$$\tau_{ij\ell}(\tilde{A},\tilde{B}) = \eta_{ij\ell}(\tilde{A},\tilde{B}) - 2[z_{ij\ell} - \pi_{ijl}(\tilde{A},\tilde{B})]. \tag{10b}$$

Thus we can solve the logistic nonlinear PCA problem by using iterative least squares. If we know how to fit $\eta_{ij\ell}(A, B)$ to a matrix by least squares, then we can also fit it logistically. In iteration k we compute the current target $\tau(A^{(k)}, B^{(k)})$ by (10b), and then we minimize (or at least improve) the least squares loss function (10a) to find $A^{(k+1)}$ and $B^{(k+1)}$.

This implies immediately that for the inner product or bilinear composition rule η we can use iterated singular value decomposition, while for the negative distance rule we can use iterated least squares multidimensional unfolding. Observe, however, that the target values in (10b) may very well be negative, which can be a problem for some multidimensional scaling algorithms. In De Leeuw [in press] it is shown how the approach can easily be extended to deal with probit, instead of logit, loss functions.

4.2. **Perfect Fit.** In general it will not be possible to find a perfect solution with zero deviance. We discuss under what conditions such a solution does exist. Consider the system of strict inequalities

$$\eta(a_i, b_{j\ell}) > \eta(a_i, b_{j\nu}) \tag{11}$$

for all (i, j, ℓ, ν) for which $z_{ij\ell} = 1$. In other words, for all i and j the largest of the $\eta(a_i, b_{j\nu})$ must be the the one corresponding to category ℓ for which $z_{ij\ell} = 1$.

Suppose system (11) has a solution $(\hat{A}, \hat{B})$, and suppose our combination rule η is homogeneous in the sense that $\eta(\lambda a_i, \lambda b_{j\ell}) = \lambda^r \eta(a_i, b_{j\ell})$ for some positive power r. Then by letting λ go to infinity we see that $\pi_{ij\ell}(\lambda\hat{A}, \lambda\hat{B})$ goes to one for all $z_{ij\ell}$ equal to one, and thus $\Delta(\lambda\hat{A}, \lambda\hat{B})$ goes to zero. We have a perfect solution, but with all points at infinity. While generally (11) will not be solvable, we can perhaps expect some points to move to infinity in the actual solutions we compute.

4.3. **Geometry of Combination Rules.** In our further analysis we concentrate on the particular combination rule using negative distance $\eta(a_i, b_{j\ell}) = -\|a_i - b_{jl}\|$. System (11) says that we want to map objects and categories into low-dimensional space in such a way that each object is closest to the category point that it falls in.

This can be illustrated nicely by using the notion of a *Voronoi diagram* [Okabe et al., 2000]. In a Voronoi diagram (for a finite number, say p, points) space is partitioned into p regions, one for each point. The cell containing the point s is the locus of all points in space that are closer to point s than to the other $p-1$ points. Voronoi cells can be bounded and unbounded, and in the Euclidean case they are polyhedral and bounded by pieces of various perpendicular bisectors. Using the $b_{j\ell}$ we can make a Voronoi diagram for each variable. Our logistic PCA, for this particular combination rule, says that each object point a_i should be in the correct Voronoi cell for each variable.

This type of representation is closely related to representation of categorical data in Guttman's MSA-I, discussed by Lingoes [1968]. It should also be emphasized that if the data are binary, then the Voronoi diagram for a variable just consist of a single hyperplane partitioning space into two regions. System (11) now says that the "yes" responses should be on one side of the hyperplane and the "no" responses should be on the other side. This is a classical version of nonlinear PCA, dating back to at least Coombs and Kao [1955], and used extensively in political science [Clinton et al., 2004].

As in Gifi [1990], we can construct variations on the basic technique by imposing constraints on the $y_{j\ell}$. If we constrain them, for example, to be on a straight line through the origin by setting $y_{j\ell s} = z_{j\ell}a_{js}$ then the bisecting hyperplanes will all be perpendicular to this line and for each variable the space will be divided into parallel strips or bands. Objects should be in the correct strip. This is the form of nonlinear PCA we already discussed in the least squares context, except that loss is measured on probabilities instead of correlations.

4.4. **Examples.** There is a example analyzing 20 "aye" and "nay" votes in the US Senate in De Leeuw [in press]. The 20 issues voted on are, of course, binary variables, which means the two Voronoi cells for each variable are half-spaces. In this section we will analyze a slightly more complicated data set, using the GALO data.

The four GALO variables have a total of 24 categories, and there are 1290 individuals. Thus the metric unfolding analysis in each majorization step must fit 30960 distances, using targets τ that can easily be negative. If we make all distances zero, which can obviously be done by collapsing all points, then the deviance becomes $1290*(\log 2+\log 9+\log 6+\log 7) = 8550$. This is, in a sense, the worst possible solution, in which all probabilities are equal.

We have written some software to optimize our loss functions. It has not been tested extensively, but so far it seems to provide convergence. It starts with the MCA solution. Remember that in MCA [Michailidis and De Leeuw, 1998] we want the a_i to be close in the least squares sense to the category centroids $b_{j\ell}$, or in the graph drawing interpretation [De Leeuw and Michailidis, 1999] we want the category stars to be small. It seems reasonable to suppose that small stars will correspond with points in their corresponding Voronoi cell. The MCA solution starts with a negative likelihood of 8490 and improves this to 8315.

In Figure 8 we draw the Voronoi cells for IQ (observe they are all open). The category points for IQ are almost on a circle (the horseshoe closes somewhat), starting first the lowest IQ category at the bottom center, and then proceeding clockwise to the higher categories. We present this solution somewhat tentatively, because both theory and algorithm are new and will require much research and refinement. It is clear, however, that at least in principle the basic theory and algorithms of Gifi [1990], which cover MCA, nonlinear PCA, and various forms of nonlinear canonical analysis, can be extended to logit and probit loss functions that optimize aspects of probabilities instead of aspects of correlation coefficients.

Appendix A. Majorization

In a majorization algorithm the goal is to minimize a function $\phi(\theta)$ over $\theta \in \Theta$, with $\Theta \subseteq \mathbb{R}^p$. Majorization requires us to construct a function $\psi(\theta, \xi)$ defined on $\Theta \times \Theta$ that satisfies

$$\phi(\theta) \leq \psi(\theta, \xi) \text{ for all } \theta, \xi \in \Theta, \tag{12a}$$

$$\phi(\theta) = \psi(\theta, \theta) \text{ for all } \theta \in \Theta. \tag{12b}$$

Thus, for a fixed ξ, $\psi(\bullet, \xi)$ is above ϕ, and it touches ϕ at the point $(\xi, \phi(\xi))$. We then say that $\psi(\theta, \xi)$ *majorizes* $\phi(\theta)$ at ξ.

There are two key theorems associated with these definitions.

