Least Squares Optimization in Multivariate Analysis

Jos M.F. ten Berge

University of Groningen

Pdf version of the monograph published by DSWO Press (Leiden, 1993)

This version (2005) is essentially the same as the original one, published in 1993 by DSWO Press (Leiden). In particular, all material has been kept on the same pages. Apart from a few typographic and language corrections, the following non-trivial changes have been made.

- ALS has been deleted four lines below (97), page 51.
- The sentence below (13) on page 8 was deleted.
- The last paragraph of page 51 has been rewritten.
- The last paragraph of page 52 has been changed.
- The line above (100) on page 53 has been changed.
- On line 14 of page 64, *X* has been changed to *A*.
- Page numbers of two references have been added on page 73.
- Question 6 of page 75 and the answer (page 77) have been changed.
- Question 13 of page 76 has been changed.
- Question 34 (and the answer) has been deleted (page 84).
- The answers to questions 36 a and b (page 84) have been rephrased.
- Question 49c of page 87 has been rephrased.

PREFACE

This book arose from a course for senior students in Psychometrics and Sociometrics. It is aimed at students who have been exposed to an introductory course in Matrix Algebra, and are acquainted with concepts like the rank of a matrix, linear (in)dependence of sets of vectors, matrix inversion, matrix partitioning, orthogonality, and eigenvectors of (real) symmetric matrices. Also, at least a superficial familiarity with Multivariate Analysis methods is deemed necessary. The methods discussed in this book are Multiple Regression Analysis, Principal Components Analysis, Simultaneous Components Analysis, MINRES Factor Analysis, Canonical Correlation Analysis, Redundancy Analysis, CANDECOMP/PARAFAC, INDSCAL, and Homogeneity Analysis. Although the purpose of each of these methods is explained in the text, previous exposure to at least some of them is recommended.

This book has a narrowly defined goal. Each of the nine methods mentioned involves a least squares minimization problem. The purpose of this book is to treat these minimum problems in a unified framework. It is this framework that matters, rather than the nine methods. The framework should provide the student with a thorough understanding of a large number of existing (alternating) least squares techniques, and may serve as a tool for dealing with novel least squares problems as they come about.

An eminent feature of the framework is that it does not rely on differential calculus. Partial derivatives play no role in it. Instead, the method of completing-the-squares is generalized to vector functions and matrix functions, to yield global minima for a variety of constrained and unconstrained least squares problems that have closed-form solutions, or serve as ingredients for monotonically convergent alternating least squares algorithms when closed-form solutions are not available. The versatility of the resulting solutions extends far beyond the specific nine methods.

Another characteristic of the book is that references are only included to the extent that they are considered helpful to the student. A full historic account of the methods has not been attempted, because that would detract from the main purpose, which is the explanation of the underlying principles of the methods. For similar reasons, examples of how the methods work in practice have been omitted. This text is meant to offer insight into the underlying (construction) principles of least squares methods, rather than explaining how to use these methods in practical applications. However, in cases where the nature of the least squares problem on which a method is based has important practical implications that are not generally known, these will be discussed.

I am obliged to Willem Heiser, Wim Krijnen, René van der Heijden, Jan Niesing, and Lidia Arends for helpful suggestions. However, a special debt of gratitude is owed to Henk Kiers for his continuous efforts to improve the exposition and organization of this book.

CONTENTS

INTRODUCTION

INTRODUCTION

A wide variety of Multivariate Analysis methods are based on least squares problems. The standard approach has been to use partial derivatives to obtain necessary conditions for a minimum, and then solving the equations involved. In the present book, a calculus-free point of departure is taken. First, a number of least squares problems that have closed-form solutions are solved by deriving lower bounds to the function to be minimized, and showing how these lower bounds can be attained. The method used to derive the lower bounds is essentially equivalent to completing the squares. The resulting solutions are globally optimal by definition.

Although unconstrained least squares problems are treated, they are outnumbered by the constrained least squares problems. Constraints of orthonormality and of limited rank play a key role in the developments. More often than not, constrained least squares problems can be transformed into equivalent constrained *trace maximization* problems. This explains why attainable upper bounds rather than lower bounds can be encountered quite often in this book which is, after all, devoted to least squares *minimization* problems.

The book is made up of four chapters. Chapter 1 deals with some results from matrix algebra that play a key role in this book, but may have received little or no attention in the matrix algebra course the readers have gone through. Chapter 2 starts with the well-known method of completing the squares to minimize a quadratic function. The theory is expanded by considering constrained quadratic functions, and functions of vectors or vector pairs. Explicit optima are derived for the unconstrained regression function, and for the linear, the quadratic, and the bilinear form, subject to unit length constraints, which can be seen as special cases of orthonormality constraints.

