

Chapter 7

Least Squares Optimal Scaling of Partially Observed Linear Systems

Jan De Leeuw

*University of California at Los Angeles, Department of Statistics.
Los Angeles, USA*

7.1 Introduction

7.1.1 Problem We compute approximate solutions to homogeneous linear systems of the form $AB = 0$. Here B is an $m \times p$ matrix, and A is an $n \times m$ matrix. Columns of A correspond with *variables*, rows with *observations*. Columns of B correspond with *equations* connecting the variables, elements of B are *coefficients*.

Both A and B are, in general, partially known and our job is to impute the unknown elements. Unknown information on the variable side can be the usual missing data, it can be in the form of unobserved (latent) variables, or it can be in the form of allowing for transformations of the variables.

Our fit criterion will be least squares. This is different from the usual procedure, which embeds the data in a statistical model and then applies some form of maximum likelihood. In particular, in the classical case, it is assumed that the rows of A are replications of a random vector \underline{a} , which satisfies $\underline{a}'B = 0$. We then transform the system to $B'\Sigma B = 0$, where $\Sigma = E(\underline{a}\underline{a}')$, and we proceed to estimate Σ under these constraints, and maybe other constraints as well (often assuming multivariate normality of \underline{a} and often integrating out the parts corresponding with missing information).

This may be fine in some situations, although in non-normal situations it is unclear why one should focus on the covariances. But in other situations the repeated independent trials assumption does not make sense, using random variables is not natural, assuming multivariate normality is ludicrous, and the framework leads to very complicated or even impossible estimation problems. We suggest a more direct approach, formulated directly in terms of the observations, and using least squares to solve matrix approximation problems.

7.1.2 Constraints As indicated above, we will look at the situation in which some of the coefficients and some of the variables are completely known, some are completely unknown, and some are partially known.

As far as the coefficients are concerned, we are mainly thinking of setting some of them equal to known constants, typically zero or one. But in principle the computations below can be easily generalized to put bound constraints on the individual coefficients, requiring them to be in a known interval, and even to equality or inequality constraints linking different coefficients.

As far as the variables are concerned, we will mostly be interested in constraints of the form $a_j \in \mathcal{K}_j \cap \mathcal{S}$, with \mathcal{K}_j a cone and with \mathcal{S} the unit sphere in \mathbf{R}^n . Cones can be used to allow for monotone transformations, spline subspaces, or monotone splines. For variables we will also explicitly discuss more complicated linking constraints, in particular we often require orthogonality of different variables. A block of variables that is required to be mutually orthogonal is called an *orthoblock*.

A key component of our approach is that we allow for variables that are completely unknown, i.e. their quantifications can be anywhere in \mathbf{R}^n . Of course, we need to have some prior knowledge in order to prevent perfect but trivial solutions to our system of equations. These unknown or latent variables will usually be linked by orthogonality constraints with other latent and observed (known or partially known) variables. Once again, we explicitly take the point of view that latent variables are a (rather extreme) form of missing data, and that the missing values can be incorporated directly in the estimation process.

7.1.3 Scope Clearly this class of bilinear systems, with the corresponding constraints, is sufficiently general to fit the linear models in LISREL, EQS, CALIS, AMOS and so on. It is also general enough to fit the models in Gifi's form of nonlinear multivariate analysis [Gifi, 1990, Michailidis and De Leeuw, 1999, Meulman and Heiser, 1999], i.e. it can be used to fit HOMALS, PRINCALS, OVERALS and so on. This illustrates that classical structural (or simultaneous) equation techniques and Gifi's nonlinear multivariate analysis techniques can be captured in a single framework (and can be fit with a single general algorithm).

7.1.4 Equivalent Systems In most cases we can rewrite the system, still in the required linear form, in such a way that the rewritten systems has the same solutions set as the original system. This does not mean, however, that the least squares solutions to the rewritten and the original system are the same. We shall see some examples of equivalent system of linear equations below. A classical example is the reduced form of simultaneous equation systems.

7.1.5 Generalizations Most of the developments below go through without modifications if we use weighted least squares, with known weights, instead of unweighted least squares. Similarly, variables can be elements of an arbitrary inner product space, instead of vectors in \mathbf{R}^n . Thus we can use sample or population distributions to compute the inner products of our variables, to study what happens in the population case, or to find out what sampling distributions of our statistics are.