Theorem A.1. *If ϕ attains its minimum on Θ at $\hat{\theta}$, then $\psi(\bullet, \hat{\theta})$ also attains its minimum on Θ at $\hat{\theta}$.*

Proof. Suppose $\psi(\tilde{\theta}, \hat{\theta}) < \psi(\hat{\theta}, \hat{\theta})$ for some $\tilde{\theta} \in \Theta$. Then, by (12a) and (12b), $\phi(\tilde{\theta}) \leq \psi(\tilde{\theta}, \hat{\theta}) < \psi(\hat{\theta}, \hat{\theta}) = \phi(\hat{\theta})$, which contradicts the definition of $\hat{\theta}$ as the minimizer of ϕ on Θ. □

Theorem A.2. *If $\tilde{\theta} \in \Theta$ and $\hat{\theta}$ minimizes $\psi(\bullet, \tilde{\theta})$ over Θ, then $\phi(\hat{\theta}) \leq \phi(\tilde{\theta})$.*

Proof. By (12a) we have $\phi(\hat{\theta}) \leq \psi(\hat{\theta}, \tilde{\theta})$. By the definition of $\hat{\theta}$ we have $\psi(\hat{\theta}, \tilde{\theta}) \leq \psi(\tilde{\theta}, \tilde{\theta})$. And by (12b) we have $\psi(\tilde{\theta}, \tilde{\theta}) = \phi(\tilde{\theta})$. Combining these three results we get the result. □

These two results suggest the following iterative algorithm for minimizing $\phi(\theta)$. Suppose we are at step k.

Step 1: Given a value $\theta^{(k)}$ construct a majorizing function $\psi(\theta, \theta^{(k)})$.

Step 2: Set $\theta^{(k+1)} = \mathop{\mathbf{argmin}}\limits_{\theta\in\Theta} \psi(\theta, \theta^{(k)})$.

Step 3: If $|\phi(\theta^{(k+1)}) - \phi(\theta^{(k)}| < \epsilon$ for some predetermined $\epsilon > 0$ stop; else go to Step 1.

In order for this algorithm to be of practical use, the majorizing function ψ needs to be easy to minimize, otherwise nothing substantial is gained by following this route.

We demonstrate next how the idea behind majorization works with a simple artificial example, chosen for its simplicity. Consider $\phi(\theta) = \theta^4 - 10\theta^2,\ \theta \in \mathbb{R}$. Because $\theta^2 \geq \xi^2 + 2\xi(\theta-\xi) = 2\xi\theta - \xi^2$ we see that $\psi(\theta, \xi) = \theta^4 - 20\xi\theta + 10\xi^2$ is a suitable majorization function. The majorization algorithm is $\theta^+ = \sqrt[3]{5\xi}$.

The algorithm is illustrated in Figure 9. We start with $\theta^{(0)} = 5$. Then $\psi(\theta, 5)$ is the dashed function. It is minimized at $\theta^{(1)} \approx 2.924$, where $\psi(\theta^{(1)}, 5) \approx 30.70$, and $\phi(\theta^{(1)}) \approx -12.56$. We then majorize by using the dotted function $\psi(\theta, \theta^{(1)})$, which has its minimum at about 2.44, equal to about -21.79. The corresponding value of ϕ at this point is about -24.1. Thus we are rapidly getting close to the local minimum at $\sqrt{5}$, with value 25. The linear convergence rate at this point is $\frac{1}{3}$.

We briefly address next some convergence issues (for a general discussion see the book by Zangwill [1969]). If ϕ is bounded above below on Θ, then the algorithm generates a bounded decreasing sequence of function values $\phi(\theta^{(k)})$, which thus converges to $\phi(\theta^{\infty})$. For example, continuity of ϕ and compactness of Θ would suffice for establishing the result. Moreover with some additional mild continuity considerations [De Leeuw, 1994] we get that $||\theta^{(k)} - \theta^{(k+1)}|| \rightarrow 0$, which in turn implies, because of a result by Ostrowski [1966], that either θ converges to a single point or that there is a continuum of limit points (all with the same function value). Hence, majorization algorithms, for all practical purposes, find local optima.

We make two final points about this class of algorithms. It is not necessary to actually minimize the majorization function in each step, it suffices to decrease it in a systematic way, for instance by taking a single step of a convergent "inner" iterative algorithm. And the rate of convergence of majorization algorithms is generally linear, in fact it is equal to the size of the second derivatives of the majorization function compared to the size of the second derivatives of the original function [De Leeuw and Michailides, (in preparation)].

Appendix B. Software

There are quite a number of software options for performing the various forms of nonlinear PCA explained in this chapter. PRINQUAL in SAS [1992] can optimize sums of the largest eigenvalues, as well as the sum of correlations and the determinant aspect. Categories [Meulman and Heiser, 1999] has CATPCA, which optimizes the classical eigenvalue criteria. In the R contributed packages we find the function homals from the homals package, which can perform nonlinear PCA for categorical variables with or without ordinal constraints using the many options inherent in the Gifi system. There are also programs for nonlinear PCA in the Guttman-Lingoes programs [Lingoes, 1973].

We are currently preparing the gifi package for R, which which will have functions to optimize arbitrary aspects of the correlation matrix and to the the nonlinear PCA of rank orders we applied in the Roskam example. It will also have the PREHOM program discussed by Bekker and De Leeuw [1988], which finds complete bilinearizable systems of scores of they exist, and the LINEALS program discussed by De Leeuw [1988]. The R code is available from the author.

Code for the logistic (and probit) versions of PCA in R is also available, in preliminary form, and will eventually be wrapped into a separate package.

The binary version has been tested quite extensively [Lewis and De Leeuw, 2004], and can be compared with similar programs for IRT analysis written mostly by educational statisticians and for roll-call analysis written mostly by political scientists.

References

P. Bekker and J. De Leeuw. Relation between Variants of Nonlinear Principal Component Analysis. In J.L.A. Van Rijckevorsel and J. De Leeuw, editors, *Component and Correspondence Analysis*. Wiley, Chichester, England, 1988.

D. Böhning and B.G. Lindsay. Monotonicity of Quadratic-approximation Algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4): 641–663, 1988.

J. Clinton, S. Jackman, and D. Rivers. The Statistical Analysis of Roll Call Data. *American Political Science Review*, 98:355–370, 2004.

C.H. Coombs and R.C. Kao. Nonmetric Factor Analysis. Engineering Research Bulletin 38, Engineering Research Institute, University of Michigan, Ann Arbor, 1955.

J. De Leeuw. Multivariate Analysis with Optimal Scaling. In S. Das Gupta and J. Sethuraman, editors, *Progress in Multivariate Analysis*, Calcutta, India, 1990. Indian Statistical Institute.

J. De Leeuw. Principal Component Analysis of Binary Data by Iterated Singular Value Decomposition. *Computational Statistics and Data Analysis*, in press.