In Chapter 3 the vector functions are generalized to matrix functions, and explicit optima are derived without constraints or under constraints of orthonormality and of low rank. The resulting least squares solutions are summarized in a table at the end of Chapter 3. Global optima for three trace maximization functions are to be found in section 3.3. The matrix theory that is used to derive these optima is treated in section 3.2, although this topic may also be considered as an extension of Chapter 1. The results of section 3.3 and the table at the end of Chapter 3 constitute the framework of the book. It can be consulted without further consideration of the matrix theory from which it was derived.

In Chapter 4 the versatility of the framework is demonstrated in terms of nine methods from Multivariate Analysis. The framework provides global optima at once for the optimization problems of Multiple Linear Regression Analysis, Principal Components Analysis, Canonical Correlation Analysis, Redundancy Analysis, and Homogeneity Analysis. For the remaining applications, alternating least squares methods are given. None of the solutions given in Chapter 4 are new, but the unified treatment is. Also, important practical implications that follow from the underlying least squares problems are discussed, when they are not generally known.

Alternating least squares methods are typically applied when globally optimal solutions are not available. The minimization problem is then split in a series of subproblems that do have (conditional) globally optimal solutions, that can be used to construct an iterative algorithm. In cases where even the latter approach fails, one may resort to Alternating Lower Squares methods or Majorization. This topic is briefly touched upon in the Epilogue. The book is concluded with 49 exercises and answers.

CHAPTER 1

SOME BASIC RESULTS FROM MATRIX ALGEBRA

1.1. THE EIGENDECOMPOSITION OF A SYMMETRIC MATRIX

Let S be a *symmetric* $q \times q$ matrix of rank r ($r \leq q$). Then S has the eigendecomposition

$$S = K\Lambda K' \tag{1}$$

with $K'K = KK' = I_q$ and Λ diagonal. This means that, for any symmetric S, an orthonormal matrix K and a diagonal matrix Λ can be found that satisfy (1).

If S is singular, that is, if $r < q$, then a parsimonious expression exists in addition to (1). We define the $r \times r$ matrix Λ_r as the diagonal matrix containing the nonzero eigenvalues of S on the diagonal, and K_r as the $q \times r$ matrix containing the associated eigenvectors of S. Then instead of (1) we can write

$$S = K_r \Lambda_r K_r' \tag{2}$$

with $K_r'K_r = I_r$ and Λ_r diagonal and nonsingular. Properties of partitioned matrices guarantee that (2) follows from (1). Conversely, (1) follows from (2): That is, expanding K_r to a square orthonormal matrix K, by adding orthonormal columns, and expanding Λ_r with zeros to a diagonal $q \times q$ matrix Λ, we obtain that $S = K\Lambda K'$. If S is nonsingular ($r = q$) then (1) and (2) coincide, because then K_r equals K, and Λ_r equals Λ.

Gramian (or *positive semidefinite*) matrices are matrices that are the product of a matrix and its transpose. They have the special feature that none of their eigenvalues are negative. Conversely, every symmetric matrix without negative eigenvalues is Gramian. A Gramian matrix S always has a square root $S^{1/2}$ defined as $K\Lambda^{1/2}K'$, where $\Lambda^{1/2}$ is the diagonal matrix containing the square roots of the diagonal elements of Λ on the diagonal. Clearly, $S^{1/2}S^{1/2}=S$.

When a Gramian matrix S has an inverse, S is called *positive definite.* In that case the inverse $K\Lambda^{-1}K'$ of S is also Gramian, and the square root of that inverse equals $K\Lambda^{-1/2}K'$. The latter matrix is the inverse of $S^{1/2}$ and it is denoted by $S^{-1/2}$. Note that $S^{-1/2}S^{-1/2} = S^{-1}$ and $S^{1/2}S^{-1/2}=I$.

Asymmetric matrices do not have an eigendecomposition of the form (1) or (2). This is immediate because (1) and (2) already imply that S is symmetric.

1.2. THE SINGULAR VALUE DECOMPOSITION OF AN ARBITRARY MATRIX

Whereas the existence of an eigendecomposition of the form (1) or (2) is limited to symmetric matrices, *every* matrix has some sort of generalized eigendecomposition, namely a singular value decomposition (SVD), also referred to as Eckart-Young decomposition or *basic structure.*

Definition. Let A be an arbitrary $p \times q$ matrix $p \geq q$ of rank r ($r \leq q$). Then the SVD of A is the decomposition

$$A = PDQ' \tag{3}$$

with $P'P = I_q$, $Q'Q = QQ' = I_q$, and D diagonal, with *nonnegative* diagonal elements arranged from high (upper left) to low (lower right). So for any $p \times q$ matrix A $(p \geq q)$ there is a triple of matrices P, D, and Q satisfying (3). The diagonal elements in D are the *singular values* of A and they are nonnegative by convention. The number of positive singular values of A equals the rank of A, because rank(A)=rank($A'A$)=rank(QD^2Q')=rank(D^2)=rank(D). The columns of P and Q are called the left and right hand singular vectors of A, respectively.

