UCLA Department of Statistics

History and Theory of Nonlinear Principal Component Analysis

Jan de Leeuw

February 11, 2011

Abstract

- Relationships between *Multiple Correspondence Analysis (MCA)* and *Nonlinear Principal Component Analysis (NLPCA)*, which is defined as PCA with Optimal Scaling (OS), are discussed. We review the history of NLPCA.
- We discuss forms of NLPCA that have been proposed over the years:
  - Shepard-Kruskal- Breiman-Friedman-Gifi PCA with optimal scaling,
  - Aspect Analysis of correlations,
  - Guttman's MSA,
  - Logit and Probit PCA of binary data, and
  - Logistic Homogeneity Analysis.
- Since I am trying to summarize 40+ years of work, the presentation will be rather dense.

Linear PCA

History

- (Linear) Principal Components Analysis (PCA) is sometimes 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 (notably from work by Cauchy and Jacobi in the mid 19th century).
- There are some modest beginnings in Galton's *Natural Inheritance* of 1889, 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 (7 physical traits of 3000 criminals) in MacDonell (1902), by a Pearson co-worker.
- There is proper attribution in: Burt, C., *Alternative Methods of Factor Analysis and their Relations to Pearson's Method of "Principle Axes"*, Br. J. Psych., Stat. Sec., 2 (1949), pp. 98-121.

Linear PCA

How To

- Hotelling's introduction of PCA follows the now familiar route of making successive orthogonal linear combinations with maximum variance. He does this by using Power iterations (without reference), discussed in 1929 by Von Mises and Pollaczek-Geiringer.
- Pearson, following Galton, used the correlation ellipsoid throughout. This seems to me the more basic approach.
- He cast 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.
- In modern notation, this means minimizing $\textbf{SSQ}(Y - XB')$ over $n \times r$ matrices X and $m \times r$ matrices B. For $r = 1$ this is the best line, etc.

Correspondence Analysis

History

- Simple Correspondence Analysis (CA) of a bivariate frequency table was first discussed, in fairly rudimentary form, by Pearson (1905), by looking at transformations linearizing regressions. See De Leeuw, *On the Prehistory of Correspondence Analysis*, Statistica Neerlandica, 37, 1983, 161–164.
- This was taken up by Hirshfeld (Hartley) in 1935, where the technique was presented in a fairly complete form (to maximize correlation and decompose contingency). This approach was later adopted by Gebelein, and by Renyi and his students in their study of maximal correlation.

Correspondence Analysis

History

- In the 1938 edition of *Statistical Methods for Research Workers* Fisher scores a categorical variable to maximize a ratio of variances (quadratic forms). This is not quite CA, because it is presented in an (asymmetric) regression context.
- Symmetric CA and the reciprocal averaging algorithm are discussed, however, in Fisher (1940) and applied by his co-worker Maung (1941a,b).
- In the early sixties the chi-square metric, relating CA to metric multidimensional scaling (MDS), with an emphasis on geometry and plotting, was introduced by Benzécri (thesis of Cordier, 1965).

Multiple Correspondence Analysis

History

- Different weighting schemes to combine quantitative variables to an index that optimizes some variance-based discrimination or homogeneity criterion were proposed in the late thirties by Horst (1936), by Edgerton and Kolbe (1936), and by Wilks (1938).
- The same idea was applied to quantitative variables in a seminal paper by Guttman (1941), that presents, for the first time, the equations defining *Multiple Correspondence Analysis (MCA)*.
- The equations are presented in the form of a row-eigen (scores), a column-eigen (weights), and a singular value (joint) problem.
- The paper introduces the "codage disjonctif complet" as well as the "Tableau de Burt", and points out the connections with the chi-square metric.
- There is no geometry, and the emphasis is on constructing a single scale. In fact Guttman warns against extracting and using additional eigen-pairs.

Multiple Correspondence Analysis

Further History

- In Guttman (1946) scale or index construction was extended to paired comparisons and ranks. In Guttman (1950) it was extended to scalable binary items.
- In the fifties and sixties 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.
- Starting in 1968, MCA was studied as a form of metric MDS by De Leeuw.
- Although the equations defining MCA were the same as those defining PCA, the relationship between the two remained problematic.
- These problems are compounded by “horse shoes” or the “effect Guttman”, i.e. artificial curvilinear relationships between successive dimensions (eigenvectors).

