LEAST SQUARES METHODS FOR FACTOR ANALYSIS

JAN DE LEEUW AND JIA CHEN

ABSTRACT. Meet the abstract. This is the abstract.

1. INTRODUCTION

Suppose we have n measurements on each of m variables. Collect these measurements in an $n \times m$ matrix Y. In this introductory section we briefly review two classical factor analysis models, more precisely linear common factor models. For a more comprehensive discussion we refer to Anderson and Rubin [1956] and to Anderson [1984].

In common factor analysis we suppose that Y is the sum of a *common part* and a *unique part*. This is analogous to discussing data as composed of a signal and a noise part (or a fit and error part) in other data analysis contexts. We write the model, informally, in algebraic form[1] as

$$\begin{aligned} \underset{n\times m}{Y} &= \underset{n\times m}{F} + \underset{n\times m}{U}, \\ \underset{n\times m}{F} &= \underset{n\times p}{H}\,\underset{p\times m}{A}, \\ \underset{n\times m}{U} &= \underset{n\times m}{E}\,\underset{m\times m}{D}. \end{aligned}$$

Date: Saturday 14th January, 2012 — 14h 33min — Typeset in LUCIDA BRIGHT.

[1]Observe that we show the dimensions of a matrix by giving the numbers of rows and columns under the symbol of the matrix.

Thus the common part consists of linear combinations of p common factors, and the unique part of linear combinations of m unique factors. There is also a notion that the common and unique parts are orthogonal or independent in some sense. This last notion will be formalized next.

2. Factor Analysis Models

There are various ways in which the general idea of factor analysis can be made more precise by formulating it as an explicit statistical model.

2.1. **Random Factor Model.**

The matrix Y is supposed to be a realization[2] of a matrix-valued random variable $\underline{Y}$.

In random score factor analysis we assume that the random variable $\underline{Y}$ has a random common part $\underline{F}$ and a random unique part $\underline{U}$. Thus

$$\underset{n\times m}{\underline{Y}} = \underset{n\times m}{\underline{F}} + \underset{n\times m}{\underline{U}} .$$

The common part is a linear combination of a number, say p, of common factors $\underline{H}$, i.e.

$$\underset{n\times m}{\underline{F}} = \underset{n\times p}{\underline{H}} \; \underset{p\times m}{A'} .$$

The unique part is a linear combination of m unique factors $\underline{E}$.

$$\underset{n\times m}{\underline{U}} = \underset{n\times m}{\underline{E}} \; \underset{m\times m}{D'} .$$

The rows of $\underline{Y}$, corresponding with the different individuals, are assumed to be independent. Moreover we assume the specific parts are uncorrelated with the common factors, and with the other specific parts. For simplicity we assume all variables are centered, i..e have expectation zero.

[2] We use the convention of underlining random variables [Hemelrijk, 1966].

2.2. **Fixed Factor Model.** The random factor model explained above was criticized soon after it was formally introduced by Lawley.

> The point is that in factor analysis different individuals are regarded as drawing their scores from *different* k-way distributions, and in these distributions the mean for each test is the true score of the individual on that test. Nothing is implies about the distribution of observed scores over a population of individuals, and one makes assumptions only about the error distributions [Young, 1940, pag. 52].

Young proposed the *fixed factor model*, which assumes

$$\underline{Y} = F + \underline{E}.$$

Now the common part is a bilinear combination of a number of common factor loadings a_{js} and common factor scores u_{is}, i.e.

$$F = UA'.$$

In the fixed model we merely assume the specific parts are uncorrelated with the other specific parts.

2.3. **Covariance Models.**

3. Estimation

3.1. **Covariance Matrix Methods.** The dominant estimation method in factor analysis is multinormal maximum likelihood for the random factor model. It was first proposed by Lawley [1939], and then popularized and programmed by Jöreskog [1967]. The negative log-likelihood measures the distance between the sample and population covariance model, and we must minimize

$$\mathcal{L}(A, D) = n \log |\Sigma| + n \ \mathbf{tr}\ \Sigma^{-1} S,$$

with S the sample covariance matrix of Y, and with $\Sigma = A\Omega A' + \Delta^2$.

In Anderson and Rubin [1956] the impressive machinery developed by the Cowles Commission was applied to both the fixed and random factor analysis model. Maximum likelihood was applied to the likelihood function of the covariance matrix, assuming multivariate normality.

3.2. **Data Matrix Methods.** Lawley's maximum likelihood procedure was criticized soon after it appeared by Young [1940], because it was inappropriate for the fixed factor model that Young favored.