A SVD for *horizontal* matrices (with more columns than rows, that is) also exists, but it requires no special treatment. Upon transposing such a matrix one can already use (3).

When A does not have full column rank, that is, when $r < q$, then again a parsimonious variant is available in addition to (3). Let P_r be defined as the $p \times r$ matrix containing the first r columns of P, Q_r as the $q \times r$ matrix containing the first r columns of Q, and D_r as the diagonal $r \times r$ matrix with diagonal elements equal to those singular values of A that are positive. Then instead of (3) one can also use the expression

$$A=P_r D_r Q_r' \qquad (4)$$

with P_r and Q_r columnwise orthonormal, and D_r diagonal with positive diagonal elements. Conversely, matrices P_r, D_r, and Q_r which satisfy (4) can be expanded to a triple P, D, and Q that satisfy (3).

The existence of a decomposition of the form (4) will now be proven *constructively*. That is, it will be shown how to obtain, for any given matrix A of rank r, a triple of matrices P_r, D_r, and Q_r that satisfy (4). First we note that $A'A$ is a *Gramian* matrix, so we may use (2) to express it as

$$A'A=K_r \Lambda_r K_r' \qquad (5)$$

with $K_r'K_r=I_r$ and Λ_r diagonal, with *positive* diagonal elements. Next we define

$$Q_r = K_r,\ D_r = \Lambda_r^{1/2},\ \text{and}\ P_r = AQ_r D_r^{-1}, \qquad (6)$$

where $\Lambda_r^{1/2}$ is the diagonal matrix having the square roots of the diagonal elements of Λ on the diagonal. Now it can be shown that the constructed matrices satisfy all requirements of (4), as follows.

Columnwise orthonormality of Q_r and P_r is immediate from (6) and D_r is clearly a diagonal matrix with positive diagonal elements. Also, we have

$$P_r D_r Q_r' = AQ_r D_r^{-1} D_r Q_r' = AQ_r Q_r' = AK_r K_r'. \qquad (7)$$

This does not yet yield (4) because, when $r<q$, $K_r K_r' \neq I$, owing to the fact that a columnwise orthonormal matrix cannot be rowwise orthonormal unless it is a

square matrix. So the solution has to be found elsewhere, that is, in (5). From (5) we have

$$A'AK_rK_r' = K\Lambda_rK_r'K_rK_r' = K\Lambda_rK_r' = A'A \tag{8}$$

and so $A'A(I–K_rK_r') = 0$. Premultiplying by $(I–K_rK_r')$ is very effective, and yields $(I–K_rK_r')A'A(I–K_rK_r') = 0$, which in turn implies that

$$A(I–K_rK_r')=0 \tag{9}$$

so $AK_rK_r' = A$. Using this and (7) jointly we have found (4) to be satisfied.

It is often believed that, for *symmetric* matrices, the SVD coincides with the eigendecomposition. This is not generally true, because eigenvalues can be negative, whereas singular values are defined to be nonnegative. Only in the case of *Gramian* matrices (symmetric, without negative eigenvalues) do SVD and eigendecomposition coincide.

When the eigendecomposition of a *symmetric* matrix is known, then the SVD can be obtained from it at once. If $S=K\Lambda K'$ and Λ contains no negative elements, then this expression is both an eigendecomposition and a SVD, with $P=K$, $Q=K$, and $D=\Lambda$. When, on the other hand, Λ does contain negative elements, then define a sign matrix T (diagonal, with diagonal elements 1 or -1) such that $T\Lambda$ has all elements nonnegative. Now $S=(KT)(T\Lambda)K'$, which is a SVD of S.

An example: The SVD of a 4×3 matrix A of rank 2. Let

$$A = \begin{pmatrix} 2.560 & 2.40 & 1.920 \\ .384 & -.64 & .288 \\ .288 & -.48 & .216 \\ 1.920 & 1.80 & 1.440 \end{pmatrix}.$$

The eigenvalues of $A'A$ are 25, 1, and 0, and the matrix containing the first two eigenvectors of $A'A$ is

$$K_2 = \begin{pmatrix} .64 & .48 \\ .60 & -.80 \\ .48 & .36 \end{pmatrix}.$$

Now define $Q_2 = K_2$, define $D_2 = \Lambda_2^{1/2} = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$, and define

$$P_2 = AQ_2D_2^{-1} = A\begin{pmatrix} .128 & .48 \\ .120 & -.80 \\ .096 & .36 \end{pmatrix} = \begin{pmatrix} .8 & 0 \\ 0 & .8 \\ 0 & .6 \\ .6 & 0 \end{pmatrix}.$$

It is readily verified that $P_2D_2Q_2' = A$, and that (4) is satisfied. Next, one may expand P_2 to obtain P, by adding $(.6 \quad 0 \quad 0 \quad -.8)'$ or $(0 \quad .6 \quad -.8 \quad 0)'$ as a third column, and Q_2 can be expanded to Q, by taking $(.6 \quad 0 \quad -.8)'$ as a third column.