We will not elaborate further on these easy generalizations, but it is good to remember that they are available at little cost.

7.2 Loss Function and Algorithm

7.2.1 Loss Function The problem studied in this paper is to minimize the *least squares loss function*

$$\sigma(A, B) = \text{tr } B'RB.$$

Here $R = A'A$ is the $m \times m$ correlation matrix of the variables.

Thus, we will look at the general class of problems in which we minimize $\sigma(A, B)$ over the A satisfying cone and orthogonality restrictions and over the B with some of the elements known.

7.2.2 Algorithm In De Leeuw [1990], we studied the problem of minimizing any real-valued function of the form $\phi(R)$, with $R = A'A$, over $a_j \in \mathcal{K}_j \cap S$. A majorization algorithm [De Leeuw, 1994, Heiser, 1995, Lange et al., 2000] was developed there for the case in which $\phi(R)$ is concave in R . Further developments of this approach are in De Leeuw [1988, 1993] and in De Leeuw et al. [1999].

In the problem studied in this paper we can define

$$\phi(R) = \min_{B \in \mathcal{B}} \text{tr } B'RB,$$

where \mathcal{B} are the matrices in $\mathbf{R}^{m \times p}$ satisfying the constraints on B . The loss function ϕ is the pointwise minimum of a family of functions linear, and thus concave, in R . Because the pointwise minimum of a family of concave functions is also concave, it follows that ϕ is concave in R . We have seen that the

majorization algorithm applies to concave functions, and thus ϕ can be minimized by the majorization algorithm. In that sense the class of techniques discussed here is an important (and so far unexplored) special case of our previous work. The main generalization is that in the current framework we do not only have the cone constraints on individual variables, but we also allow for orthoblocks of variables.

We give a brief outline of the algorithm in the general case. Algorithms for specific systems will be discussed in more detail below. For the majorization algorithm we need the subgradient of the loss function. For ϕ the subgradient $\partial\phi(\tilde{R})$ is the convex hull of the matrices $\tilde{B}\tilde{B}'$, where \tilde{B} is any matrix minimizing $\text{tr } B'\tilde{R}B$ over $B \in \mathcal{B}$. This follows from the formula for the subgradient of a pointwise minimum of concave functions [Hiriart-Urruty and Lemarechal, 1993, Theorem 4.4.2]. Observe that computing \tilde{B} is a quadratic minimization problem. This is why it is easy to incorporate bound and linear inequality constraints on the coefficients, because we will remain in the convex quadratic programming framework.

The subgradient inequality tells us that $\phi(R) \leq \text{tr } R\tilde{G}$, for any $\tilde{G} \in \partial\phi(\tilde{R})$. The algorithm selects a subgradient \tilde{G} , and minimizes the majorization function $\text{tr } R\tilde{G}$ by cycling over all variables (or orthoblocks of variables). This gives us a new R^+ , and we have

$$\phi(R^+) \leq \text{tr } R^+\tilde{G} \leq \text{tr } \tilde{R}\tilde{G} = \phi(\tilde{R}).$$

In majorization theory this is called the *sandwich inequality*, and it forces convergence of the sequence of loss function values (and through that, using Zangwill [1969], convergence of the sequence of solutions). Observe that convergence will be to a stationary point of the algorithm, which is a point satisfying the stationary equations of the minimization problem. This is likely to be a local minimum, because it is a minimum with respect to each block of variables, but there is no guarantee that we find the global minimum.

Observe that if we update a variable, or an orthoblock, we have to recompute the subgradient at the new point, i.e. we have to recompute the optimal B for current A , before we proceed to the next variable or block.

7.2.3 Subproblems for Variables Suppose we are optimizing over the single variable $a_1 \in \mathcal{K}_1 \cap \mathcal{S}$, keeping A_2 , the rest of A , fixed at its current values. Partition A and \tilde{G} in the obvious way. Then

$$\text{tr } R\tilde{G} = \tilde{g}_{11} + 2a_1' A_2 \tilde{g}_1 + \text{tr } A_2' A_2 G_{22}.$$

Only the second term depends on a_1 . De Leeuw [1990] shows that the new optimal a_1 can be found by projecting $\tilde{h} = A_2 \tilde{g}_1$ on the cone \mathcal{K}_1 and then normalizing the projection to unit length.

