Contributions of Dutch Matricians to Computational Statistics

Kohei Adachi, Osaka University, Japan
Patrick J. F. Groenen, Erasmus University Rotterdam, the Netherlands

1. Introduction

We introduce the term matricians for those researchers that have taken matrix-intensive approaches to form statistical computing procedures. The Netherlands has a number of matricians, which is exemplified by Magnus and Neudecker’s (1994) famous book on matrix differential calculus. This paper focuses on the contributions of the Dutch matricians rooted in psychometrics, whose works have been effectively contributed to computational statistics.

2. Maximizing Trace Functions

For a given $n \times p$ matrix $X$, some multivariate analysis procedures are formulated as problems of maximizing trace functions of parameter matrices subject to orthonormality constraints. A simple example is

$$
\max_w f(W) = \text{tr}W'XWX \text{ s.t. } W'W = I_m.
$$

(1)

A classic solution is to use the Lagrangian function $\phi(W) = f(W) - \text{tr}L(W'W - I_m)$: the system of equations $\partial\phi(W)/\partial W = O$ and $W'W = I_m$ gives the necessary condition for the solution (e.g., Magnus & Neudecker, 1994). But, these conditions are not sufficient and, thus, further efforts are needed to find the solution.

For such problems, Jos Ten Berge elaborated on a more efficient approach where inequalities take central roles: attainable upper bounds. Its logic is as follows: first, the inequality $f(W) \leq UL$ (upper limit) is presented, then $f(W) = UL$ for $W = \hat{W}$ is shown for proving that $\hat{W}$ is the solution. The upper limit of some trace functions are given by Ten Berge’s (1993) theorem:

If $M$ is a $p \times p$ sub-orthonormal matrix with $\text{rank}(M) = m \leq p$ and $D = \text{diag}\{d_1, \ldots, d_p\}$ is a $p \times p$ diagonal matrix with $d_1 \geq \ldots \geq d_p \geq 0$, then

$$
f(M) = \text{tr}MD \leq \text{tr}D_m = d_1 + \ldots + d_m \leq \text{tr}D
$$

(2)

with $D_m = \text{diag}\{d_1, \ldots, d_m\}$ the first $m \times m$ diagonal block of $D$. Here, a sub-orthonormal matrix refers to the one which can be completed to be an orthonormal matrix by appending rows, columns, or both, or is orthonormal, with the product of such matrices known to be sub-orthonormal.

Using the theorem, the solution of (1) is easily found as follows: using the eigenvalue decomposition $X'X = LA'L'$, $L/L = I$, and $L = \text{diag}\{\lambda_1 \geq \ldots \geq \lambda_r\}$ with $r = \text{rank}(X'X)$, the function $f(W)$ in (1) is rewritten as $f(W) = \text{tr}W'LA'^2L'W = \text{tr}W'WW'LA^2$. Since $W'WW'$ is sub-orthonormal and $\text{rank}(W'WW') \leq r$, we have $f(W) = \text{tr}W'XWX \leq \text{tr}A_m^2$ with $A_m = \text{diag}\{\lambda_1, \ldots, \lambda_m\}$. Here, the upper bound is attained for $W = L_m$ with $L_m$ contains the first $m$ columns of $L$. In similar manner, we can use Ten Berge’s theorem for the problems $\max_A \text{tr}A'X$ s.t. $A'A = I_p$ and $\max_{A,B} \text{tr}A'XB$ s.t. $A'A = B'B = I$.

3. Majorization

Some procedures do not have explicit solutions as in the last section. For such cases, iterative algorithms are used, which are classified into some types. One of them is a majorization method, which has been introduced by De Leeuw (1977) in multivariate data analysis, in particular, for multidimensional scaling (MDS, see, e.g., Borg & Groenen, 2005). It can be formulated as minimizing

$$
\min_A f(A) = \sum_{ij}(x_{ij} - ||a_i - a_j||)^2 = c - 2\sum_{ij}x_{ij}||a_i - a_j|| + n\text{tr}A'JA
$$

(3)
for a nonnegative and symmetric matrix $X$ with $n = p$, and $J = I_n - n^{-1}1_n1_n'$. Let $B$ be the current estimate. Then, the majorization method uses an auxiliary function $g(A, B)$ that satisfies $f(B) = g(B, B)$, $f(A) \leq g(A, B)$, and $g(A, B)$ being easier to handle than $f(A)$. Let $A_{\text{new}}$ be an update such that $g(A_{\text{new}}, A) \leq g(A, A)$. These inequalities lead to

$$f(A_{\text{new}}) \leq g(A_{\text{new}}, B) \leq g(B, B) = f(B).$$  

(4)

In the next iteration, $B$ is set to $A_{\text{new}}$ so that the function values decrease monotonically. The auxiliary function corresponding to (3) can be defined as

$$g(A, B) = c - 2\text{tr}A'Q(B)B + n\text{tr}A'JA$$  

(5)

with $Q(B) = \Sigma_{i,j}q_{ij}(B)(e_i - e_i)(e_j - e_j)'$. Here, $e_i$ is the $i$th column of $I_p$, and $q_{ij}(B) = x_{ij}/\|b_i - b_j\|$ if $b_i \neq b_j$; $x_{ij}(B) = 0$ otherwise. For a given $B$, (5) is easily minimized over $B$. The algorithm has been elaborated by Groenen (1993) and Heiser (1991).