Finally, upon constructing $D = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, we have $A=PDQ'$ and the SVD (3) has been obtained.

1.3. THE SCHWARZ INEQUALITY

Below, inequalities will play a key role in the treatment of functions. A particularly useful inequality for the inner product of two vectors **x** and **y** is the inequality $-(\mathbf{x}'\mathbf{x})^{1/2}(\mathbf{y}'\mathbf{y})^{1/2} \leq \mathbf{x}'\mathbf{y} \leq (\mathbf{x}'\mathbf{x})^{1/2}(\mathbf{y}'\mathbf{y})^{1/2}$, known as the Schwarz inequality or the Cauchy-Schwarz inequality. A proof for this inequality is as follows:

From the inequality

$$[\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1/2} - \mathbf{y}(\mathbf{y}'\mathbf{y})^{-1/2}]\ '\ [\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1/2} - \mathbf{y}(\mathbf{y}'\mathbf{y})^{-1/2}] \geq 0 \tag{10}$$

we have that $1+1-2\mathbf{x}'\mathbf{y}/\{(\mathbf{x}'\mathbf{x})^{1/2}(\mathbf{y}'\mathbf{y})^{1/2}\} \geq 0$, and hence

$$\mathbf{x}'\mathbf{y} \leq (\mathbf{x}'\mathbf{x})^{1/2}(\mathbf{y}'\mathbf{y})^{1/2}. \tag{11}$$

Upon replacing subtraction in (10) by addition, a similar inequality shows that

$$\mathbf{x}'\mathbf{y} \geq -(\mathbf{x}'\mathbf{x})^{1/2}(\mathbf{y}'\mathbf{y})^{1/2}. \tag{12}$$

Combining (11) and (12) yields the Schwarz inequality:

$$-(\mathbf{x}'\mathbf{x})^{1/2}(\mathbf{y}'\mathbf{y})^{1/2} \leq \mathbf{x}'\mathbf{y} \leq (\mathbf{x}'\mathbf{x})^{1/2}(\mathbf{y}'\mathbf{y})^{1/2}. \tag{13}$$

The Schwarz inequality holds as an equality when $\mathbf{x}$ is proportional to $\mathbf{y}$ or $-\mathbf{y}$. That is, if $\mathbf{x} = \lambda\mathbf{y}$ (for some positive scalar λ), then $\mathbf{x}'\mathbf{y} = (\mathbf{x}'\mathbf{x})^{1/2}(\mathbf{y}'\mathbf{y})^{1/2}$, and if $\mathbf{x} = -\lambda\mathbf{y}$ then $\mathbf{x}'\mathbf{y} = -(\mathbf{x}'\mathbf{x})^{1/2}(\mathbf{y}'\mathbf{y})^{1/2}$. The Schwarz inequality can be used, for instance, to prove that a correlation coefficient must always be in between –1 and 1.

1.4. HADAMARD PRODUCT, KRONECKER PRODUCT AND VEC NOTATION

The Hadamard product * of two matrices of the *same* order is defined as the elementwise product. So $(A*B)$ has as its ij-th element $a_{ij}b_{ij}$. An example:

When

$$A = \begin{pmatrix} 3 & 1 \\ 1 & -1 \\ 2 & 0 \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & 2 \\ -1 & 0 \\ 3 & 1 \end{pmatrix} \text{ then } (A*B) = \begin{pmatrix} 3 & 2 \\ -1 & 0 \\ 6 & 0 \end{pmatrix} \qquad (14\ )$$

The Kronecker product $A \otimes B$ of a $p \times q$ matrix A and a $k \times l$ matrix B is the $pk \times ql$ matrix consisting of pq submatrices $a_{ij}B$, i=l,...,p; j=l,...,q.

