Title
History of Nonlinear Principal Component Analysis

Author
De Leeuw, Jan

Publication Date
2013-03-01

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, LaTeX.

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 $\mathbf{SSQ}(Y - XB')$ 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 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 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).

3. FORMS OF NONLINEAR PCA

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] Bock) 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 $\mathbf{SSQ}(Y - XB')$, 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].

In 1973 De Leeuw, Young, and Takane started the ALSOS project, with 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], in CATPCA in SPSS [SPSS, 1989], and in the `R` package `homals` De Leeuw and Mair [2009].

Winsberg and Ramsay [1983] published a PCA-OS version using monotone spline transformations. Koyak [1987], using the ACE smoothing methodology of Breiman and Friedman [1985], introduced `mdrace`.

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 $\sigma(X,Y)=\sum_{j=1}^m \mathbf{SSQ}(X-G_jY_j)$, 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 ℓ for variable j is proportional to $\beta_{j\ell}\exp\{\phi(x_i, y_{j\ell})\}$, where ϕ 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

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 Series in Probability and Mathematcal Statistics, chapter 1, pages 1–31. Wiley, Chichester, England, 1988. URL `http://www.stat.ucla.edu/~deleeuw/janspubs/1988/chapters/bekker_deleeuw_C_88.pdf`.

R.D. Bock. Methods and Applications of Optimal Scaling. Psychometric Laboratory Report 25, L.L. Thurstone Psychometric Laboratory, University of North Carolina, Chapell Hill, N.C., 1960.

L. Breiman and J. H. Friedman. Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, 80:580–619, 1985.

C. Burt. Alternative Methods of Factor Analysis and their Relations to Pearson's Method of "Principle Axes". *British Journal of Psychology, Statistical Section*, 2:98–121, 1949.

J.D. Carroll. A Generalization of Canonical Correlation Analysis to Three or More Sets of Variables. In *Proceedings of the 76th Annual Convention of the American Psychological Association*, pages 227–228, Washington, D.C., 1968. American Psychological Association.

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.

B. Cordier. *L'Analyse Factorielle des Correspondances*. Thèse de Troisieme Cycle, Université de Rennes, 1965.

J. De Leeuw. Canonical Discriminant Analysis of Relational Data. Research Note 007-68, Department of Data Theory FSW/RUL, 1968. URL http://www.stat.ucla.edu/~deleeuw/janspubs/1968/reports/deleeuw_R_68e.pdf.

J. De Leeuw. *Canonical Analysis of Categorical Data*. PhD thesis, University of Leiden, The Netherlands, 1973. Republished in 1984 by DSWO-Press, Leiden, The Netherlands.

J. De Leeuw. Nonlinear principal component analysis. In H. Caussinus, P. Ettinger, and R. Tomassone, editors, *COMPSTAT 1982*, pages 77–86, Vienna, Austria, 1982. Physika Verlag. URL http://www.stat.ucla.edu/~deleeuw/janspubs/1982/chapters/deleeuw_C_82.pdf.

J. De Leeuw. On the Prehistory of Correspondence Analysis. *Statistica Neerlandica*, 37:161–164, 1983. URL http://www.stat.ucla.edu/~deleeuw/janspubs/1983/articles/deleeuw_A_83b.pdf.

J. De Leeuw. Multivariate analysis with linearizable regressions. *Psychometrika*, 53:437–454, 1988. URL http://www.stat.ucla.edu/~deleeuw/janspubs/1988/articles/deleeuw_A_88a.pdf.

J. De Leeuw. Gifi Goes Logistic. Preprint Series 449, UCLA Department of Statistics, Los Angeles, CA, 2005. URL http://www.stat.ucla.edu/~deleeuw/janspubs/2005/reports/deleeuw_R_05a.pdf. SCASA 2005 Keynote.

J. De Leeuw. Principal Component Analysis of Binary Data by Iterated Singular Value Decomposition. *Computational Statistics and Data Analysis*, 50(1):21–39, 2006a. URL http:/www.stat.ucla.edu/~deleeuw/janspubs/2006/articles/deleeuw_A_06b.pdf.

J. De Leeuw. Nonlinear principal component analysis and related techniques. In M. Greenacre and J. Blasius, editors, *Multiple*

Correspondence Analysis and Related Methods, chapter 4, pages 107–133. Chapman and Hall, Boca Raton, FA, 2006b. URL http://www.stat.ucla.edu/~deleeuw/janspubs/2006/chapters/deleeuw_C_06b.pdf.

J. De Leeuw and P. Mair. Homogeneity Analysis in R: the Package homals. *Journal of Statistical Software*, 31(4):1–21, 2009. URL http://www.stat.ucla.edu/~deleeuw/janspubs/2009/articles/deleeuw_mair_A_09a.pdf.