> Such a distribution is specified by the means and variances of each test and the covariances of the tests in pairs; it has no parameters distinguishing different individuals. Such a formulation is therefore inappropriate for factor analysis, where factor loadings of the tests and of the individuals enter in a symmetric fashion in a bilinear form [Young, 1940, pag. 52].

Young proposed to minimize the log-likelihood of the data

$$\mathcal{L}(U,A,D)=n\log|D|+\mathbf{tr}\ (Y-UA')'D^{-1}(Y-UA')$$

where D is a *known* diagonal matrix with column (variable) weights. The solution is given by a weighted singular value decomposition of Y.

The basic problem with Young's method is that it supposes the weights to be known. One solution, suggested by Lawley [1942], is to estimate them along with the loadings and scores.

Lawley suggests to alternate minimization over (U,A), which is done by weighted singular value decomposition, and minimization over diagonal D, which simply amounts to computing the average sum of squares of the residuals for each variable. Iterating these two minimizations produces a block relaxation algorithm intended to minimize the negative log-likelihood. Although the algorithm

obviously produces a decreasing sequence of loss function values, it does not work.

> A rather disconcerting feature of the new method is, however, that iterative numerical solutions of the estimation equations either fail to converge, or else converge to unacceptable solutions in which one of more of the measurements have zero error variance. It is apparently impossible to estimate scale as well as location parameters when so many unknowns are involved [Whittle, 1952, pag. 224].

In fact, if we look at the loss function we can see it is unbounded below. We can choose scores to fit one variable perfectly, and then let the corresponding variance term approach zero [Anderson and Rubin, 1956].

Several other remedies have been proposed to rescue the weighted least squares methods. Whittle [1952] suggested to take D proportional to the variances of the variables. This amounts to doing a singular value decomposition of the standardized variables. Jöreskog [1962] makes the more reasonable choice of setting D proportional to the reciprocals of the diagonals of the inverse of the covariance matrix of the variables (i.e. to the residual variances when regressing each variable on the others). Of course in these approaches the weights themselves depend on the data Y, which means that simple weighted least squares theory does not apply.

An original approach was suggested by McDonald [1979]. Also see Etezadi-Amoli and McDonald [1983]. He proposes to maximize the determinant of the correlation matrix of the matrix of residuals $R = Y - UA'$. This criterion can be derived by using the fact that if we minimize over diagonal D, then

$$\min_D \mathcal{L}(U, A, D) = n \log |\mathbf{diag} R'R|,$$

while if we minimize over unrestricted S we have

$$\min_S \mathcal{L}(U, A, S) = n \log |R'R|,$$

The difference of the two is the logarithm of the determinant of the correlation matrix of the residuals. The approach is clearly scale-free, and the maximum of zero is attained if we can make the residuals exactly uncorrelated. Computational and statistical properties of this so-called *maximum likelihood ratio method* are quite complicated, however.

4. Unweighted Least Squares

4.1. General Considerations.

4.2. Non-negativity Constrains.

5. Least Squares on the Covariances

5.1. **Loss.** The least squares loss function used in LSFAC is

$$\phi(A, D) \triangleq \frac{1}{2}\mathbf{SSQ}(C - AA' - D). \tag{1}$$

where $A \in \mathbb{R}^{m\times p}$ and $D \in \mathbb{D}^m$, the diagonal matrices of order m.

5.2. **Projections.** We also define the two projected or concentrated loss functions, in which one set of parameters is "minimized out",

$$\phi(A) \triangleq \min_{D\in\mathbb{D}^m} \phi(A, D) = \sum_{1\leq j<\ell\leq m} (c_{j\ell} - a_j' a_\ell)^2, \tag{2a}$$

and

$$\phi(D) \triangleq \min_{A\in\mathbb{R}^{m\times p}} \phi(A, D) = \frac{1}{2}\sum_{s=p+1}^{m} \lambda_s^2(C - D). \tag{2b}$$

Note that ϕ is used as a generic symbol for these LSFAC loss functions, because it will be clear from the context which ϕ we are using.

5.3. **Algorithms.** There have been four major approaches to minimizing this loss function.

5.3.1. *Thomson's Principal Factor Analysis.* PFA [Thomson, 1934] is an alternating least squares (ALS) method [De Leeuw, 1994], in which we alternate minimizing over A for D fixed at its current value and minimizing over D for A fixed at its current value.

The minimum over D for fixed A is attained at $D = \mathbf{diag}(C - AA')$. If $C - D = K\Lambda K'$ is the eigen-decomposition of $C - D$, and we write Λ_p and K_p for the p largest eigenvectors and corresponding eigenvectors, then the minimum over A for fixed D is $A = K\Lambda^{\frac{1}{2}}$. If fewer than p eigenvalues are positive, then the negative elements in Λ_p are replaced by zeroes.