For example, if

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 5 \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & 2 \\ -1 & 0 \\ 3 & 1 \end{pmatrix} \qquad (15)$$

then

$$(A \otimes B) = \left(\begin{array}{cc|cc} 1 & 2 & 2 & 4 \\ -1 & 0 & -2 & 0 \\ 3 & 1 & 6 & 2 \\ \hline 3 & 6 & 5 & 10 \\ -3 & 0 & -5 & 0 \\ 9 & 3 & 15 & 5 \end{array}\right) = \begin{pmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{pmatrix}, \qquad (16)$$

whereas, on the other hand,

$$(B \otimes A) = \left(\begin{array}{cc|cc} 1 & 2 & 2 & 4 \\ 3 & 5 & 6 & 10 \\ \hline -1 & -2 & 0 & 0 \\ -3 & -5 & 0 & 0 \\ \hline 3 & 6 & 1 & 2 \\ 9 & 15 & 3 & 5 \end{array}\right) = \begin{pmatrix} b_{11}A & b_{12}A \\ b_{21}A & b_{22}A \\ b_{31}A & b_{32}A \end{pmatrix}. \qquad (17)$$

By Vec(*A*) we mean the vector containing all columns of *A*, one below another. Accordingly, if.

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 5 \end{pmatrix}, \text{ then } \mathrm{Vec}(A) = \begin{pmatrix} 1 \\ 3 \\ 2 \\ 5 \end{pmatrix} \text{ and } \mathrm{Vec}(A') = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 5 \end{pmatrix}. \qquad (18)$$

The following result makes it possible to express the elements of a matrix product of the form *ABC* as a function of Vec(*B*):

$$\mathrm{Vec}(ABC) = (C' \otimes A)\mathrm{Vec}(B). \qquad (19)$$

A proof can be found in Magnus and Neudecker (1991, pp. 30-31) and will not be given here. Instead, the result will be demonstrated in an example.
Let

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 5 \\ 1 & 1 \end{pmatrix}, B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \text{ and } C = \begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & 3 \end{pmatrix}. \qquad (20)$$

Then

$$ABC = \begin{pmatrix} 7 & 10 & 16 \\ 18 & 26 & 42 \\ 4 & 6 & 10 \end{pmatrix} \qquad (21)$$

and $\mathrm{Vec}(ABC) = (7 \quad 18 \quad 4 \quad 10 \quad 26 \quad 6 \quad 16 \quad 42 \quad 10)'$.

On the other hand, $(C' \otimes A)\mathrm{Vec}(B)$ is

$$\begin{pmatrix} 1 \\ 3 \\ 2 \\ 4 \end{pmatrix} = \text{Vec}(B)$$

$$(C'\otimes A) = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 3 & 5 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 3 & 5 \\ 0 & 0 & 1 & 1 \\ -2 & -4 & 3 & 6 \\ -6 & -10 & 9 & 15 \\ -2 & -2 & 3 & 3 \end{pmatrix} \quad \begin{pmatrix} 7 \\ 18 \\ 4 \\ 10 \\ 26 \\ 6 \\ 16 \\ 42 \\ 10 \end{pmatrix} = (C'\otimes A)\,\text{Vec}(B), \qquad (22)$$

which demonstrates the validity of (19).

An important reason for adopting the Vec notation is the property that the sum of squares of the elements of *ABC* equals the sum of squares of the elements of Vec(*ABC*), because these elements are the same. They are merely arranged differently. In addition, the following property will be used in the sequel:

$$\text{Vec} \sum_i (A_i B_i C_i) = \sum_i \text{Vec}(A_i B_i C_i). \qquad (23)$$

That is, the 'Vec of a sum is the sum of the Vecs'.

CHAPTER 2

FUNCTIONS OF VECTORS

2.1. FUNCTIONS AND EXTREME VALUES

A function is a rule by which an arbitrary number x is assigned a function value f(x). This general definition is rather abstract and hence not very helpful. It is easier to think of a function as some sort of a machine. Upon feeding an arbitrary number x into the machine (like dropping a coin), one obtains another number f(x) in return (like a bottle of coke). An example will be instructive. The function f, defined by the rule $f(x) = x^2 + 2$, squares the number fed into the machine, and adds 2 to that square. If x=0 is inserted, you get 2 out of it; if 3 goes in, 11 will come out, and so on. The number x that goes into the machine is called *the argument* of the function, and the number f(x) that comes out is called the associated *function value.* The symbol that is used to refer to the argument of a function is completely arbitrary. That is, when $f(x) = x^2+ 2$ and $g(a) = a^2+2$ then f and g are identical functions.

Techniques of data analysis often are based on maxima or minima of functions. For instance, it may be desired to *maximize* the variance accounted for, or to *minimize* the amount of error. Typically, such problems are approached by formulating the desired state of affairs in terms of a utility function, and searching for an extreme value (maximum or minimum) of this function.

A very familiar method of finding extreme values of a function is by differentiation. This method yields equations, the solutions of which produce values of x where the tangent to the function is in horizontal direction. Such values of x correspond to maxima, minima, or saddle points, which means that additional effort may be required to determine the precise location of a maximum or minimum. This is a first limitation of differentiation. Another limitation is that, even if it is known, for instance, that a function has a maximum for a certain x, it may still be unclear whether this maximum is local or global. The function depicted in Figure 1 has a local maximum for x=1.65, a local minimum for x= .73, and a global maximum for x = –.85.

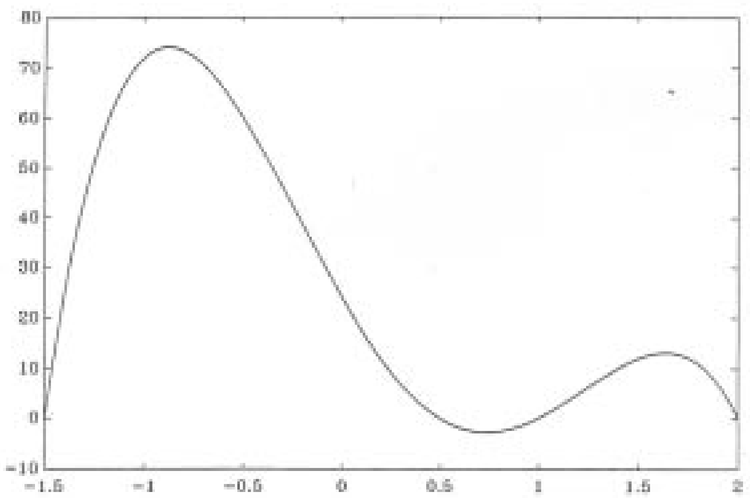

Figure 1. Graph of a function with local and global optima.

Clearly, a local maximum is useless when the highest possible function value (global maximum) is desired. Similarly, a local minimum is useless when one is after the lowest possible function value (global minimum).

Beside differentiating, there is another approach to determining maxima and minima of a function, namely, by deriving an upper or lower bound that can be attained. This will first be demonstrated in an example. Suppose that we want to find the global minimum of the function $f(x)=x^2-6x+11$. Then we may write $f(x)=(x-3)^2+2$, which shows that $f(x)\geq 2$. As a result, 2 is a lower bound to the function f. In the graph, a lower bound is represented by a horizontal line below which the function never comes. When we make a graph of $f(x)=(x-3)^2+2$, it can be seen that every function value is above the lower bound except at $x = 3$, where $f(x) = 2$. So there is a point (viz. $x = 3$), where the function *attains* its lower bound. It follows at once that this point yields the global minimum. We have found an *attainable* lower bound, and along with it, we get the global minimum in the bargain. Similarly, detecting an attainable upper bound gives the global maximum of the function at once. For instance, the global maximum of $f(x) = -x^2+8x-11$ can be found by writing $f(x)$ as

$f(x) = -(x-4)^2+5 \leq 5$, an upper bound that is attained for $x=4$. Hence, 5 is the global maximum of $f(x)$.

The maximum and minimum problems to be treated in the present and the following chapters will be handled, whenever possible, by finding attainable upper or lower bounds. Differentiation will be avoided throughout. In the sequel, a global minimum or maximum will be referred to simply as minimum or maximum, respectively, unless stated otherwise.

When dealing with upper or lower bounds to functions, two important facts should be borne in mind. First, it is of vital importance to realize that a bound that cannot be attained has no use whatsoever in tracing a maximum or minimum of a function. In the example $f(x) = (x-3)^2+ 2$ we have $f(x) \geq 2$ and hence also $f(x) > 1$. However, the lower bound 1 is never attained, so it is not the minimum of f. Only the highest lower bound can be of interest. In the same vein, amidst upper bounds only the lowest can be interesting.

It is of even greater importance to see that a lower or upper bound must not depend on *x*. Writing $f(x) = -x^2+8x -11$ as $f(x) = -(x-3)^2+(2x-2) \leq (2x-2)$ does yield an inequality (with equality obtained for $x=3$), but $(2x-2)$ is not an upper bound. It is just another function of *x*, which never attains a value below $f(x)$. The graph of this other function is by no means a horizontal line. Also, the point $x=3$ is a point where $f(x) = 2x-2$, but that is anything but the maximum of $f(x)$. A joint picture of $f(x)$ and the function $g(x) = 2x-2$ will show this quite clearly. It can be concluded that, to find an upper bound, we are to search for an expression for $f(x)$ of the form $f(x) \leq$ *something that does not depend on x*.