J. De Leeuw and J.L.A. Van Rijckevorsel. Homals and princals: Some generalizations of principal components analysis. In *Data Analysis and Informatics*, Amsterdam, 1980. North Holland Publishing Company. URL http://www.stat.ucla.edu/~deleeuw/janspubs/1980/chapters/deleeuw_vanrijckevorsel_C_80.pdf.

J. De Leeuw, G. Michailidis, and D. Y. Wang. Correspondence analysis techniques. In S. Ghosh, editor, *Multivariate Analysis, Design of Experiments, and Survey Sampling*, pages 523–547. Marcel Dekker, 1999. URL http://www.stat.ucla.edu/~deleeuw/janspubs/1999/chapters/deleeuw_michailidis_wang_C_99.pdf.

H.A. Edgerton and L.E. Kolbe. The Method of Minimum Variation for the Combination of Criteria. *Psychometrika*, 1:183–187, 1936.

R.A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, London, England, 1938.

R.A. Fisher. The Precision of Discriminant Functions. *Annals of Eugenics*, 10:422–429, 1940.

F. Galton. *Natural Inheritance*. MacMillan and Co, London, GB, 1889.

H. Gebelein. Das Statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *Zeitschrift für Angewandte Mathematik Und Mechanik*, 21:364–379, 1941.

A. Gifi. *Nonlinear Multivariate Analysis*. Wiley, Chichester, England, 1990.

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, pages 321–348. Social Science Research Council, New York, 1941.

L. Guttman. An Approach for Quantifying Paired Comparisons and Rank Order. *Annals of Mathematical Statistics*, 17:144–163, 1946.

L. Guttman. The Principal Components of Scale Analysis. In S.A. Stouffer and Others, editors, *Measurement and Prediction*. Princeton University Press, Princeton, 1950.

L. Guttman. Metricizing Rank-ordered or Unordered Data for a Linear Factor Analysis. *Sankhya*, A21:257–268, 1959.

L. Guttman. Multidimensional Structuple Analysis (MSA-I) for the Classification of Cetacea: Whales, Porpoises and Dolphins. In J.-F. Marcotorchino, J.-M. Proth, and J. Janssen, editors, *Data Analysis in Real Life Environment: Ins and Outs of Solving Problems*, volume 8 of *Advanced Series in Management*, pages 45–54, Amsterdam and New York, 1985. North Holland Publishing Company.

H.O. Hirschfeld. A Connection between Correlation and Contingency. *Proceedings Cambridge Philosophical Society*, 31:520–524, 1935.

P. Horst. Obtaining a Composite Measure from a Number of Different Measures of the Same Attribute. *Psychometrika*, 1:53–60, 1936.

H. Hotelling. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24:417–441,498–520, 1933.

R. Koyak. On Measuring Internal Dependence in a Set of Random Variables. *Annals of Statistics*, 15:1215–1228, 1987.

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

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

J.A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, New York, N.Y., 2007.

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

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

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

W.R. MacDonell. On Criminal Anthropometry and the Identification of Criminals. *Biometrika*, 1:177–227, 1902.

P. Mair and J. De Leeuw. A General Framework for Multivariate Analysis with Optimal Scaling: The R Package aspect. *Journal of Statistical Software*, 32(9):1–23, 2010. URL `http://www.stat.ucla.edu/~deleeuw/janspubs/2010/articles/mair_deleeuw_A_10.pdf`.

K. Maung. Discriminant Analysis of Tocher's Eye Colour Data for Scottish School Children. *Annals of Eugenics*, 11:64–76, 1941a.

K. Maung. Measurement of Association in a Contingency Table with Special Reference to the Pigmentation of Hair and Eye Colour of Scottish School Children. *Annals of Eugenics*, 11:189–223, 1941b.

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

K. Pearson. On Certain Points Connected with Scale Order in the Case of a Correlation of Two Characters which for Some Arrangement Give a Linear Regression Line. *Biometrika*, 5:176–178, 1906.

K.T. Poole and H. Rosenthal. A Spatial Model for Legislative Roll Call Analysis. *American Journal of Political Science*, 29(2):357–384, 1985.

A. Renyi. On Measures of Dependence. *Acta Mathematica Academy Sciences Hungary*, 10:441–451, 1959.

E.E. 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.

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

R. Von Mises and H. Pollackzek-Geiringer. Practische Verfahren der Gleichungs-auflösung. *Zeitschrift für Angewandte Mathematik Und*

Mechanik, 9:58–79 and 152–164, 1929.

S.S. Wilks. Weighting Systems for Linear Functions of Correlated Variables when there is no Dependent Variable. *Psychometrika*, 3:23–40, 1938.

S. Winsberg and J. O. Ramsay. Monotone Spline Transformations for Dimension Reduction. *Psychometrika*, 48:575–595, 1983.

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. URL `http://www.stat.ucla.edu/~deleeuw/janspubs/1978/articles/young_takane_deleeuw_A_78.pdf`.

G. U. Yule. On the Methods of Measuring Association Between Two Attributes. *Journal of the Royal Statistical Society*, 75:107–170, 1912.

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`