J. De Leeuw. Block Relaxation Methods in Statistics. In H.H. Bock, W. Lenski, and M.M. Richter, editors, *Information Systems and Data*

Analysis, Berlin, 1994. Springer Verlag.

J. De Leeuw. Monotonic Regression. In Brian S. Everitt and David C. Howell, editors, *Encycxlopedia of Statistics in Behavioral Science*, volume 3, pages 1260–1261. Wiley, 2005.

J. De Leeuw. Multivariate Analysis with Linearizable Regressions. *Psychometrika*, 53:437–454, 1988.

J. De Leeuw. Nonlinear Principal Component Analysis. In H. Caussinus Et Al., editor, *COMPSTAT 1982*, Vienna, Austria, 1982. Physika Verlag.

J. De Leeuw and J.J. Meulman. Principal Component Analysis and Restricted Multidimensional Scaling. In W. Gaul and M. Schader, editors, *Classification as a Tool of Research*, Amsterdam, London, New York, Tokyo, 1986. North-Holland.

J. De Leeuw and G. Michailides. *Block Relaxation and Majorization Algorithms in Statistics*. Springer, (in preparation).

J. De Leeuw and G. Michailidis. Graph Layout Techniques and Multidimensional Data Analysis. In T. Bruss and L. LeCam, editors, *Festschrift for Thomas S. Ferguson*. Institute of Mathematical Statistics, 1999.

J. De Leeuw, G. Michailidis, and D. Y. Wang. Correspondence Analysis Techniques. In S. Ghosh, editor, *Multivariate Analysis, Design of Experiments, and Survey Sampling*. Marcel Dekker, 1999.

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

J.C. Gower and D.J. Hand. *Biplots*. Number 54 in Monographs on Statistics and Applied Probability. Chapman and Hall, 1996.

L. Guttman. The Quantification of a Class of Attributes: A Theory and Method of Scale Construction. In P. Horst, editor, *The Prediction of Personal Adjustment*. Social Science Research Council, New York, New York, 1941.

L. Guttman. Image Theory for the Structure of Quantitative Variables. *Psychometrika*, 18:277–296, 1953.

A.S. Householder and G. Young. Matrix Approximation and Latent Roots. *American Mathematical Monthly*, 45:165–171, 1938.

I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer Verlag, Second edition, 2002.

J.B. Kruskal and R.N. Shepard. A Nonmetric Variety of Linear Factor Analysis. *Psychometrika*, 39:123–157, 1974.

J. Lewis and J. De Leeuw. A General Method for Fitting Spatial Models of Politics. Technical report, UCLA Department of Statistics, 2004.

J.C. Lingoes. *The Guttman-Lingoes Nonmetric Program Series*. Mathesis Press, 1973.

J.C. Lingoes. The Multivariate Analysis of Qualitative Data. *Multivariate Behavioral Research*, 3:61–94, 1968.

J.C. Lingoes and L. Guttman. Nonmetric Factor Analysis: a Rank Reducing Alternative to Linear Factor Analysis. *Multivariate Behavioral Research*, 2:485–505, 1967.

J. J. Meulman and W.J. Heiser. *SPSS Categories 10.0.* SPSS Inc., Chicago, Illinois, 1999.

G. Michailidis and J. De Leeuw. The Gifi system for Descriptive Multivariate Analysis. *Statistical Science*, 13:307–336, 1998.

A. Okabe, B. Boots, K. Sugihara, and S.N. Chiu. *Spatial Tessellations.* Wiley, second edition, 2000.

A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, N.Y., 1966.

K. Pearson. On Lines and Planes of Closest Fit to Systems of Point is Space. *Philosophical Magazine (6)*, 23:559–572, 1901.

J. L. Peschar. *School, Milieu, Beroep*. Tjeek Willink, Groningen, The Netherlands, 1975.

E.E.CH.I. Roskam. *Metric Analysis of Ordinal Data in Psychology*. PhD thesis, University of Leiden, 1968.

SAS. SAS/STAT Software: Changes and Enhancements. Technical Report P-229, SAS Institute Inc., Cary, North Carolina, 1992.

L.L. Thurstone. *Multiple Factor Analysis*. University of Chicago Press, Chicago, Illinois, 1947.

L.R. Tucker. Intra-individual and Inter-individual Multidimensionality. In H. Gulliksen and S. Messick, editors, *Psychological Scaling: Theory and Applications*. Wiley, 1960.

J.L.A. Van Rijckevorsel. *The Application of Fuzzy Coding and Horseshoes in Multiple Correspondence Analysis*. PhD thesis, University of Leiden, The Netherlands, 1987. Also published in 1987 by DSWO-Press, Leiden, The Netherlands.

F.W. Young, Y. Takane, and J. De Leeuw. The Principal Components of Mixed Measurement Level Multivariate Data: an Alternating Least Squares Method with Optimal Scaling Features. *Psychometrika*, 45:279–281, 1978.

W. I. Zangwill. *Nonlinear Programming: a Unified Approach*. Prentice-Hall, Englewood-Cliffs, N.J., 1969.

base	height	diag	area	circf	b/h	h/b
1	1	1.41	1	4	1.00	1.00
2	2	2.82	4	8	1.00	1.00
3	3	4.24	9	12	1.00	1.00
4	4	5.66	16	16	1.00	1.00
5	5	7.07	25	20	1.00	1.00
6	6	8.49	36	24	1.00	1.00
7	7	9.90	49	28	1.00	1.00
8	8	11.31	64	32	1.00	1.00
9	9	12.73	81	36	1.00	1.00
10	10	14.14	100	40	1.00	1.00
11	10	14.87	110	42	1.10	0.91
12	9	15.00	108	42	1.33	0.75
13	8	15.26	104	42	1.63	0.62
14	7	15.65	98	42	2.00	0.50
15	6	16.16	90	42	2.50	0.40
16	5	16.76	80	42	3.20	0.31
17	4	17.46	68	42	4.25	0.23
18	3	18.24	54	42	6.00	0.17
19	2	19.10	38	42	9.50	0.11
20	1	20.02	20	42	20.00	0.05