2.2. CONSTRAINED MAXIMA AND MINIMA

Above, a function has been described as a machine in which an arbitrary number (argument) is inserted, to obtain another number (function value). When the function is $f(x) = (x-3)^2+2$, the number that is obtained is never less than 2, and it is 2 when $x=3$ is the number inserted. Therefore, 2 is the minimum of f(x), and it is attained for x=3. However, this function does not have a maximum. As we choose x farther away from 3, f(x) will reach an increasingly high function value. There is no upper bound to f(x), let alone an upper bound that can be attained.

The situation may change dramatically, when the argument of the function is constrained. For instance, if the requirement (constraint) is imposed that the x to be inserted in the machine has to be in the interval [–1, 1], that is, if the constraint $x^2 \leq 1$ is imposed, then f(x) does have a maximum. This can be verified as follows: From $x^2 \leq 1$ we have $f(x) = x^2 - 6x + 11 \leq 1 - 6x + 11 = -6x + 12$. This does not yet give an upper bound: We still need to lose the term $-6x$. That is not very difficult. It follows from $x \geq -1$ that $-x \leq 1$ so $-6x \leq 6$, and $f(x) \leq 6+12=18$. As a result, 18 is an upper bound to f(x) in the interval $-1 \leq x \leq 1$. Now the question is whether or not this bound can be attained in that interval. To answer that question, we need to verify if the inequalities that were used along the road can hold as equalities for a certain x in the interval. That is, we search for an x in the interval $x^2 \leq 1$ such that $x^2 = 1$ and $-x = 1$. The answer ($x = -1$) is immediate. So f(–1) = 18, and 18 is the maximum of f(x) subject to the constraint $-1 \leq x \leq 1$.

Another example. Suppose that we want to minimize $f(x) = -x^2+3$ over the interval $x^2 \leq 4$, see Figure 2. There is no unconstrained minimum, but the

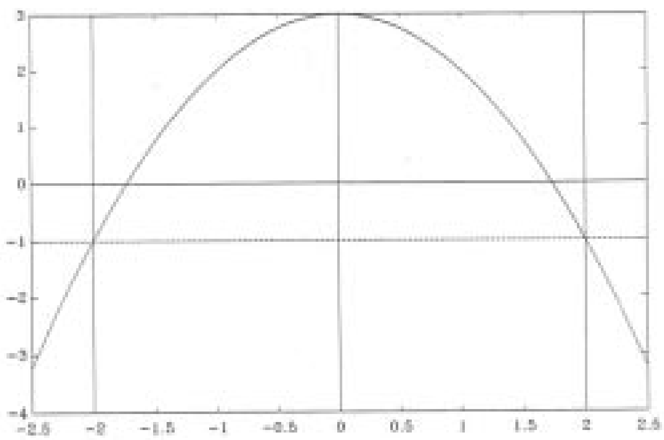

Figure 2. Graph of the function $f(x) = -x^2+3$.

constrained minimum does exist. Specifically, write f(x) = $-x^2+3 \geq -4+3 = -1$ (using $-x^2 \geq -4$), so -1 is a lower bound to f(x) on the interval $-2 \leq x \leq 2$. In the graph of the unconstrained f(x) it should be clear that -1 is not a lower bound at all. However, for those x values admitted, -1 is an attainable lower bound and hence it is the constrained minimum of f(x). Incidentally, this minimum is not unique, because it is attained both for $x = 2$ and for $x = -2$.

2.3. **VECTOR FUNCTIONS AND EXTREME VALUES**

So far, a function has been viewed as some sort of machine in which a number x is inserted, upon which another number f(x) comes out. The theory is now generalized by considering vector functions. A vector function is a rule which assigns to any vector **x** (of a specified order) a number f(x). The *argument* of the function now is a *vector* **x** instead of a *number* x. The *function value* f(**x**) is a number as it was before. An example of a vector function is $f(\mathbf{x})=3x_1+4x_2$. The argument is the vector $\mathbf{x}=(x_1\ \ x_2)'$. If $\mathbf{x}=(1\ \ 2)'$ is inserted into the machine, the number 11 comes out; if $\mathbf{x} = (1\ \ 1)'$ is inserted, the number 7 comes out, and so on. The function f given here has no maximum or minimum. However, if the constraint $\mathbf{x}'\mathbf{x}=1$ is imposed, there certainly are extreme values. This will be shown shortly. First, however, the reader is invited to guess what vector **x** with $\mathbf{x}'\mathbf{x}=1$ gives the highest possible value for $f(\mathbf{x}) = 3x_1+4x_2$.

take some time to reflect ....

Have you come up with the answer $\mathbf{x}' = (0\ 1)$? Then you are not the only one who did. Whether or not the answer is correct will be revealed shortly. First, an attainable upper bound will be determined.