Because $D = \mathbf{diag}(C - AA')$ we always have $D \lesssim \mathbf{diag}(C)$, but there is no guarantee that convergence is to a D for which both $D \gtrsim 0$ and $C - D \gtrsim 0$.

5.3.2. *Comrey's MRFA.* Comrey [1962] proposed minimum residual factor analysis, which was a (non-rigorous) attempt to minimize the projected loss function. The method was put on a more solid footing by Zegers and Ten Berge [1983].

5.3.3. *Harman's MINRES.* In MINRES [Harman and Jones, 1966; Harman and Fukuda, 1966] we project out D. We use ALS to minimize the projected loss function $psi(A)$ from (2a) over A, using the m rows as blocks.

5.3.4. *Gradient and Newton Methods.* Gradient methods can be most conveniently applied by projecting out A. Thus we work with $\phi(D)$ from (2b), and minimize over D. Now use

$$\begin{aligned}\mathcal{D}_j\lambda_s(D) &= -z_{js}^2,\\ \mathcal{D}_j z_{\ell s}(D) &= z_{js}(C - D - \lambda_s I)^+_{jl},\end{aligned}$$

where z_s is the normalized eigenvector corresponding with eigenvalue λ_s and $(C-D-\lambda_s)^+_{jl}$ is the (j,ℓ) element of the Moore-Penrose inversee of $C-D-\lambda_s I$. This directly gives formulas for the first and second derivatives of the loss function.

$$\mathcal{D}_j\phi(D) = -\sum_{s=p+1}^{m}\lambda_s z_{js}^2,$$

$$\mathcal{D}_{j\ell}\phi(D) = \sum_{s=p+1}^{m}\left(z_{\ell s}^2 z_{js}^2 - 2\lambda_s z_{js} z_{ls}(C-D-\lambda_s I)^+_{j\ell}\right)$$

6. Least Squares on the Data Matrix

6.1. **Loss.** The loss function used in LSFAY is

$$\psi(X,U,A,D) = \frac{1}{2}\mathbf{SSQ}(Y - XA' - UD).$$

We minimize over $X \in \mathbb{R}^{n\times p}, U \in \mathbb{R}^{n\times m}, A \in \mathbb{R}^{m\times p}$ and $D \in \mathbb{R}^{m\times m}$, under the conditions $X'X = I, U'U = I, X'U = 0$.

Our approach may seem to be quite similar to the approach proposed by Paul Horst in his book [Horst, 1965]. Where we differ from Horst is in the additional assumptions that D is diagonal and that U has the same size as the data Y. This puts us solidly in the common factor analysis framework. Horst, on the contrary, only makes the assumption that there is a small number of common and residual factors, and he then finds them by truncating the singular value decomposition. Separating common and unique factors is be done later by using rotation techniques. For Horst factor analysis is just principal component analysis with some additional interpretational tools.

6.2. **Advantages.** LSFAY over LSFAC: independence more realistic, optimal scaling easy (FACTALS), D always positive

6.3. **Projection.** We use the result in Appendix A to define a projected version of the LSFAY loss function.

$$\psi(A,D) = \frac{1}{2}\min_{U,X}\mathbf{SSQ}(Y - XA' - UD) = \\ = \frac{1}{2}\mathbf{SSQ}(Y) + \frac{1}{2}\mathbf{SSQ}(A \mid D) - \sum_{s=1}^{m}\sigma_s(YA \mid YD),$$

where the $\sigma_s(YA \mid YD)$ are the ordered singular values of $(YA \mid YD)$. Note that $(YA \mid YD)$ is $n \times (m+p)$, but its rank is less than or equal to m. Thus at least p of the singular values are zero.

The singular values are the square roots of the ordered eigenvalues λ_s of

$$E \triangleq \begin{pmatrix} A'CA & A'CD \\ DCA & DCD \end{pmatrix}. \tag{3}$$

Thus we can also write

$$\psi(A,D) = \frac{1}{2}\mathbf{tr}(C) + \frac{1}{2}\mathbf{SSQ}(A) + \frac{1}{2}\mathbf{SSQ}(D) - \sum_{s=1}^{m}\sqrt{\lambda_s(E)}.$$

6.4. **Algorithms.**

6.4.1. *Alternating Least Squares.* The first algorithm to minimize our loss function is of the ALS type. It was first proposed by De Leeuw [2004], and has since then been used by Unkel and Trendafilov [2010]; Trendafilov and Unkel [2011].