TABLE 1. Rectangles

	SOC	EDU	CLI	MAT	EXP	CUL	IND	TST	PHY
1	1	5	7	3	2	4	6	9	8
2	1	3	2	7	6	4	5	8	9
3	1	6	5	3	8	2	4	7	9
4	1	5	4	7	6	2	3	8	9
5	7	1	4	3	6	8	2	9	5
6	6	1	2	5	3	7	8	4	9
7	2	1	4	5	3	8	6	7	9
8	4	1	2	8	3	5	9	6	7
9	4	1	3	5	7	6	8	2	9
10	3	1	2	4	6	8	9	7	5
11	4	1	8	3	7	6	2	5	9
12	3	2	1	5	6	8	7	4	9
13	2	9	1	6	8	3	4	5	7
14	2	7	1	4	3	9	5	6	8
15	7	2	1	3	5	8	9	4	6
16	5	7	8	1	3	9	4	2	6
17	5	9	8	1	2	7	6	3	4
18	9	6	5	1	3	7	8	2	4
19	9	6	7	2	1	8	3	4	5
20	8	3	7	2	1	9	4	5	6
21	7	2	8	5	1	9	6	4	3
22	8	7	6	3	1	9	2	5	4
23	8	6	5	2	1	9	4	7	3
24	8	7	5	2	1	9	6	4	3
25	7	3	6	2	1	9	8	4	5
26	4	7	9	5	1	8	2	3	6
27	5	6	8	2	1	9	4	7	3
28	1	8	9	2	3	7	6	4	5
29	2	5	6	4	8	1	7	3	9
30	2	5	4	3	6	1	8	7	9
31	5	3	2	9	4	1	6	7	8
32	4	5	6	2	8	7	1	3	9
33	5	7	9	3	2	8	1	4	6
34	6	3	7	2	8	5	1	4	9
35	8	5	7	4	2	9	1	3	6
36	2	6	5	4	3	7	1	8	9
37	5	8	9	2	3	7	1	4	6
38	8	7	3	4	2	9	5	6	1
39	5	6	7	2	4	9	8	3	1

TABLE 2. Roskam Psychology Subdiscipline Data

Area	Plot Code
Social Psychology	SOC
Educational and Developmental Psychology	EDU
Clinical Psychology	CLI
Mathematical Psychology and Psychological Statistics	MAT
Experimental Psychology	EXP
Cultural Psychology and Psychology of Religion	CUL
Industrial Psychology	IND
Test Construction and Validation	TST
Physiological and Animal Psychology	PHY

TABLE 3. Nine Psychology Areas

Aspect	λ_1	λ_2	λ_3	λ_4	r_{23}
Sum of Correlations	2.147	0.987	0.637	0.229	0.767
Sum of Squared Correlations	2.149	0.998	0.648	0.204	0.791
Sum of Cubed Correlations	2.139	0.934	0.730	0.198	0.796
Largest Eigenvalue	2.157	0.950	0.682	0.211	0.784
Sum of Two Largest Eigenvalues	1.926	1.340	0.535	0.198	0.795
Sum of Three Largest Eigenvalues	1.991	1.124	0.688	0.196	0.796
Squared Multiple Correlation with Advice	2.056	1.043	0.703	0.196	0.796
Sum of Squared Multiple Correlations	1.961	1.302	0.538	0.199	0.795
Determinant	2.030	1.220	0.551	0.199	0.796

