HISTORY OF NONLINEAR PRINCIPAL COMPONENT ANALYSIS

JAN DE LEEUW

ABSTRACT. We discuss several forms of Nonlinear Principal Component Analysis (NLPCA) that have been proposed over the years: Linear PCA with optimal scaling, aspect analysis of correlations, Guttman's MSA, Logit and Probit PCA of binary data, and Logistic Homogeneity Analysis. They are compared with Multiple Correspondence Analysis (MCA), which we also consider to be a form of NLPCA.

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], Pages 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 3000 criminals) by a Pearson co-worker 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

Date: Monday 25th February, 2013 — 11h 10min — Typeset in TIMES ROMAN.

Key words and phrases. Template, L\LaTeXX.
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. In modern notation, this means minimizing $SSQ(Y - XB')$ over $n \times r$ matrices $X$ and $m \times r$ matrices $B$. For $r = 1$ this is the best line, for $r = 2$ it is the best plane, and so on.

2. CA AND MCA

The history of CA and MCA is reviewed expertly in the chapter by [Lebart and Saporta 2013]. We merely give some additional references that serve to connect MCA with NLPCA.

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 [Rényi 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 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].

Then in the early sixties 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].

2.2. **Multiple Correspondence Analysis.** 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], Edgerton and Kolbe [1936], and by 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 analyzing qualitative data for the war department. He presents, for the first time, the equations defining 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”, 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 explicitly against extracting and using additional eigenpairs.

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. See Lebart and Saporta [2013] 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 “horse shoes” or the “effect Guttman”, i.e. by artificial curvilinear relationships between successive dimensions (eigenvectors).
There are various ways in which we can introduce nonlinearity into PCA. First, we could seek indices which are non-linear combinations of variables that discriminate maximally in some sense. This generalizes the weighting approach of Hotelling. Second, we could find nonlinear combinations of unobserved components that are close to the observed variables. This generalizes the reduced rank approach of Pearson. Third, we could look for transformations of the variables that optimize the linear PCA fit. This is known (term of Bock [1960]) as the optimal scaling (OS) approach.

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. NLPCA WITH OPTIMAL SCALING

Guttman [1959] observed that if we require that the regression 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 $SSQ(Y - XB')$, as before, but now we minimize over components $X$, loadings $B$, and also over 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 (ALS) type, where optimal transformation and low-rank matrix approximation are alternated until convergence.
4.0.1. Software. PCA-OS became interesting after it became 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 non-metric breakthrough to construct the first PCA-OS programs around 1962. The 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]; [Lingoes, 1973].


5. NLPCA IN THE GIFI PROJECT

The Gifi project followed the ALSOS project. It’s 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 technique that fitted into the system were nonlinear regression, canonical analysis, and PCA.

The Gifi loss function is \( \sigma(X, Y) = \sum_{j=1}^{m} \text{SSQ}(X - G_j Y_j) \), which must be minimized over \( n \times p \) scores \( X \) for the \( n \) objects satisfying \( X'X = I \), and over \( k_j \times p \) category quantifications \( Y_j \) of the \( m \) variables. The \( G_j \) are the indicator matrices, coding category membership of the objects (“codage disjonctif complet”), where variable \( j \) has \( k_j \) categories. In the context of generalized
canonical analysis, this is identical to the loss function proposed 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 $Y_j$.

NLPCA results by imposing the restriction that $Y_j = z_j a_j'$, i.e. the category quantification are of rank one. Further restrictions can require the single quantifications $z_j$ to be either linear, polynomial, or monotonic functions of the original measurements. The monotonic case gives nonmetric PCA in the classical Kruskal-Shepard sense, the linear case gives classical linear PCA.

The relation between MCA and NLPCA was further investigated in a series of papers by De Leeuw and his students [De Leeuw 1982; Bekker and De Leeuw 1988; De Leeuw 1988; De Leeuw et al. 1999; De Leeuw 2006b]. The research is centered on assuming simultaneous linearizability of the regressions. This condition 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 NLPCA’s on a sequence of related correlation matrices. All solutions to the MCA equations are also solutions to the NLPCA equations. This also elucidates the horseshoe or Guttman effect and the role of rank-one constraints.
6. NLPCA Using Pavings

There is a geometrical approach NLPCA and MCA, which has given rise to other 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. In NLPCA we want these category subsets to be either small (relative to the whole set) or we want them to be 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 centroid. In PCA-OS we want separation of the subsets by parallel hyperplanes, and loss is defined as squared Euclidean distance to approximate separating hyperplanes. Total loss measures how well our smallness or separation criteria are satisfied over all variables.

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]). In particular Guttman’s MSA-I can be thought of as a form of NLPCA which has a way of measuring separation by using a pseudo-topological definition of inner and outer points of the subsets defined by the categories. There is an illustration in Guttman [1985].

7. NLPCA Using Aspects

In Mair and De Leeuw [2010] the R package aspect is described. This implements theory from De Leeuw [1988], 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 the variables. The correlation matrix itself is a function of the quantifications or transformations of the variables. The software maximize the aspect over transformations, by using majorization methods,
which are guaranteed to converge if the aspect is a convex function of the correlation matrix.

In MCA the aspect is the largest eigenvalue. Each MCA dimension provides a stationary value of the aspect. In PCA-OS the aspect is the sum of the largest $p$ eigenvalues. We can also easily define regression and canonical analysis in terms of the aspects they optimize. The R package also has a loss function defined as the sum of the differences between the squared correlation ratios and the 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].

8. 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 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 favor 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 $G_j$. Let us assume the probability that individual $i$ chooses alternative $\ell$ for variable $j$ is proportional to $\beta_{j\ell} \exp\{\phi(x_i, y_{j\ell})\}$, where $\phi$ is either the inner product, the negative Euclidean distance, or the negative squared Euclidean distance between vectors $x_i$ and $y_{j\ell}$. Assuming independent residuals we can now write down the negative log likelihood and minimize it over object scores and category quantifications.
This formulation allows for all the restrictions used in the Gifi project, replacing least squares by maximum likelihood and ALS by majorization [De Leeuw 2005]. The technique 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.

REFERENCES


L. Guttman. The Quantification of a Class of Attributes: A Theory and Method of Scale Construction. In P. Horst, editor, The Prediction of


L. Lebart and G. Saporta. Elements about the History of Correspondence Analysis and Multiple Correspondence Analysis. 2013.


SPSS. *SPSS Categories*. SPSS, Chicago, IL, 1989.


Department of Statistics, University of California, Los Angeles, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu