ORTHOGONAL AND INDEPENDENT COMPONENT ANALYSIS

JAN DE LEEUW

ABSTRACT. This is a unified presentation of Principal Component Analysis (PCA), Factor Analysis (FA), and Independent Component Analysis (ICA), in both linear and nonlinear versions. We do not embed the techniques in an inferential framework with unobservable random variables or probabilities but present the as matrix approximation methods.

1. LS APPROXIMATION IN THE LINEAR CASE

1.1. **Linear PCA.** Suppose Y is an $n \times p$ data matrix. In PCA we minimize

$$\sigma_Y(X,A) = \mathbf{SSQ}(Y - XA') \tag{1}$$

over the $n \times r$ matrices X of *component scores* and the $p \times r$ matrices A of *component loadings*. We typically identify the problem by requiring $X'X = I$, which implies that $r \leq p$. The solution is usually found by singular value decomposition of Y, and the minimum loss function value is the sum of the $p - r$ residual singular values.

Instead of measuring loss directly by fitting the approximation to the data, we can also fit a closely related approximation to the cross products. If $Y = XA'$, with $X'X = I$, then $C = Y'Y = AA'$. Thus alternatively we can minimize

$$\sigma_C(A) = \mathbf{SSQ}(C - AA'). \tag{2}$$

Date: Tuesday 12^{th} July, 2011 — 14h 59min — Typeset in TIMES ROMAN.

The solution is found from the eigenvalue decomposition of C, and the minimum loss function value is the sum of the $p-r$ residual eigenvalues, i.e. the sum of squares of the $p-r$ residual singular values of Y.

In this most basic case approximating the data matrix and approximating the cross products give the same results. But this equivalence no longer holds for more complicated cases in which there are missing data, or transformations of the variables, or constraints on the loadings.

1.2. **Linear FA.** In FA we minimize

$$\sigma_Y(X,A,U,D) = \mathbf{SSQ}(Y - XA' - UD) \tag{3}$$

over the *common factor scores* X and *common factor loadings* A as before, but in addition we minimize over the $n \times p$ matrix U of *unique factor scores* and the $p \times p$ diagonal matrix of *unique loadings*. We again assume (in exploratory factor analysis) that $X'X = I$, and in addition that $X'U = 0$ and $U'U = I$. The minimization problem is more complicated in this case [De Leeuw, 2004; Unkel and Trendafilov, 2010].

Again there is a cross product version, which is derived by "ignoring errors". That phrase simply means that if $Y = XA' + UD$, with $(X \mid U)$ orthonormal, then $C = Y'Y = AA' + D^2$. In fact the "Fundamental Theorem of Factor Analysis" tells us that the reverse implication is true as well, although the scores $(X \mid U)$ cannot be recovered uniquely from Y, A, and D. See Appendix B for the precise result. The cross product loss is

$$\sigma_C(A,D) = \mathbf{SSQ}(C - AA' - D^2). \tag{4}$$

In the FA case minimizing data loss (3) and cross product loss (4) generally give different results, except in the case of perfect fit.

1.3. **General.** Both linear PCA and linear FA can be thought of as minimizing loss functions (1) or (2), but with different constraints on the scores and loadings. PCA is basically unconstrained, the orthogonality constraints are just used for identification. In FA, because of the diagonality constraints on the unique loadings, the number of factors is larger than the number of variables. This causes some indeterminacy, but it does not really modify the basic iterative algorithms. Thus it makes sense to think of Linear Component Analysis (LCA) as a general matrix approximation technique, with various restrictions on the loadings and scores, and with the number of components not necessarily smaller than the number of variables. This also makes Confirmatory Factor Analysis, with linear constraints on the common factor loadings, a form of LCA. In addition, Nonnegative Components Analysis, with non-negative loadings, fits in this framework as well.

In the same way as we have derived loss functions defined on the cross products, we could also derive loss functions derived on higher order multivariate moments or cumulants. But because we require only orthogonality of scores, not independence, this is not particularly natural and just leads to complicated fitting problems. By "ignoring errors" we see, for example, that

(5a) $$\sum_{i=1}^{n} y_{ij}y_{i\ell}y_{ik} = \sum_{s=1}^{r}\sum_{t=1}^{r}\sum_{u=1}^{r} h_{stu}a_{js}a_{\ell t}a_{ku},$$

with “core array”

(5b) $$h_{stu} = \sum_{i=1}^{n} x_{is}x_{it}x_{iu}.$$

Orthonormality does not give any additional structure to the core array, and the minimization of the loss function defined on third order moments will be complicated (although it could be used to make FA solutions determinate).

1.4. **Extensions.**

1.4.1. *Weighted Loss.*

There are other ways of measuring distance between Y and XA', or between C and AA'.

1.4.2. *Asymptotically Distribution Free Loss.*

For data matrix distances various power norms have been tried, regularization penalty terms have been added, and so on. This leads to robust and sparse versions of component analysis. The basic idea of direct matrix approximation is preserved.