If we are optimizing over an orthoblock of variables A_1 , then by the same reasoning

$$\text{tr } R\tilde{G} = \text{tr } \tilde{G}_{11} + 2\text{tr } A_1' A_2 \tilde{G}_{12} + \text{tr } A_2' A_2 \tilde{G}_{22}.$$

Thus the optimal A_1 is found by solving the orthogonal Procrustus problem for $\tilde{H} = A_2 \tilde{G}_{12}$. If \tilde{H} has full column rank, and $\tilde{H} = K\Lambda L'$ is its singular value decomposition, then the optimal A_1 is KL' . In the Appendix we generalize this to singular \tilde{H} , a generalization we will need in our factor analysis algorithm.

We discuss two more subproblems we can come across in implementing the general algorithm. First, we want to find $a_1 \in \mathcal{K}_1 \cap \mathcal{S}$ orthogonal to the block A_2 , but not necessarily to the block A_3 . Clearly

$$\text{tr } R\tilde{G} = a_1' A_2 \tilde{g}_2 + a_1' A_3 \tilde{g}_3 + \text{terms not dependent on } a_1$$

To find the optimal a_1 we need to project

$$\tilde{h} = (I - A_2(A_2' A_2)^{-1} A_2') A_3 \tilde{g}_3$$

on \mathcal{K}_1 , and normalize. We do not assume here that A_2 is an orthoblock.

The second subproblem asks for an orthoblock A_1 which is orthogonal to block A_2 , but not necessarily to block A_3 . With the same reasoning as in the previous subproblem, we now have to apply Procrustus to

$$\tilde{H} = (I - A_2(A_2' A_2)^{-1} A_2') A_3 \tilde{G}_{31}.$$

7.2.4 Algorithm Flow The flow of the algorithm is to partition A into blocks that are either orthoblocks or single variables. We optimize the variables over the first block, optimize over B , optimize variables in the second block, optimize over B , and so on. It is not necessary to go through the blocks of variables in order, in fact we can change the order or use “free-steering” methods in which we cycle through the blocks in random order [De Leeuw and Michailides, 1999], but in general it is necessary to update B every time we update a block of variables. This may be inefficient if computing B is much more complicated than updating blocks of variables. There are variations of the algorithm possible based on using majorization a second time, this time to bound.

It is sometimes also possible, and even advisable, to decompose B into a number of blocks and apply block relaxation to optimize over B . Thus many variations are possible with our general “algorithm model”.

7.3 Examples

Let us look at some “classical ” systems to see how they fit into our framework and to see what loss functions and algorithms we can expect to discover (or rediscover). In discussing regression and various other special cases of our general framework, we use additional notation which is natural for the problem at hand and consequently obvious.

7.3.1 Regression It has always been somewhat ambiguous what regression analysis wants of the residuals. On the one hand, we want the residuals to be “small”, on the other hand we want them to be “unsystematic”, which means unrelated to the predictors (and may mean more). Clearly being small does not imply being unsystematic, and vice versa.

This means that we can distinguish two regression problems in our framework. The first system is

$$\begin{bmatrix} y & | & X & | & e \end{bmatrix} \begin{bmatrix} 1 \\ -\beta \\ -\sigma \end{bmatrix} = 0$$

or $y = X\beta + \sigma e$, where we require $e'X = 0$.

Minimizing loss gives the usual regression statistics (regression coefficients, residuals, residual sum of squares), which actually make the minimum loss equal to zero. There is no room for improvement of the loss in this case by using optimal cone transformations of the variables. If we have missing information in X and/or y , we can just fill it in arbitrarily, and we will still have zero loss. This simply repeats the obvious: if we project a vector orthogonally on a subspace then the residuals are orthogonal to the subspace, and thus the decomposition we seek is always possible.

Now let us “ignore errors”, i.e. remove e from the system. We concentrate on making the residuals small. Thus

$$\begin{bmatrix} y & | & X \end{bmatrix} \begin{bmatrix} 1 \\ \beta \end{bmatrix} = 0.$$

The solution for β , for given X and y , is again the usual vector of regression coefficients, but now, of course, the minimum loss is non-zero, and we can use transformations, quantifications, or imputations to attain a better fit. In other words, we can require y and/or the columns of X to vary over cones.