Kiers (2002) has presented majorization algorithms for minimizing

$$\min_A f(A) = \text{tr}XA + \sum_k \text{tr}Y_kAZ_kA'$$  

(6)

for any choice of $X$, $Y_k$, and $Z_k (k = 1, \ldots, K)$ possibly subject to constraints on $A$. The algorithms are useful in that (6) include a variety of loss functions as special cases. For example, Adachi (2014) in Japan has presented an algorithm for generalized least squares factor analysis using Kiers’ (2002) method.

In a number of cases, auxiliary functions are found using the Cauchy-Schwarz inequality. The EM algorithm can also be included as a family member of majorization.

4. Homogeneity Analysis

Gifi (1990) is a famous book on nonlinear multivariate analysis with an emphasis on homogeneity analysis. Gifi is not the name of a single person, but refers to a group of Dutch matricians, amongst them De Leeuw, Heiser, Meulman, Van der Geer, and others. Homogeneity analysis can be formulated as

$$\min_{F,W} \sum_k \|F - X_k W_k\|^2 \text{ s.t. } n^{-1}F'F = I_m.$$  

(7)

with $X_k n \times p_k$ data matrix, $F$ being $n \times m$ and $W = [W_1', \ldots, W_K']' (p \times m)$.

A special feature of (7) is that a parameter matrix $F$ is fitted to $X_k W_k$, that is, a part that is also dependen on the unknown parameters $W_k$. It differs from the usual least-squares functions that model data directly by a function of unknown parameters. However, (7) allows us to easily comprehend the implication of analyzing multi-set data ($X_1, \ldots, X_K$): $F$ serves as a matrix that summarizes the similar part of the $X_1 W_1, \ldots, X_K W_K$.

Homogeneity analysis can be regarded as a generalization of canonical correlation analysis and also as optimal scoring of categories when $X_k$ is a binary indicator matrix of observations by categories.

5. Three-mode Component Analysis

For three-way array $X = \{x_{ijr}; i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K\}$, the three-mode principal component analysis (3MPCA) called Tucker3 or Tucker decomposition is formulated as

$$\min_{A,B,C,G} \|X - \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R (a_{p} \circ b_{q} \circ c_{r})g_{pqr}\|^2.$$  

(9)

Here, $A = (a_p) = [a_1, \ldots, a_p] (I \times P)$, $B = (b_q) = [b_1, \ldots, b_Q] (J \times Q)$, $C = (c_r) = [c_1, \ldots, c_R] (K \times R)$, $\cdot \circ \cdot \circ \cdot \circ$ denotes the tensor product with $P \leq I$, $Q \leq J$, and $R \leq K$, and $\|\cdot\|^2$ denotes here the sum of squares of all elements in the three-way array. Tucker3 was originally
proposed by the American Tucker (1966). However, he did not presented an exact solution, but only an approximate one. Later, the alternating least-squares (ALS) algorithms for minimizing (9) directly were developed by Dutch matricians. Such an algorithm was firstly presented by Kroonenberg and De Leeuw (1980). More efficient algorithms were developed by Dutch matricians (e.g., Kiers, Kroonenberg, & Ten Berge, 1992). For describing the ALS algorithms, the tensor expression in (9) is transformed in the matrix form as

\[
\min_{A,B,C,G} \left| \left| X - AG(C \otimes B) \right| \right|^2.
\] (9')

Here, \( X = [X_1, \ldots, X_K] \) and \( G = [G_1, \ldots, G_R] \) with \( X_k = (x_{ijk}) \) and \( G_r = (g_{pq}) \).

Without loss of generality, some of the elements in \( G \) can be set to zero. Such mathematical properties have been explored by several Dutch matricians (e.g., Ten Berge & Kiers, 1999; Tendideiro, Ten Berge & Kiers, 2009). A less restrictive version of Tucker3, called Tucker2, relaxes the model part as \( \sum_p \sum_q (a_p \circ b_q \circ g_{pq}) \) with \( g_{pq} = [g_{pq1}, \ldots, g_{pqK}]' \). Ikemoto and Adachi (2016) in Japan have presented sparse Tucker2 with the cardinality of \( g_{pq} \) constrained. Tucker3 and nonnegative matrix factorization have been integrated in Japanese Riken Brain Science Institute (Cichocki, Zdunek, Phan, & Amari, 2009).

6. Matrix Decomposition Factor Analysis

In the classic formulation of factor analysis (FA), the factors are treated as latent random variables. Recently, a very different formulation of FA has been presented in which common and unique factors are regarded as unknown fixed matrices: FA is formulated as

\[
\min_{Z,B} \left| \left| X - FA' - UB\right| \right|^2 = \left| \left| X - ZB' \right| \right|^2 \text{ s.t. } 1_n'Z = 0_m' \text{ and } n^{-1}Z'Z = I_m ,
\] (10)

with \( Z = [F, U] \), \( B = [A, \Psi] \), and \( \Psi \) being diagonal. To the best of our knowledge, this was first presented by Prof. Henk A. L. Kiers in 2001, which is found in Sočan (2003). Independently, De Leeuw (2004) has described (10). Unkel and Trendafilov (2010) at the Open University in the UK, have reviewed FA formulations with (10). Adachi (2012) presented the algorithm for obtaining the optimal \( B \) only if sample covariance matrix \( S = n^{-1}XX' \). Further, Adachi and Trendafilov (2015) have incorporated the cardinality constraint in \( A \) to produce a sparse FA procedure. Recently, new algorithms for (10) are presented by Stegeman (2016).

References


Leiden, The Netherlands: DSWO.