For cross product matrix distances a fairly general approach has been proposed by Swain [1975] in the case of exploratory linear FA. We have R code for this. Most of the Swain alternative distances are variations of multinormal maximum likelihood (i.e. the KL distance), but there are also asymptotically distribution free methods which use weighted least squares, with weights computed from fourth order product moments.

1.4.3. *LPCA and LFA with Missing Data.*

1.4.4. *LPCA and LFA with Optimal Scaling.*

2. Independent Component Analysis

We can go from orthogonality to independence. In this context, "independence" of the components is

$$\sum_{i=1}^{n} w_i \prod_{j=1}^{m} x_{is}^{\alpha_s} = \prod_{j=1}^{m} \sum_{i=1}^{n} w_i x_{is}^{\alpha_s} \tag{6}$$

where w is a vector with non-negative weights adding up to one. This constraint is imposed for a number of vector of integer powers α. This generalizes orthogonality, and is equivalent to it if we only consider product moments up to order two.

It is difficult to directly impose the constraints (6) when fitting LCA to the data matrix. But the constraints become useful when deriving loss functions based on higher order product moments (or, preferably, cumulants). Using (6) dramatically simplifies the core array of (5b).

This does leads directly to Linear Independent Component Analysis [Hyvärinen et al., 2001; Comon and Jutten, 2010], and to multilinear decomposition of multidimensional arrays of cumulants. New algorithms for LICA are being developed at UCLA Statistics in the dissertation project of Irina Kukuyeva.

3. Polynomial Component Analysis

We can go from linearity to polynomiality, giving Polynomial Component Analysis, with both orthogonal and independent versions (POCA and PICA).

The approximation we are fitting is, for column j of the data matrix,

$$y_j \approx P_j(x_1, \cdots, x_r),$$

with the P_j multivariate polynomials and with the columns of X either orthonormal or "independent". Again, in the case of constraints (6), we would not impose them directly but switch to a loss function defined on the higher order product moments.

New algorithms for POCA are being developed at UCLA Statistics in the dissertation project of Kekona Sorensen.

Appendix A. Linear Factor Analysis

In LCA with data-loss we minimize $\sigma(X,A) = \mathbf{SSQ}(Y - XA')$ over $X'X = I$ and A, possibly with constraints on A. Y is $n\times m$, X is $n\times p$, and A is $m\times p$. We do not exclude cases in which $p > m$, and/or in which $\mathbf{rank}(Y) < \min(n,m)$.

Define $\sigma(\star,A) = \min_{X'X=I}\sigma(X,A)$. Then

$$\sigma(\star,A) = \mathbf{SSQ}(Y) + \mathbf{SSQ}(A) - 2\max_{X'X=I}\mathbf{tr}\ X'YA$$

Suppose $Z \stackrel{\Delta}{=} YA$ is an $n\times p$ matrix of rank r. Consider the problem of maximizing $\mathbf{tr}\ X'Z$ over the $n\times p$ matrices X satisfying $X'X = I$. This is known as the *Procrustus* problem, and it is usually studied for the case $n \geq m = r$. We want to generalize to $n \geq m \geq r$.

For this, following De Leeuw [2004], we use the singular value decomposition

$$Z = \begin{bmatrix} \underset{n\times r}{K_1} & \underset{n\times(n-r)}{K_0} \end{bmatrix} \begin{bmatrix} \underset{r\times r}{\Lambda} & \underset{r\times(p-r)}{0} \\ \underset{(n-r)\times r}{0} & \underset{(n-r)\times(p-r)}{0} \end{bmatrix} \begin{bmatrix} \underset{r\times p}{L_1'} \\ \underset{(p-r)\times p}{L_0'} \end{bmatrix}.$$

Theorem A.1. *The maximum of* $\mathbf{tr}\ X'Z$ *over* $n\times p$ *matrices* X *satisfying* $X'X = I$ *is* $\mathbf{tr}\ \Lambda$*, and it is attained for any* X *of the form* $X = K_1L_1' + K_0VL_0'$*, where* V *is any* $(n-r)\times(p-r)$ *matrix satisfying* $V'V = I$*.*

Proof. Using a symmetric matrix of Lagrange multipliers M leads to the stationary equations $Z = XM$, which implies $Z'Z = M^2$ or $M = \pm\ (Z'Z)^{1/2}$. It also implies that at a solution of the stationary equations $\mathbf{tr}\ X'Z = \pm\ \mathbf{tr}\ \Lambda$. The negative sign corresponds with the minimum, the positive sign with the maximum.