This leads to techniques implemented, for example, in ACE [Breiman and Friedman, 1985], TRANSREG [SAS, 1992], or CATREG [Meulman and

Heiser, 1999]. Our majorization algorithm is identical to the algorithm used in these programs.

Clearly we can extend this to multivariate regression. Now

$$\begin{bmatrix} Y & | & X & | & E \end{bmatrix} \begin{bmatrix} I \\ -\Gamma \\ -\Sigma \end{bmatrix} = 0,$$

and we typically require $E'X = 0$. This is entirely tautological again, unless we impose constraints such as $E'E = I$ and Σ is diagonal, or impose constraints on the coefficients in Γ .

For path analysis we have

$$\begin{bmatrix} Y & | & X & | & E \end{bmatrix} \begin{bmatrix} I - \Theta \\ -\Gamma \\ -\Sigma \end{bmatrix} = 0,$$

with Θ upper triangular, $E'X = 0$, $E'E = I$ and Σ diagonal. Again this is a saturated system which can always be solved perfectly, and we need additional constraints to make imputing missing information interesting. Observe we can rewrite this system in the reduced form

$$\begin{bmatrix} Y & | & X & | & E \end{bmatrix} \begin{bmatrix} I \\ -\Gamma(I - \Theta)^{-1} \\ -\Sigma(I - \Theta)^{-1} \end{bmatrix} = 0$$

which introduces some awkward nonlinear constraints on the parameters.

For both multivariate regression and path analysis, the “ignore errors” versions are also quite straightforward. We should emphasize here that “ignoring errors” is somewhat of a misnomer. It is only defined relative to a more complicated system which does have one or more additional orthoblocks of unobserved variables. It can also be misleading to emphasize that blocks such as X and Y are “observed”, while E is “unobserved”. In general, all blocks of variables have both known and unknown elements, and subsets of both X and Y can be “unobserved” as well.

7.3.2 Factor Analysis Again, the notation is adapted to fit the problem. As in regression, we can take two approaches to residuals. We can make them “as

small as possible”, and we can make them “as unsystematic as possible”. First we explore making residuals unsystematic.

The system, in matrix form, is

$$\begin{bmatrix} Y & | & U & | & E \end{bmatrix} \begin{bmatrix} I \\ -\Gamma \\ -\Delta \end{bmatrix} = 0,$$

or $Y = U\Gamma + E\Delta$, where $U'E = 0$, $U'U = I$, and $E'E = I$, and where Δ is diagonal. U is $n \times f$, where f is the number of common factors.

Minimizing the least squares loss function is a form of factor analysis, but the loss function is not the familiar one. In “classical” least squares factor analysis, as described in Young [1940], Whittle [1952] and Jöreskog [1962], the unique factors E are not parameters in the loss function. Instead the unique variances are used to weight the residuals of each observed variable (as suggested by maximum likelihood).

Our loss function, which is

$$\sigma(Y, U, E, \Gamma, \Delta) = \|Y - U\Gamma - E\Delta\|^2$$

can be better understood by defining the $(m + f) \times m$ matrix

$$T = \begin{bmatrix} \Gamma \\ \Delta \end{bmatrix},$$

and by observing that

$$\begin{aligned} \min_{U, E} \sigma(Y, U, E, \Gamma, \Delta) &= \sum_{j=1}^m [\lambda_j^2(Y) + \lambda_j^2(T) - 2\lambda_j(YT')] \quad (1) \\ &\geq \sum_{j=1}^m (\lambda_j(Y) - \lambda_j(T))^2. \end{aligned}$$

Here the $\lambda_j(\bullet)$ are the ordered singular values of their matrix argument. See the Appendix, and for the inequality use the theorem on the singular values of a matrix product [Horn and Johnson, 1991 Theorem 3.3.14]. Thus we see that our loss function is just one (orthogonally invariant) way to measure how similar $Y'Y$ is to $T'T = \Gamma'\Gamma + \Delta^2$.

There is another conceptual problem that has prevented the straightforward use of our least squares loss function, although it seems such a natural choice in the alternating least squares framework of Takane, Young and De Leeuw [Young, 1981] or the nonlinear multivariate analysis framework of Gifi [1990].