Nonlinear PCA

What ?

PCA can be made non-linear in various ways.

1. First, we could seek indices which discriminate maximally and are non-linear combinations of variables. This generalizes the weighting approach (Hotelling).
2. Second, we could find nonlinear combinations of components that are close to the observed variables. This generalizes the reduced rank approach (Pearson).
3. Third, we could look for transformations of the variables that optimize the linear PCA fit. This is known (term of Darrell Bock) as the *optimal scaling (OS)* approach.

Nonlinear PCA

Forms

- The first approach has not been studied much, although there are some relations with Item Response Theory.
- The second approach is currently popular in Computer Science, as "nonlinear dimension reduction". I am currently working on a polynomial version, but there is not 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 its history and discuss the main results.

Nonlinear PCA

PCA with OS

- Guttman observed in 1959 that if we require that the regression between monotonically transformed variables are linear, then the transformations are uniquely defined. In general, however, we need approximations.
- The loss function for PCA-OS is $\mathbf{SSQ}(Y - XB')$, as before, but now we minimize over components X, loadings B, and transformations Y.
- 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* type, where optimal transformation and low-rank matrix approximation are alternated until convergence.

PCA-OS

History of programs

- Shepard and Kruskal used the monotone regression machinery of non-metric MDS to construct the first PCA-OS programs around 1962. The paper describing the technique was not published until 1975.
- Around 1970 versions of PCA-OS (sometimes based on Guttman's rank image principle) were developed by Lingoes and Roskam.
- In 1973 De Leeuw, Young, and Takane started the ALSOS project, with resulted in PRINCIPALS (published in 1978), and PRINQUAL in SAS.
- In 1980 De Leeuw (with Heiser, Meulman, Van Rijckevorsel, and many others) started the Gifi project, which resulted in PRINCALS, in SPSS CATPCA, and in the R package homals by De Leeuw and Mair (2009).
- In 1983 Winsberg and Ramsay published a PCA-OS version using monotone spline transformations.
- In 1987 Koyak, using the ACE smoothing methodology of Breiman and Friedman (1985), introduced mdrace.

PCA/MCA

The Gifi Project

The Gifi project followed the ALSOS project. It has or had as its explicit goals:

1. Unify a large class of multivariate analysis methods by combining a single loss function, parameter constraints (as in MDS), and ALS algorithms.
2. Give a very general definition of component analysis (to be called *homogeneity analysis*) that would cover CA, MCA, linear PCA, nonlinear PCA, regression, discriminant analysis, and canonical analysis.
3. Write code and analyze examples for homogeneity analysis.

Gifi

Loss of Homogeneity

The basic Gifi loss function is

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

- The $n \times k_j$ matrices G_j are the *data*, coded as *indicator matrices* (or *dummies*). Alternatively, G_j can be a B-spline basis. Also, G_j can have zero rows for missing data.
- X is an $n \times p$ matrix of *object scores*, satisfying the *normalization conditions* $X'X = I$.
- Y_j are $k_j \times p$ matrices of *category quantifications*. There can be *rank, level* and *additivity constraints* on the Y_j.

Gifi

ALS

The basic Gifi algorithm alternates

1. $X^{(k)} = \mathbf{ORTH}(\sum_{j=1}^m G_j Y_j^{(k)})$.
2. $Y_j^{(k+1)} = \underset{Y_j \in \mathscr{Y}_j}{\mathbf{argmin}}\, \mathbf{tr}\, (\hat{Y}_j^{(k+1)} - Y_j)' D_j (\hat{Y}_j^{(k+1)} - Y_j)$.

We use the following notation.

- Superscript (k) is used for iterations.
- $ORTH()$ is any orthogonalization method such as QR, Gram-Schmidt, or SVD.
- $D_j = G_j' G_j$ are the *marginals*.
- $\hat{Y}_j^{(k+1)} = D_j^{-1} G_j' X^{(k)}$ are the *category centroids*.
- The constraints on Y_j are written as $Y_j \in \mathscr{Y}_j$.