Table 4. GALO Example with Aspects

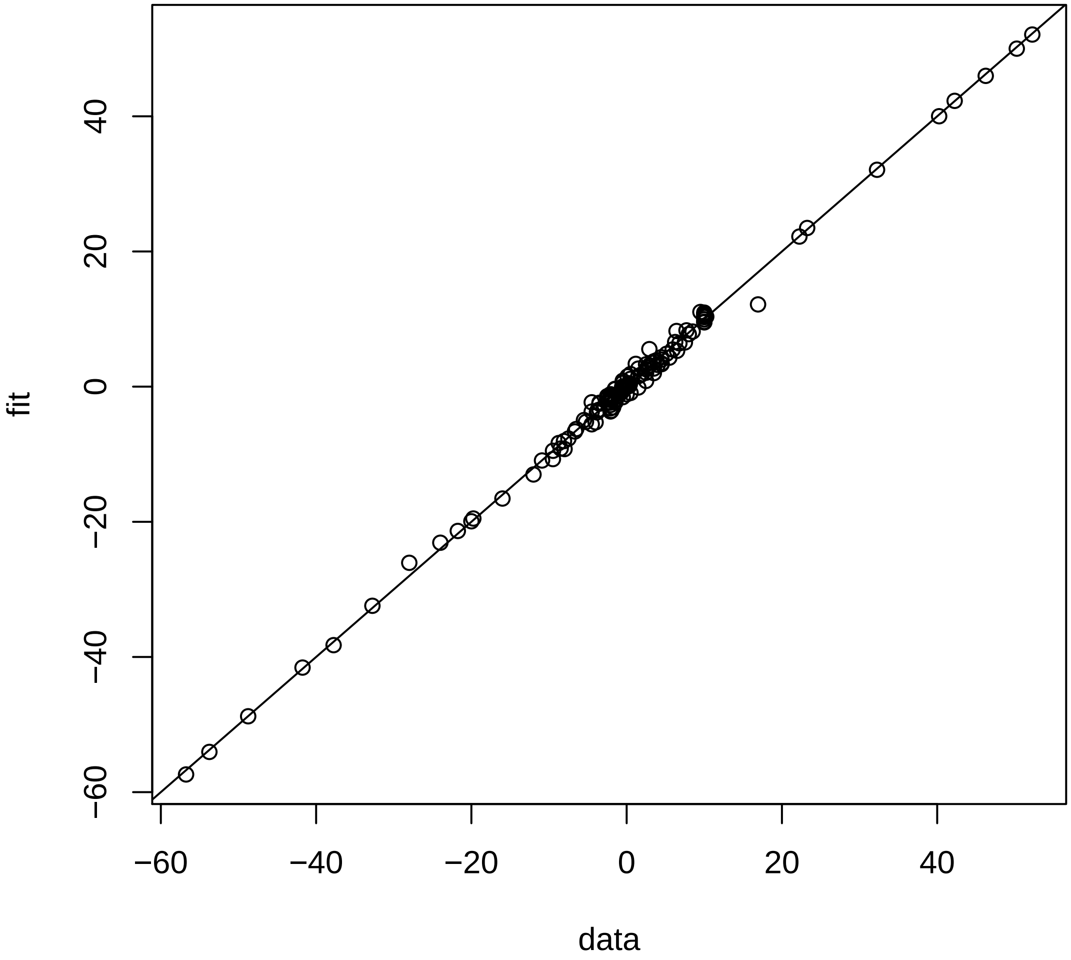

FIGURE 1. PCA Fit for Rectangles

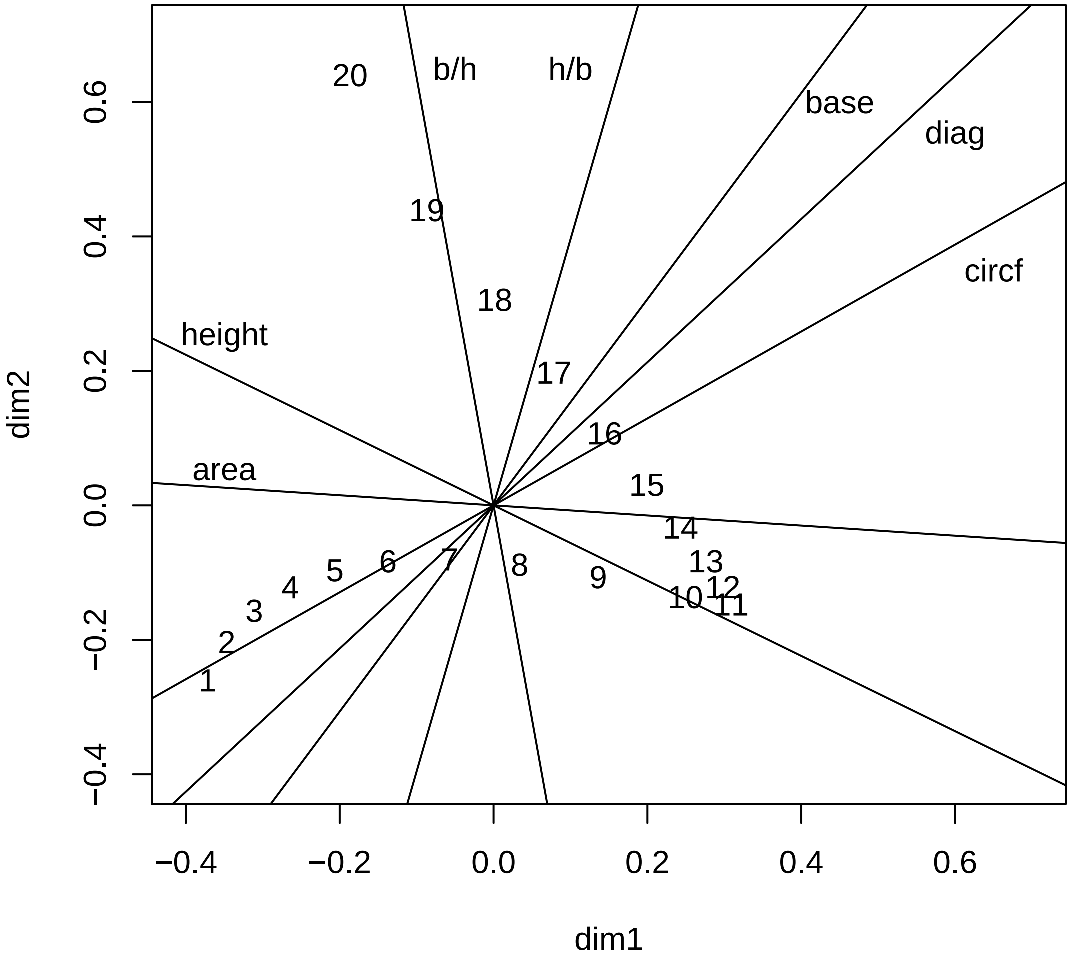

FIGURE 2. PCA Solution for Rectangles