Factor score indeterminacy implies that generally the solution of U and E will not be unique, even for fixed loadings Γ and uniquenesses Δ . For that reason, Takane et al. [1979] go out of their way to redefine the least squares loss function in terms of the correlation matrix of the observed variables. This makes the algorithm quite complicated, with as a consequence several errors, and several proposed corrections to fix it [Mooijart, 1984, Nevells, 1989, Kiers et al., 1993].

But non-uniqueness of factor scores is not really a problem from the algorithmic point of view. Because all solutions to the augmented Procrustus problem are in a closed set in matrix space, we still have convergence of our majorization algorithm [Zangwill, 1969], and accumulation points of the sequences we generate will still be stationary points.

Implementation is quite straightforward. In fact, we can think of two obvious ways to implement the part of the algorithm that updates the common and unique factor scores. In the first we update the U and E blocks successively, in the second we update them simultaneously by treating them as a single orthonormal block. In the first case we alternate solving the Procrustus problems for $(Y - E\Delta)\Gamma'$ and $(Y - U\Gamma)\Delta$, in the second case we solve the augmented Procrustus problem for YT' . Updating loadings and uniquenesses is simple, because the optimal Γ is simply $U'Y$ and the optimal Δ is $\text{diag}(E'Y)$. It is also easy, in this algorithm, to incorporate the types of constraints typical for confirmatory factor analysis.

We can also apply the “ignoring errors” strategy to factor analysis. Remove E and we have

$$\begin{bmatrix} Y & | & U \end{bmatrix} \begin{bmatrix} I \\ -A \end{bmatrix} = 0,$$

or $Y = UA$, where $U'U = I$. Minimizing the least squares loss function is principal component analysis with optimal scaling, as implemented in PRINCALS [Gifi, 1990], CATCPA [Meulman and Heiser, 1999], PRINQUAL [SAS, 1992], or MDRACE [Koyak, 1987]. From our viewpoint the two techniques are quite close, the difference is incorporating the residuals in the loss function and requiring them to be an orthoblock, or minimizing the residuals and maybe looking at them after the analysis is done. It is of some interest that requiring the matrix of unique variances Δ to be scalar in our least squares factor analysis algorithm does *not* lead to principal component analysis.

7.4 MIMIC

Regression and factor analysis are relatively simple systems. The first step towards making life more complicated is the MIMIC system [Jöreskog and

Goldberger, 1975]. It is

$$[X \mid Y \mid U \mid E \mid F] \begin{bmatrix} 0 & -\Theta \\ I & 0 \\ -\Gamma & I \\ -\Delta & 0 \\ 0 & -\Omega \end{bmatrix} = 0,$$

or

$$\begin{aligned} Y &= U\Gamma + E\Delta, \\ U &= X\Theta + F\Omega. \end{aligned}$$

There are presumably various additional assumptions, which make $(E|F)$ or even $(U|E|F)$ into an orthoblock, and which make Δ and Ω diagonal. Many variations are possible, in particular the familiar variations which “ignore the errors” E and/or F . We will not go into implementation details here, but just point out the flexibility and the painless incorporation of transformations of the observed variables or imputations of the missing data.

Of course systems of this form, and more complicated ones along the usual structural equations lines, can have identification problems. This does not prevent the algorithm from doing its work properly, but it still remains a valid topic to be studied, at least if one is interesting in interpreting and using the computed coefficients.

Also, as we mentioned above, systems can be manipulated algebraically and rewritten as equivalent but different systems. Suppose, for example, that $F = 0$, for simplicity. Then we substitute $U = X\Theta$, and rewrite the system as

$$[X \mid Y \mid E] \begin{bmatrix} -\Xi \\ I \\ -\Delta \end{bmatrix} = 0,$$

where $\Xi = \Theta\Gamma$. This is now reduced rank regression, or redundancy analysis [Reinsel and Velu, 1998]. Our majorization algorithm can be used to minimize $\|Y - X\Xi - E\Delta\|^2$ with $X'E = 0$ and $E'E = I$. In this case optimizing over regression coefficients must of course take the rank restrictions on Ξ into account. Again the loss function is different from previous ones, because it explicitly treats residuals as additional parameters in the matrix decomposition.