Now

$$M = \begin{bmatrix} \underset{p\times r}{L_1} & \underset{p\times(p-r)}{L_0} \end{bmatrix} \begin{bmatrix} \underset{r\times r}{\Lambda} & \underset{r\times(p-r)}{0} \\ \underset{(p-r)\times r}{0} & \underset{(p-r)\times(p-r)}{0} \end{bmatrix} \begin{bmatrix} \underset{r\times pm}{L_1'} \\ \underset{(p-r)\times p}{L_0'} \end{bmatrix}.$$

If we write X in the form

$$X = \begin{bmatrix} \underset{n\times r}{K_1} & \underset{n\times(n-r)}{K_0} \end{bmatrix} \begin{bmatrix} \underset{r\times p}{X_1} \\ \underset{(n-r)\times p}{X_0} \end{bmatrix}$$

then $Z = XM$ can be simplified to

$$X_1 L_1 = I,$$

$$X_0 L_1 = 0,$$

with in addition, of course, $X_1'X_1 + X_0'X_0 = I$. It follows that $X_1 = L_1'$ and

$$\underset{(n-r)\times p}{X_0} = \underset{(n-r)\times(p-r)}{V} \underset{(p-r)\times p}{L_0'},$$

with $V'V = I$. Thus $X = K_1 L_1' + K_0 V L_0'$. □

Corollary A.2.

$$\sigma(\star, A) = \mathbf{SSQ}(Y) + \mathbf{SSQ}(A) - 2\sum_{s=1}^{p} \lambda_s(A'CA),$$

where the λ_s are the square roots of the ordered eigenvalues of $A'CA$.

Proof. Directly from the Theorem. □

Corollary A.2 shows that the optimal factor loadings are continuous, and usually differentiable, functions of the cross product matrix C, even if we use data loss. This means that if the cross products are asymptotically normal, then so are the loadings estimates. Computing their asymptotic standard errors is tedious, but straightforward.

Theorem A.1 shows that the factor scores are never unique if $p > r$, no matter what the constraints on A are. In the FA literature this is known as the “indeterminacy problem”, and it has often been discussed in rather mysterious terms.

Appendix B. The Fundamental Theorem of Factor Analysis

There is a theorem closely related to theorem A.1 which is known, or used to be known, as the "fundamental theorem of factor analysis". It took the cumulative efforts of many fine minds, starting with Spearman, about 25 years to come up with a proof of this theorem. The fact that it follows easily from the singular value decomposition shows the power of modern matrix algebra tools. We use the same reasoning and notation as in appendix A.

Theorem B.1. *Suppose* $\underset{n\times m}{Y}$ *and* $\underset{m\times p}{A}$ *are such that* $X'X = AA'$. *Then there is an* $\underset{n\times p}{X}$ *such that* $X'X = I$ *and* $Y = XA'$.

Proof. From $Y'Y = AA'$ we know that A has singular value decomposition

$$A = \begin{bmatrix} \underset{m\times r}{L_1} & \underset{m\times(m-r)}{L_0} \end{bmatrix} \begin{bmatrix} \underset{r\times r}{\Lambda} & \underset{r\times(p-r)}{0} \\ \underset{(m-r)\times r}{0} & \underset{(m-r)\times(p-r)}{0} \end{bmatrix} \begin{bmatrix} \underset{r\times p}{V_1'} \\ \underset{(p-r)\times p}{V_0'} \end{bmatrix},$$

where $r \leq p$ is the rank of both Y and A. Observe that the left singular vectors of A are the right singular vectors of Y.

Now we still have to solve $Y = XA'$. Write

$$X = \begin{bmatrix} \underset{n\times r}{K_1} & \underset{n\times(n-r)}{K_0} \end{bmatrix} \begin{bmatrix} \underset{r\times p}{X_1} \\ \underset{(n-r)\times p}{X_0} \end{bmatrix}.$$

Then $Y = XA'$ simplifies to

$$I = X_1 V_1,$$
$$0 = X_0 V_1,$$

with in addition, of course, $X_1'X_1 + X_0'X_0 = I$. It follows that $X_1 = V_1'$ and

$$\underset{(n-r)\times p}{X_0} = \underset{(n-r)\times(p-r)}{W} \; \underset{(p-r)\times p}{V_0'},$$

with $W'W = I$. Thus $X = K_1 V_1' + K_0 W V_0'$. □

REFERENCES

P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation*. Academic Press, Amsterdam, The Netherlands, 2010.

J. De Leeuw. Least Squares Optimal Scaling of Partially Observed Linear Systems. In K. van Montfort, J. Oud, and A. Satorra, editors, *Recent Developments on Structural Equation Models*, chapter 7. Kluwer Academic Publishers, Dordrecht, Netherlands, 2004. URL `http://preprints.stat.ucla.edu/360/lserr.pdf`.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons Inc., New York, N.Y., 2001.

A.J. Swain. A Class of Factor Analysis Estimation Procedures with Common Asymptotic Sampling Properties. *Psychometrika*, 40:315–335, 1975.

S. Unkel and N.T. Trendafilov. A Majorization Algorithm for Simultaneous Parameter Estimation in Robust Exploratory Factor Analysis. *Computational Statistics and Data Analysis*, 2010.

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`