We start with an initial estimate $A^{(0)}$ and $D^{(0)}$ and then alternate

$$\begin{bmatrix} X^{(k)} & \mid & U^{(k)} \end{bmatrix} \in \mathbf{Procrustus}\begin{bmatrix} YA^{(k)} & \mid & YD^{(k)} \end{bmatrix}, \tag{4a}$$

$$A^{(k+1)} = Y'X^{(k)}, \tag{4b}$$

$$D^{(k+1)} = \mathbf{diag}(Y'U^{(k)}). \tag{4c}$$

The Procrustus transformation of a matrix is defined in terms of its singular value decomposition. If the $n \times m$ matrix T has rank m and singular value decomposition $T = K\Lambda L'$, then we define

Procrustus$(T) \triangleq KL'$. Alternatively, we can use the Moore-Penrose inverse and the matrix symmetric square root, because

$$\textbf{Procrustus}(T) = \sqrt{(TT')^+}T = T\sqrt{(T'T)^+}.$$

If $\textbf{rank}(T) = r < \min(n, m)$ then we define the Procrustus transformation as a (closed) set of matrices. See Appendix A for the details.

It is important that we can use the symmetric square root to construct a version of the algorithm that does not depend on the number of observations n, and that can be applied to examples that are only given as covariance or correlation matrices. We can combine the equations in (4) to

$$A^{(k+1)} = CA^{(k)}\sqrt{(E^{(k)})^+}, \tag{5a}$$

$$D^{(k+1)} = \textbf{diag}(CD^{(k)}\sqrt{(E^{(k)})^+}). \tag{5b}$$

This version of the algorithm no longer uses Y, only C. It can be thought of as an adapted version of Bauer-Rutishauser simultaneous iteration [Rutishauser, 1969].

6.4.2. *Gradient and Newton Methods.* Suppose the eigenvector z_s of E in (3) corresponding with λ_s, is partitioned, by putting the first p elements in v_s and the last m elements in w_s. Then [De Leeuw, 2007]

$$\frac{\partial\sqrt{\lambda_s(E)}}{\partial a_{jr}} = \frac{1}{\sqrt{\lambda_s(E)}}v_{rs}c_j'(Av_s + Dw_s),$$
$$\frac{\partial\sqrt{\lambda_s(E)}}{\partial d_{jj}} = \frac{1}{\sqrt{\lambda_s(E)}}w_{js}c_j'(Av_s + Dw_s),$$

where c_j is column j of C. Collecting terms gives

$$\mathcal{D}_1\sigma(A, D) = A - CA\sqrt{E^+},$$
$$\mathcal{D}_2\sigma(A, D) = D - \textbf{diag}(CD\sqrt{E^+}),$$

which shows that the alternating least squares algorithm (5) can be written as a gradient algorithm with constant step-size

$$A^{(k+1)} = A^{(k)} - \mathcal{D}_1\sigma(A^{(k)}, D^{(k)}),$$
$$D^{(k+1)} = D^{(k)} - \mathcal{D}_2\sigma(A^{(k)}, D^{(k)}).$$

To derive a suitable form of Newton's algorithm we do an extra projection. Let

$$\stackrel{\Delta}{=} \min_{X,U,A}$$

References

T.W. Anderson. Estimating Linear Statistical Relationships. *The Annals of Statistics*, 12(1):1–45, 1984.

T.W. Anderson and H. Rubin. Statistical Inference in Factor Analysis. In J. Neyman, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5*, pages 111–150. University of California Press, 1956.

A.L. Comrey. The Minimum Residual Method of Factor Analysis. *Psychological Records*, 11:15–18, 1962.

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. URL http://preprints.stat.ucla.edu/131/131.ps.gz.

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.

J. De Leeuw. Derivatives of Generalized Eigen Systems with Applications. Preprint Series 528, UCLA Department of Statistics, 2007. URL http://preprints.stat.ucla.edu/528/derivatives.pdf.

J. Etezadi-Amoli and R.P. McDonald. A Second Generation Nonlinear Factor Analysis. *Psychometrika*, 48:315–342, 1983.

H.H. Harman and Y. Fukuda. Resolution of the Heywood Case in the MINRES Solution. *Psychometrika*, 31:563–571, 1966.

H.H. Harman and W.H. Jones. Factor Analysis by Minimizing Residuals (MINRES). *Psychometrika*, 31:351–368, 1966.

J. Hemelrijk. Underlining Random Variables. *Statistica Neerlandica*, 20:1–7, 1966.

P. Horst. *Factor Analysis of Data Matrices*. Holt, Rinehart and Winston, 1965.

K. G. Jöreskog. Some Contributions to Maximum Likelihood Factor Analysis. *Psychometrika*, 32:443–482, 1967.

K.G. Jöreskog. On the Statistical Treatment of Residuals in Factor Analysis. *Psychometrika*, 27(4):335–354, 1962.