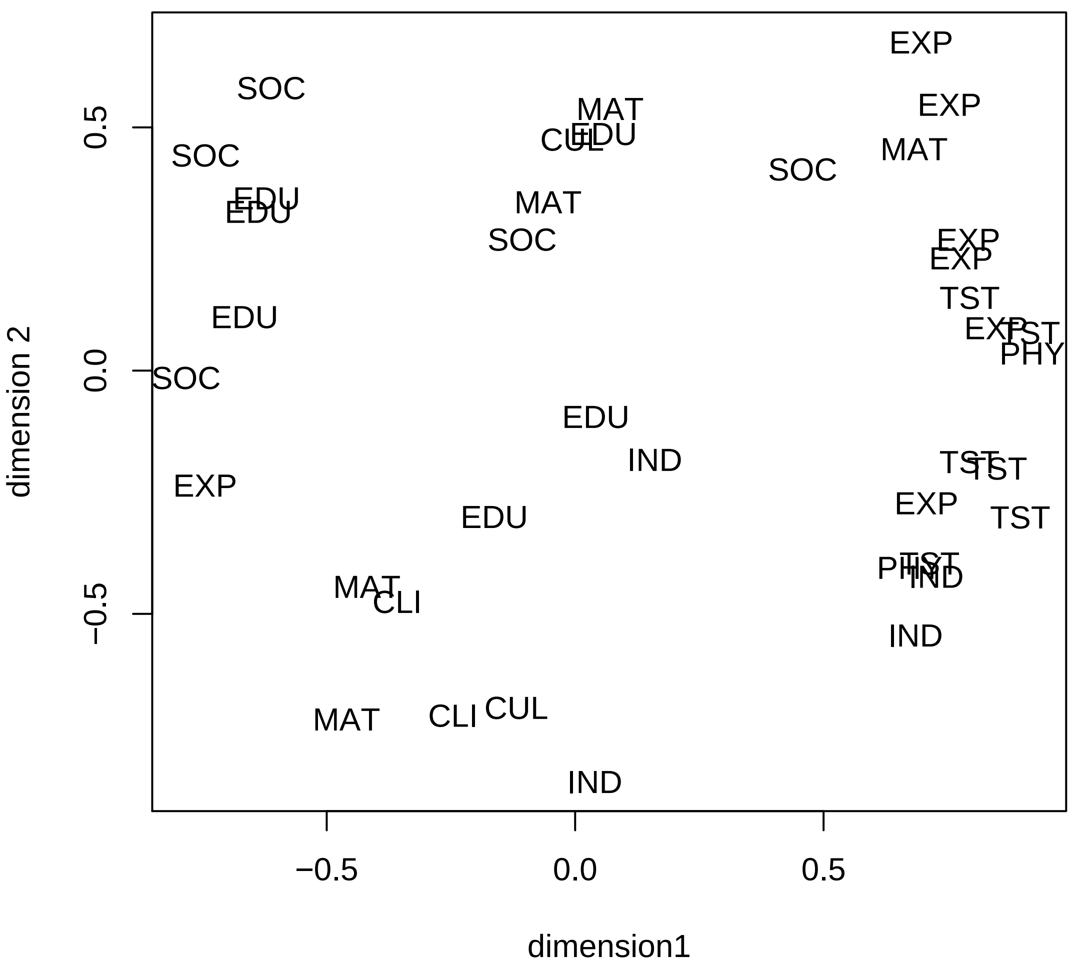

Figure 3. Roskam Data: Linear PCA

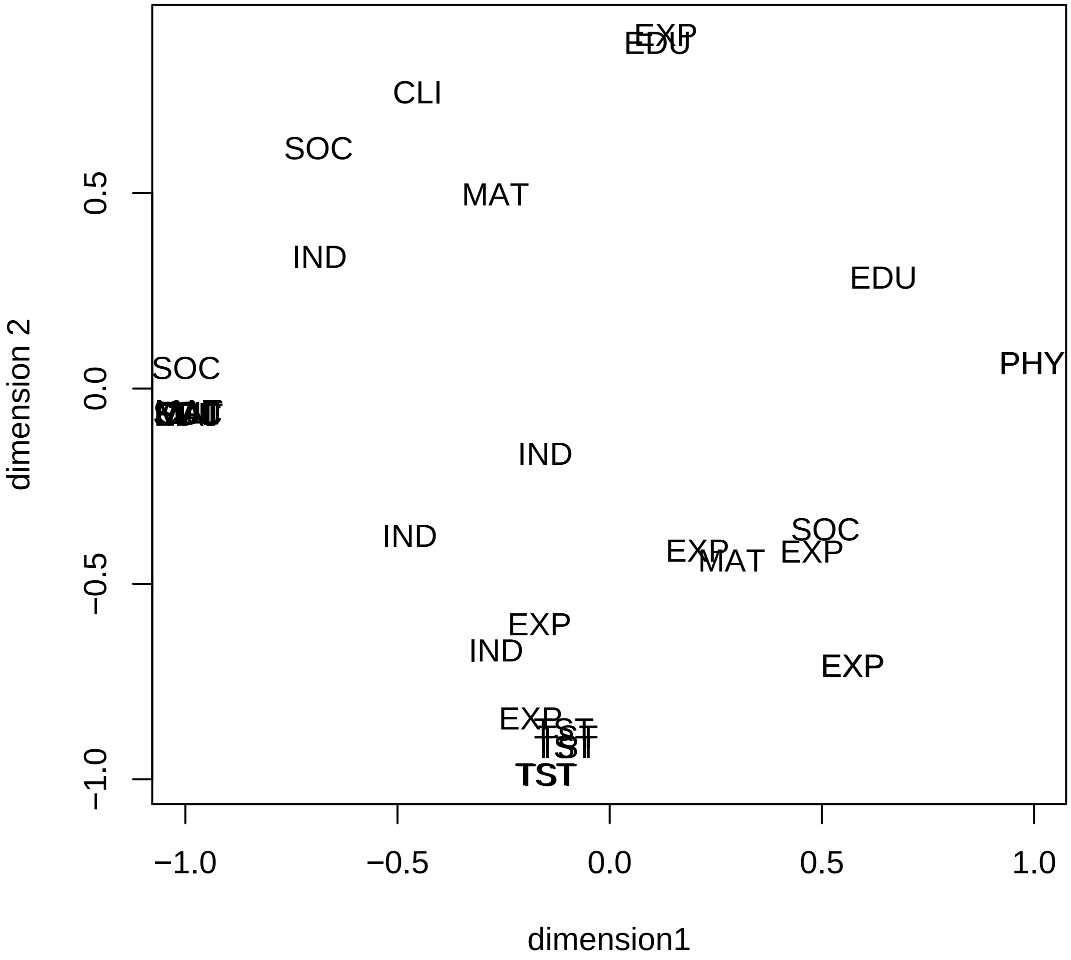

FIGURE 4. Roskam Data: Nonlinear PCA