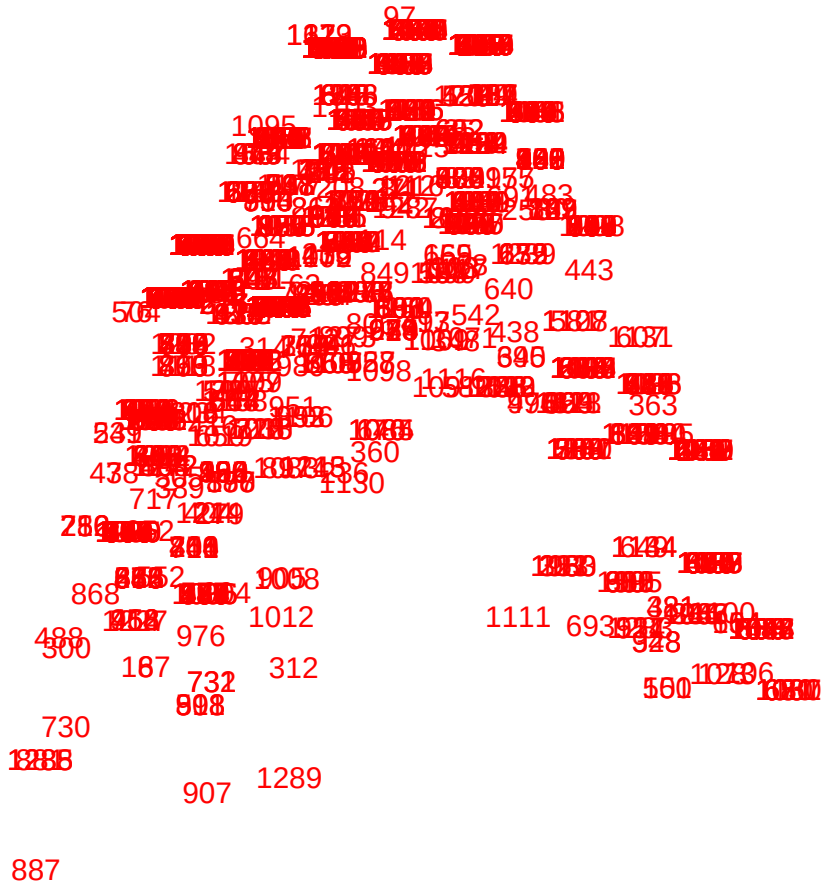
Gifi

Star Plots

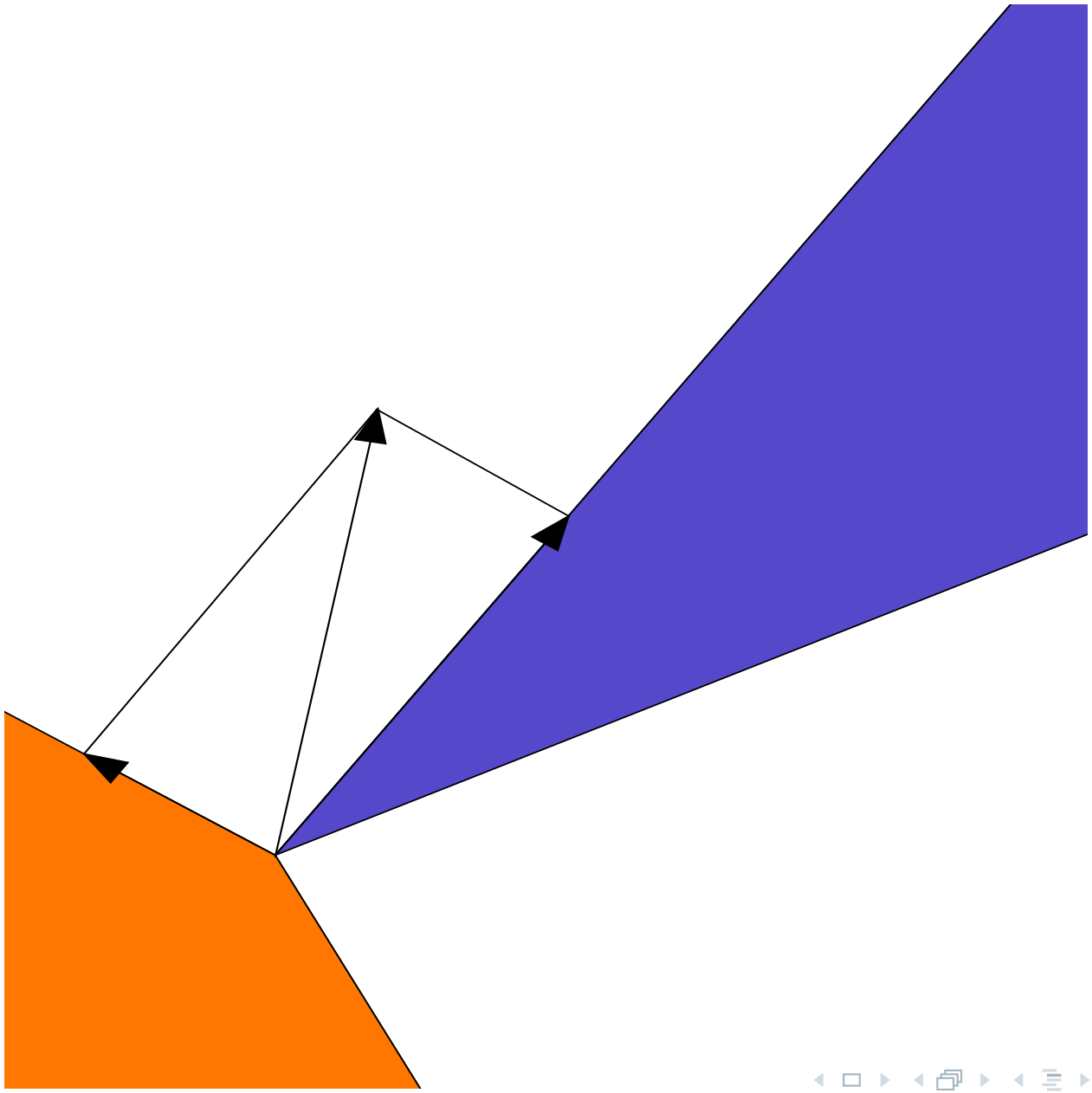
Let's look at some movies.

- GALO: 1290 students, 4 variables. We show both MCA and NLPCA.
- Senate: 100 senators, 20 votes. Since the variables are binary, MCA = NLPCA.

objplot galo



objplot senate



Gifi

Single Variables

- If there are no constraints on the Y_j homogeneity analysis is MCA.
- We will not go into additivity constraints, because they take us from PCA and MCA towards regression and canonical analysis. See the `homals` paper and package.
- A *single variable* has constraints $Y_j = z_j a_j'$, i.e. category quantifications are of rank one. In a given analysis some variables can be single while other can be multiple (unconstrained). More generally, there can be *rank constraints* on the Y_j.
- This can be combined with *level constraints* on the *single quantifications* z_j, which can be numerical, polynomial, ordinal, or nominal.
- If all variables are single homogeneity analysis is NLPCA (i.e. PCA-OS). This relationship follows from the form of the loss function.

Gifi

Multiple Variables

- There is another relationship, which is already implicit in Guttman (1941). If we transform the variables to maximize the dominant eigenvalue of the correlation matrix, then we find both the first MCA dimension and the one-dimensional nominal PCA solution.
- But there are deeper relations between MCA and NLPCA. These were developed in a series of papers by De Leeuw and co-workers, starting in 1980. Their analysis also elucidates the “Effect Guttman”.
- These relationships are most easily illustrated by performing an MCA of a continuous standardized multivariate normal, say on m variables, analyzed in the form of a Burt Table (with doubly-infinite subtables). Suppose the correlation matrix of this distribution is the $m \times m$ matrix $R = \{r_{j\ell}\}$.