D.N. Lawley. The Estimation of Factor Loadings by the Method of Maximum Likelihood. *Proceedings of the Royal Society of Edinburgh*, 60:64–82, 1939.

D.N. Lawley. Further Investigations in Factor Estimation. *Proceedings of the Royal Society of Edinburgh*, 61:176–185, 1942.

R.P. McDonald. The Simultaneous Estimation of Factor Loadings and Scores. *British Journal of Mathematical and Statistical Psychology*, 32:212–228, 1979.

H. Rutishauser. Computational Aspects of F.L. Bauer's Simultaneous Iteration Method. *Numerische Mathematik*, 13:4–13, 1969.

G.H. Thomson. Hotelling's Method Modfiied to Give Spearman's *g*. *Journal of Educational Psychology*, 25:366–374, 1934.

N.T. Trendafilov and S. Unkel. Exploratory Factor Analysis of Data Matrices with more Variables than Observations. *Psychometrika*, 2011.

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

P. Whittle. On Principal Components and Least Squares Methods in Factor Analysis. *Skandinavisk Aktuarietidskrift*, 35:223–239, 1952.

G. Young. Maximum Likelihood Estimation and Factor Analysis of Data Matrices. *Psychometrika*, 6(1):49–53, 1940.

F.E. Zegers and J.M.F. Ten Berge. A Fast and Simple Computational Methods of Minimum Residual Factor Analysis. *Multivariate Behavioral Research*, 18:331–340, 1983.

Appendix A. Augmented Procrustus

Suppose X is an $n \times m$ matrix of rank r. Consider the problem of maximizing $\mathbf{tr}\ U'X$ over the $n \times m$ matrices U satisfying $U'U = I$. 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, we use the singular value decomposition

$$X = \begin{bmatrix} K_1 & K_0 \\ {\scriptstyle n\times r} & {\scriptstyle n\times(n-r)} \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ {\scriptstyle r\times r} & {\scriptstyle r\times(m-r)} \\ 0 & 0 \\ {\scriptstyle (n-r)\times r} & {\scriptstyle (n-r)\times(m-r)} \end{bmatrix} \begin{bmatrix} L_1' \\ {\scriptstyle r\times m} \\ L_0' \\ {\scriptstyle (m-r)\times m} \end{bmatrix}.$$

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

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

Now

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

If we write U in the form

$$U = \begin{bmatrix} K_1 & K_0 \\ {\scriptstyle n\times r} & {\scriptstyle n\times(n-r)} \end{bmatrix} \begin{bmatrix} U_1 \\ {\scriptstyle r\times m} \\ U_0 \\ {\scriptstyle (n-r)\times m} \end{bmatrix}$$

then $X = UM$ can be simplified to

$$\begin{aligned} U_1L_1 &= I, \\ U_0L_1 &= 0, \end{aligned}$$

with in addition, of course, $U_1'U_1 + U_0'U_0 = I$. It follows that $U_1 = L_1'$ and

$$\underset{(n-r)\times m}{U_0} = \underset{(n-r)\times(m-r)}{V} \underset{(m-r)\times m}{L_0'},$$

with $V'V = I$. Thus $U = K_1L_1' + K_0VL_0'$. □

Appendix B. The Fundamental Theorem of Factor Analysis

There is a closely related theorem 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.

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

Proof. From $X'X = AA'$ 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 X and A. Observe that the left singular vectors of A are the right singular vectors of X.

Now we still have to solve $X = UA'$. Write

$$U = \begin{bmatrix} \underset{n\times r}{K_1} & \underset{n\times(n-r)}{K_0} \end{bmatrix} \begin{bmatrix} \underset{r\times p}{U_1} \\ \underset{(n-r)\times p}{U_0} \end{bmatrix}.$$

Then $X = UA'$ simplifies to

$$I = U_1V_1,$$

$$0 = U_0V_1,$$

with in addition, of course, $U_1'U_1 + U_0'U_0 = I$. It follows that $U_1 = V_1'$ and

$$\underset{(n-r)\times p}{U_0} = \underset{(n-r)\times(p-r)}{W} \; \underset{(p-r)\times p}{V_0'},$$

with $W'W = I$. Thus $U = K_1V_1' + K_0WV_0'$. □

Appendix C. Code

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

E-mail address, Jia Chen: joshchen@stat.ucla.edu

URL, Jia Chen: http://directory.stat.ucla.edu/jia-chen