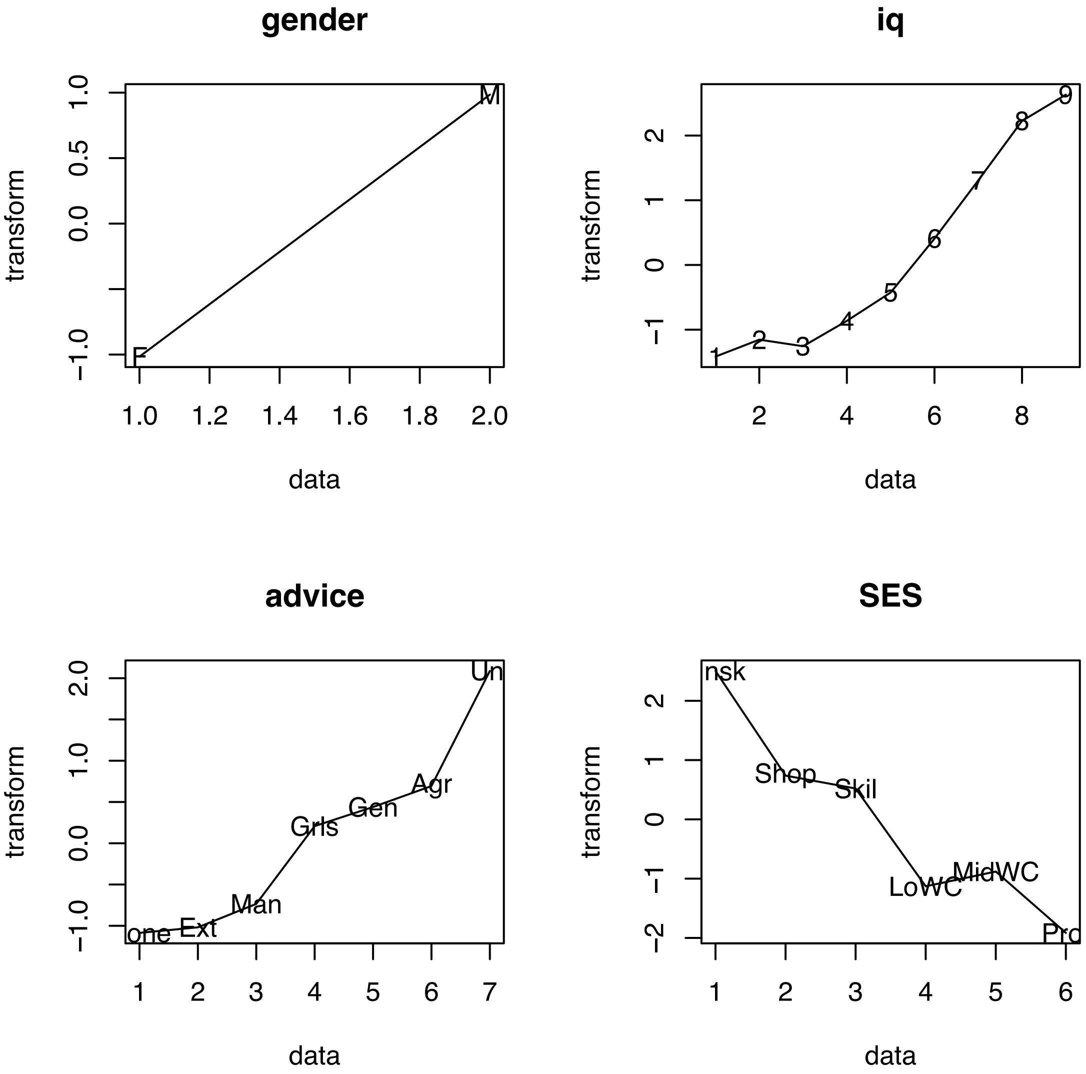

FIGURE 5. GALO Data: Transformations

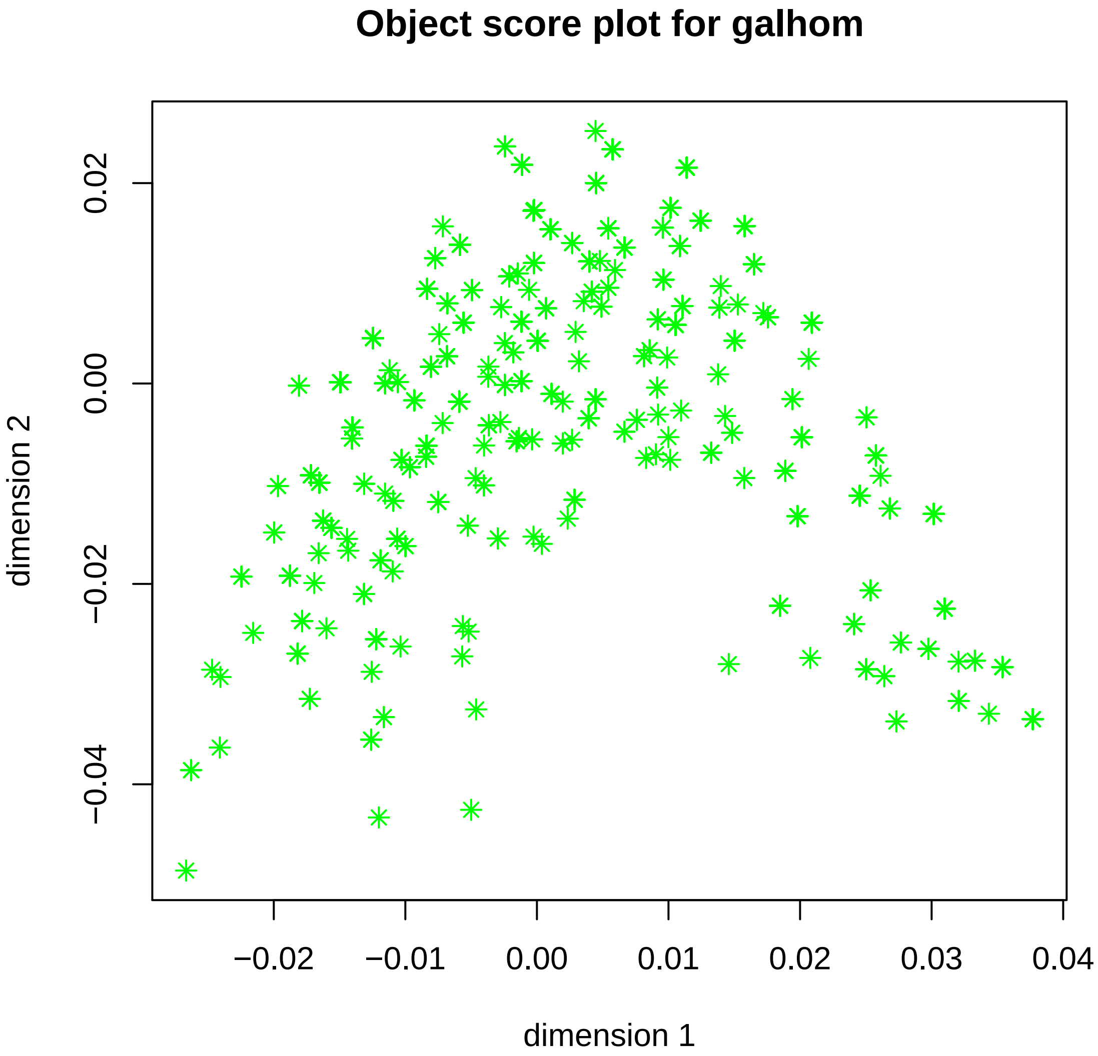

FIGURE 6. GALO Data: MCA

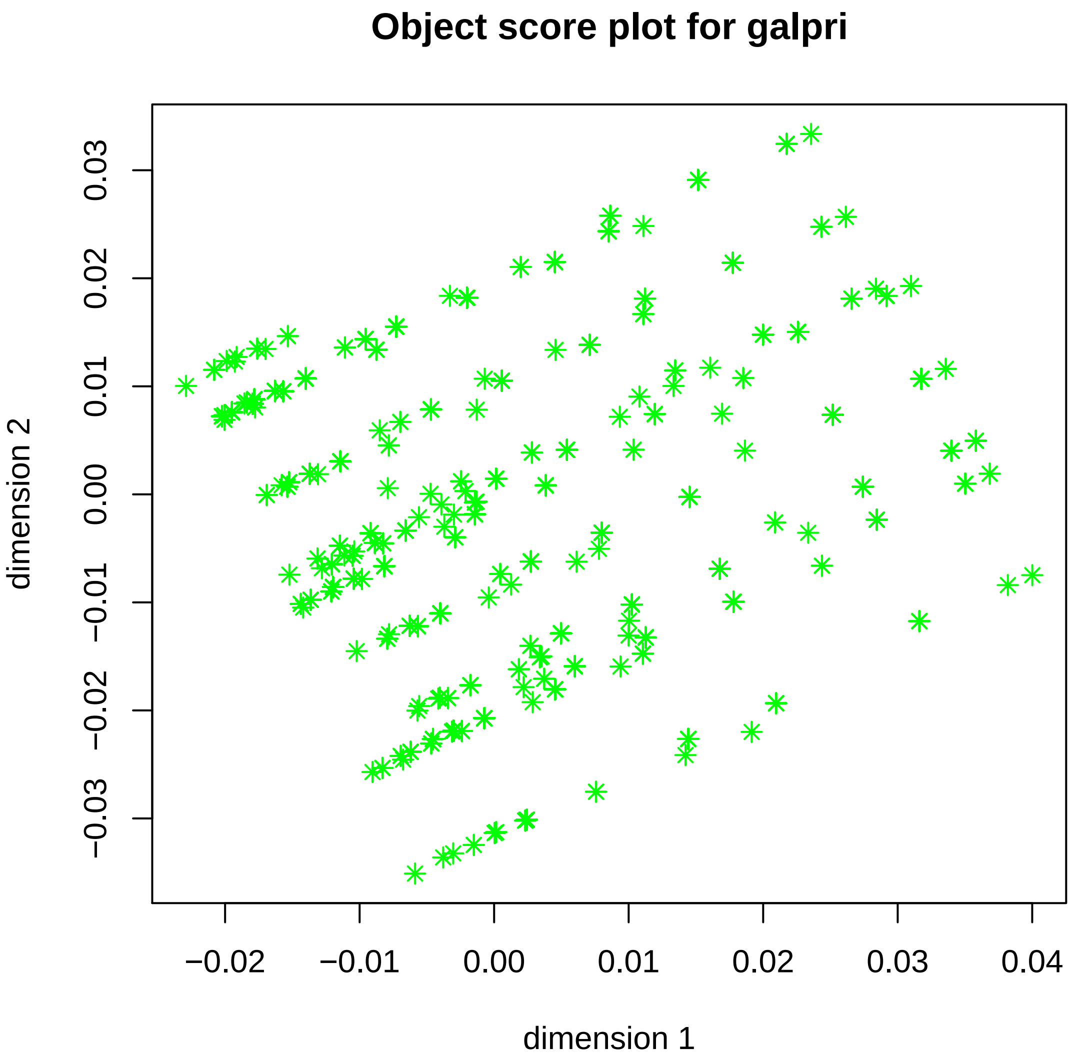