It is clear that the approach we have illustrated here on the MIMIC system can be extended to even more complicated systems studied in LISREL and related techniques. Nothing essential really changes, although implementation details and lists of possible options and variations may become quite messy

7.5 Nonlinear Multivariate Analysis

Let us now switch to systems studied in nonlinear multivariate analysis [Gifi, 1990]. The system for a general form of homogeneity analysis is

$$\left[\begin{array}{c|c|c|c|c} X & Q_1 & \cdots & Q_m & \end{array} \right] \begin{bmatrix} I & I & I & \cdots & I \\ -\Gamma_1 & 0 & 0 & \cdots & 0 \\ 0 & -\Gamma_2 & 0 & \cdots & 0 \\ 0 & 0 & -\Gamma_3 & \cdots & 0 \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & 0 & -\Gamma_m \end{bmatrix} = 0,$$

or $X = Q_j \Gamma_j$ for $j = 1, \dots, m$. The $n \times s$ matrix X and each of the $n \times r_j$ matrices Q_j is an orthoblock. Moreover all columns of Q_j are assumed to be in the same subspace \mathcal{L}_j of \mathbb{R}^n . This is because the orthoblock Q_j corresponds with a single variable in the original data, which we allow to have multiple quantifications. The rank of the quantification for variable j is r_j , which is usually either one (single quantification) or s (multiple quantification).

By writing the system in this way, we cover both HOMALS (a.k.a. multiple correspondence analysis, in which the Q_j are known and quantifications are multiple) and PRINCALS (a.k.a nonlinear principal component analysis, in which the Q_j are partially known and all quantifications are single). See Bekker and De Leeuw [1988] for further details on this. In fact a slight modification also allows us to include generalized canonical correlation analysis OVERALS, in which each column of the coefficient matrix has more than one non-zero Γ matrix, although each row still only has a single non-zero Γ matrix. Thus there is a partition of the indices $1, \dots, m$ into sets of variables J_1, \dots, J_ℓ such that

$$X = \sum_{j \in J_\nu} Q_j \Gamma_j,$$

for all $\nu = 1, \dots, \ell$. As explained in Gifi [1990], this makes it easy to include canonical analysis and canonical discriminant analysis as special cases.

The actual majorization algorithm from OVERALS is implemented in R by De Leeuw and Ouwehand [2003]. The implementation has one additional feature, to deal with ordinal variables in the case of multiple quantifications. We can require Q_j to be an orthoblock, with all its columns in the subspace L_j , but with, in addition, its first column in a cone $\mathcal{K}_j \cup L_j$. Thus we can require the leading quantification of a variable to be ordinal, while the remaining quantifications are orthogonal to the leading one.

If all variables are single, then homogeneity analysis becomes principal component analysis. Combining this with the ideas from the previous section show how explicit orthoblocks or errors can be introduced into the technique, so that we obtain versions of factor analysis or canonical analysis with orthogonal error.

References

- Bekker, P., & De Leeuw, J. (1988). Relation between variants of nonlinear principal component analysis. In J.L.A. Van Rijkevorsel & J. De Leeuw (Eds.), *Component and correspondence analysis*. Chichester: Wiley.
- Breiman, L., & Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580-619.
- De Leeuw, J. (1988). Multivariate analysis with linearizable regressions. *Psychometrika*, 53, 437-454.
- De Leeuw, J. (1990). Multivariate analysis with optimal scaling. In S. Das Gupta & J. Sethuraman (Eds.), *Progress in multivariate analysis*. Calcutta: Indian Statistical Institute.
- De Leeuw, J. (1993). Some generalizations of correspondence analysis. In C.M. Cuadras & C.R. Rao (Eds.), *Multivariate analysis: Future directions 2*. Amsterdam: North-Holland.
- De Leeuw, J. (1994). Block relaxation methods in statistics. In H.H. Bock, W. Lenski & M.M. Richter (Eds.), *Information systems and data analysis*. Berlin: Springer.
- De Leeuw, J., & Michailidis, G. (1999). *Block relaxation algorithms in statistics*. <http://www.stat.ucla.edu/deleeuw/block.pdf>.
- De Leeuw, J., & Michailidis, G., & Wang, D.Y. (1999). Correspondence analysis techniques. In S. Gosh (Ed.), *Multivariate analysis, design of experiments, and survey sampling*. New York: Marcel Dekker.
- De Leeuw, J. & Ouwehand, A. (2003). *Homogeneity analysis in R* (Technical Report, UCLA Statistics). Los Angeles: UCLA Statistics Department.

- Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.
- Heiser, W. (1995). Convergent computing by iterative majorization: Theory and applications in multidimensional data analysis. In W.J. Krzanowski (Ed.), *Recent advantages in descriptive multivariate analysis*. Oxford: Clarendon.
- Horn, R.A., & Johnson, C.R. (1991). *Topics in matrix analysis*. Cambridge: Cambridge University Press.
- Jöreskog, K.G. (1962). On the statistical treatment of residuals in factor analysis. *Psychometrika*, 27, 335-354.
- Jöreskog, K.G., & Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70, 631-639.
- Kiers, H.A.L., Takane, Y., & Mooijaart, A. (1993). A monotonically convergent algorithm for FACTALS. *Psychometrika*, 58, 567-574.
- Koyak, R. (1987). On measuring internal dependence in a set of random variables. *Annals of Statistics*, 15, 1215-1228.
- Lange, K., Hunter, D., & Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics*, 9, 1-59.
- Meulman, J.J. & Heiser, W.J. (1999). *SPSS Categories 10.0*. Chicago: SPSS.
- Michailidis, G., & De Leeuw, J. (1999). The Gifi system for descriptive multivariate analysis. *Statistical Science*, 13, 307-336.
- Mooijaart, A. (1984). The nonconvergence of FACTALS: A nonmetric common factor analysis. *Psychometrika*, 49, 143-145.
- Nevels, K. (1989). An improved solution for FACTALS: a nonmetric common factor analysis. *Psychometrika*, 54, 339-343.
- Reinsel, G.C., & Velu, R.P. (1998). *Multivariate reduced-rank regression* (Lecture notes in statistics No. 136). Berlin: Springer.
- SAS. (1992). *SAS/STAT software: Changes and enhancements* (Technical report P-229). Cary NC: SAS Institute.
- Takane, Y., Young, F.W., & De Leeuw, J. (1979). Nonmetric common factor analysis: An alternating least squares method with optimal scaling. *Behaviormetrika*, 6, 45-56.
- Whittle, P. On principle components and least squares methods in factor analysis. *Skandinavisk Aktuarietidskrift*, 35, 223-239.
- Young, F.W. (1981). Quantitative analysis of qualitative data. *Psychometrika*, 46, 357-388.
- Young, G. (1940). Maximum likelihood estimation and factor analysis. *Psychometrika*, 6, 49-53.
- Zangwill, W.I. (1969). *Nonlinear programming: A unified approach*. Englewood-Cliffs NJ: Prentice-Hall.

Appendix.

Augmented Procrustus

Suppose X is an $n \times m$ matrix of rank r . Consider the problem of maximizing $\text{tr } U'X$ 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

$$X = \begin{bmatrix} K_1 & K_0 \\ n \times r & n \times (n-r) \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ r \times r & r \times (m-r) \\ 0 & 0 \\ (n-r) \times r & (n-r) \times (m-r) \end{bmatrix} \begin{bmatrix} L'_1 \\ r \times m \\ L'_0 \\ (m-r) \times m \end{bmatrix}.$$

Theorem 1 *The maximum of $\text{tr } U'X$ over $n \times m$ matrices U satisfying $U'U = I$ is $\text{tr } \Lambda$, and it is attained for any U of the form $U = K_1 L'_1 + K_0 V L'_0$, 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 $\text{tr } U'X = \pm \text{tr } \Lambda$. The negative sign corresponds with the minimum, the positive sign with the maximum.

Now

$$M = \begin{bmatrix} L_1 & L_0 \\ m \times r & m \times (m-r) \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ r \times r & r \times (m-r) \\ 0 & 0 \\ (m-r) \times r & (m-r) \times (m-r) \end{bmatrix} \begin{bmatrix} L'_1 \\ r \times m \\ L'_0 \\ (m-r) \times m \end{bmatrix}.$$

If we write U in the form

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

then $X = UM$ can be simplified to

$$\begin{aligned} U_1 L_1 &= I, \\ U_0 L_1 &= 0, \end{aligned}$$

with in addition, of course, $U'_1 U_1 + U'_0 U_0 = I$. It follows that $U_1 = L'_1$ and

$$\begin{matrix} U_0 & & V & & L'_0 \\ (n-r) \times m & = & (n-r) \times (m-r) & & (m-r) \times m \end{matrix},$$

with $V'V = I$. Thus $U = K_1 L'_1 + K_0 V L'_0$.