Gifi

Multinormal MCA

- Suppose $I = R^{[0]}, R = R^{[1]}, R^{[2]}, \ldots$ is the infinite sequence of Hadamard (elementwise) powers of R.
- Suppose $\lambda_j^{[s]}$ are the m eigenvalues of $R^{[s]}$ and $y_j^{[s]}$ are the corresponding eigenvectors.
- The eigenvalues of the MCA solution are the $m \times \infty$ eigenvalues $\lambda_j^{[s]}$.
- The MCA eigenvector corresponding to $\lambda_j^{[s]}$ consists of the m functions $y_{j\ell}^{[s]} \mathcal{H}_s$, with $\mathcal{H}_s$ the s^{th} normalized Hermite polynomial.
- An MCA eigenvector consists of m linear transformations, or m quadratic transformations, and so on. There are m linear eigenvectors, m quadratic eigenvectors, and so on.

Gifi

Multinormal MCA

- The same theory applies to what Yule (1900) calls "strained multinormal" variables z_j, in which there exists diffeomorphisms ϕ_j such that $\phi_j(z_j)$ are jointly multinormal (an example are Gaussian copulas).
- And the same theory also applies, except for the polynomial part, when separate transformations of the variables exists that linearize all bivariate regressions (this generalizes a result of Pearson from 1905).
- Under all these scenarios, MCA solutions are NLPCA solutions, and vice versa.
- With the provision that NLPCA solutions are always selected from the same $R^{[s]}$, while MCA solutions come from all $R^{[s]}$.
- Also, generally, the dominant eigenvalue is $\lambda_1^{[1]}$ and the second largest one is either $\lambda_1^{[2]}$ or $\lambda_2^{[1]}$. In the first case we have a horseshoe.

Gifi

Bilinearizability

- The "joint bilinearizibility" also occurs (trivially) if $m = 2$, i.e. in CA, and if $k_j = 2$ for all j, i.e. for binary variables.
- If there is joint linearizability then the joint first-order asymptotic normal distribution of the induced correlation coefficients does not depend on the standard errors of the computed optimal transformations (no matter if they come from MCA or NLPCA or any other OS method).
- There is additional horseshoe theory, due mostly to Schriever (1986), that uses the Krein-Gantmacher-Karlin theory of total positivity. It is not based on families of orthogonal polynomials, but on (higher-order) order relations.
- This was, once again, anticipated by Guttman (1950) who used finite difference equations to derive the horseshoe MCA/NLPCA for the binary items defining a perfect scale.

More

Pavings

- If we have a mapping of n objects into $\mathbb{R}^p$ then a categorical variable can be used to label the objects.
- The subsets corresponding to the categories of the variables are supposed to be *homogeneous*.
- This can be formalized either as being *small* or as being *separated* by lines or curves. There are many ways to quantify this in loss functions.
- MCA (multiple variables, star plots) tends to think small (within vs between), NLPCA tends to think separable.
- Guttman's MSA defines outer and inner points of a category. An outer point is a closest point for any point not in the category.
- The closest outer point for an inner point should belong to the same category as the inner point. This is a nice "topological" way to define separation, but it is hard to quantify.

More

Aspects

- The aspect approach (De Leeuw and Mair, JSS, 2010, using theory from De Leeuw, 1990) goes back to Guttman's 1941 original motivation.
- An aspect is any function of all correlations (and/or the correkation ratios) between m transformed variables.
- Now choose the transformations/quantifications such that the aspect is maximized. We use majorization to turns this into a sequence of least squares problems.
- For MCA the aspect is the largest eigenvalue, for NLPCA it is the sum of the largest p eigenvalues.
- Determinants, multiple correlations, canonical correlations can also be used as aspects.
- Or: the sum of the differences of the correlation ratios and the squared correlation coefficients.
- Multinormal, strained multinormal, and bilinearizability theory applies to all aspects.

More

Logistic

- Instead of using least squares throughout, we can build a similar system using logit or probit log likelihoods. This is in the development stages.
- The basic loss function, corresponding to the Gifi loss function, is

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell=1}^{k_j} g_{ij\ell}\log\frac{\exp\{-\phi(x_i,y_{j\ell})\}}{\sum_{\nu=1}^{k_j}\exp\{-\phi(x_i,y_{j\nu})\}}$$

  where the data are indicators, as before. The function ϕ can be distance, squared distance, or negative inner product.
- This emphasizes separation, because we want the x_i closest to the $y_{j\ell}$ for which $g_{ij\ell}=1$.
- We use majorization to turn this into a sequence of reweighted least squares MDS or PCA problems.