Figure 7. GALO Data: Nonlinear PCA

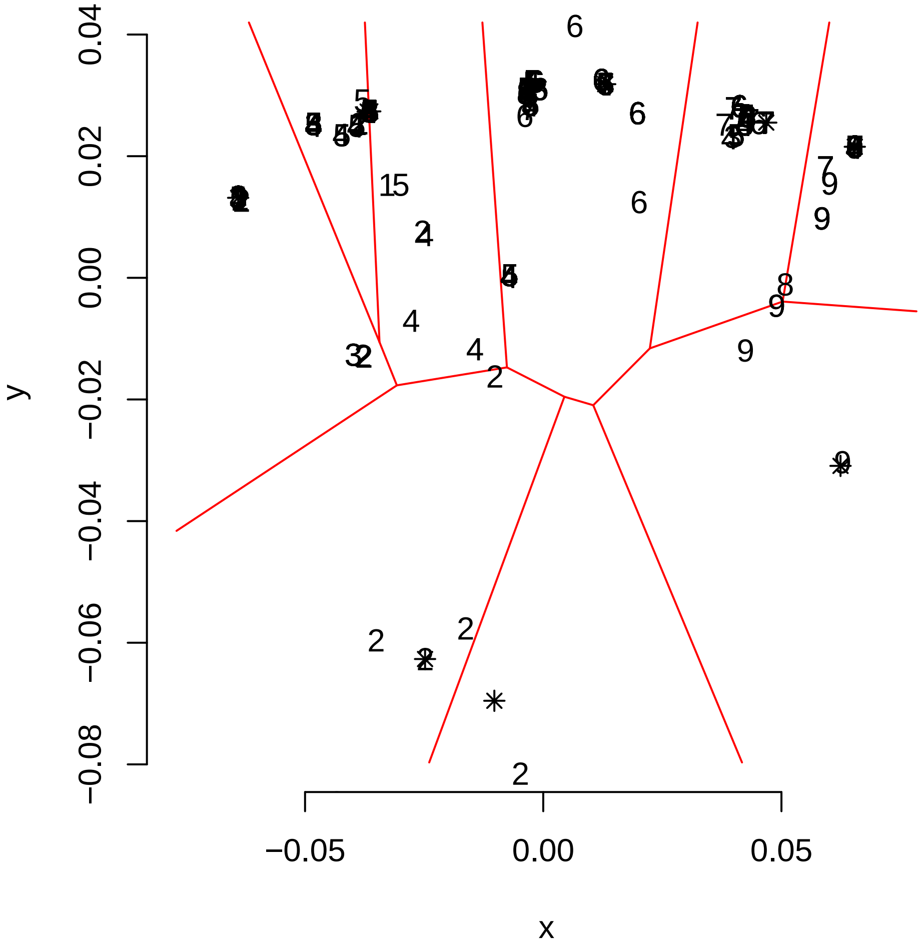

FIGURE 8. GALO Data, Intelligence, Logistic PCA

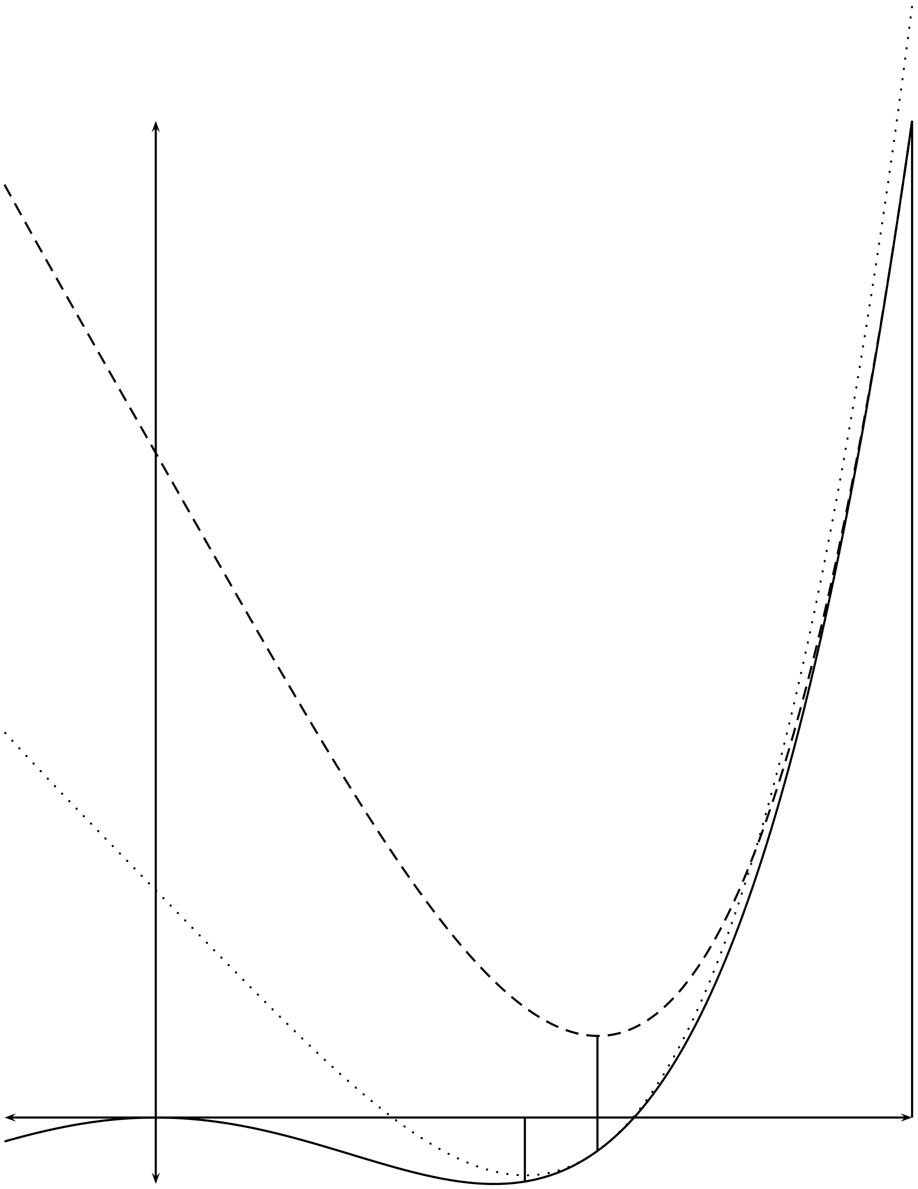

Figure 9. Majorization Example