Upon defining the constant vector $\mathbf{y}=(3\ \ 4)'$, we can write f(**x**) as $f(\mathbf{x})=\mathbf{x}'\mathbf{y}$. The Schwarz inequality tells us that $\mathbf{x}'\mathbf{y} \leq (\mathbf{x}'\mathbf{x})^{½}(\mathbf{y}'\mathbf{y})^{½}$. This does not yet provide an upper bound, because the right hand side of the inequality still depends on **x**. However, the constraint $\mathbf{x}'\mathbf{x}=1$ implies that, for those vectors to be admitted, always $(\mathbf{x}'\mathbf{x})^{½} = 1$. This allows us to use the inequality $\mathbf{x}'\mathbf{y} \leq (\mathbf{y}'\mathbf{y})^{½} = \sqrt{25} = 5$. Therefore, 5 is an upper bound to $\mathbf{x}'\mathbf{y}$ subject to the constraint $\mathbf{x}'\mathbf{x}=1$. This upper bound can be attained. To verify this, consider the derivation of the Schwarz inequality. The inequality $\mathbf{x}'\mathbf{y} \leq (\mathbf{x}'\mathbf{x})^{½}(\mathbf{y}'\mathbf{y})^{½}$ holds as an equality if and only if (10) holds as an equality. That occurs only if the vector $[\mathbf{x}(\mathbf{x}'\mathbf{x})^{-½} - \mathbf{y}(\mathbf{y}'\mathbf{y})^{-½}]$ is a zero vector, that is, if $\mathbf{x}$ is proportional to $\mathbf{y}$. So we choose $\mathbf{x}$ proportional to $\mathbf{y}=(3 \;\; 4)'$ and such that $\mathbf{x}'\mathbf{x}=1$. As a result, $\mathbf{x}$ is the unit length rescaled version of $\mathbf{y}$, that is, $\mathbf{x} = \mathbf{y}(\mathbf{y}'\mathbf{y})^{-½} = .2\mathbf{y} = (.6 \;\; .8)'$. Evaluating the value of $\mathbf{x}'\mathbf{y}$ yields $\mathbf{x}'\mathbf{y} = .6\times3 + .8\times4 = 1.8 + 3.2 = 5$. It follows that 5 is an *attainable* upper bound, hence it is the desired maximum. Note that the maximum is attained for $\mathbf{x} = (.6 \;\; .8)'$, and that the solution $\mathbf{x} = (0 \;\; 1)'$ is not optimal, yielding a function value as low as 4. Congratulations to those who came up with the right answer.

In the example above, the vector $\mathbf{y}$ of constants was known in advance to be $\mathbf{y} = (3 \;\; 4)'$. It is of utmost importance, however, to see that the solution $\mathbf{x}=\mathbf{y}(\mathbf{y}'\mathbf{y})^{-½}$ is valid quite generally for any conceivable, unknown vector $\mathbf{y}$, as long as it is independent of $\mathbf{x}$: From $\mathbf{x}'\mathbf{x}=1$ it follows that $\mathbf{x}'\mathbf{y}\leq(\mathbf{x}'\mathbf{x})^{½}(\mathbf{y}'\mathbf{y})^{½} =(\mathbf{y}'\mathbf{y})^{½}$ (upper bound) and we have $\mathbf{x}'\mathbf{y}=(\mathbf{y}'\mathbf{y})^{½}$ (the upper bound is attained) if $\mathbf{x}=\mathbf{y}(\mathbf{y}'\mathbf{y})^{-½}$, as is easy to verify. From now on, extreme values of vector functions will always be determined in such general form.

The function $f(\mathbf{x}) = \mathbf{x}'\mathbf{y}$ is called a *linear form.* Having dealt with maxima of linear forms, we now turn to *extreme values* of two different vector functions, namely the *quadratic form,* and the *regression function.*

A vector function of the form

$$g(\mathbf{x}) = \mathbf{x}'A\mathbf{x} \tag{24}$$

where A is a known or unknown fixed (square) matrix, is called a *quadratic form.* The argument is again a vector $\mathbf{x}$, and the function value is, of course, a number. For instance, if $A=\begin{pmatrix} 2 & 0 \\ 3 & -1 \end{pmatrix}$, then $g(\mathbf{x})$ is the function

$$g(\mathbf{x}) = 2x_1^2 + 3x_1x_2 - x_2^2 . \tag{25}$$

Again, it is desired to find the maximum of the function subject to the constraint $\mathbf{x}'\mathbf{x} = 1$. First, assume that $A=A'$. This can be done without loss of generality. That is, if $A \neq A'$, we write

$$g(\mathbf{x}) = \mathbf{x}'A\mathbf{x} = ½\mathbf{x}'A\mathbf{x} + ½\mathbf{x}'A\mathbf{x} = ½\mathbf{x}'A\mathbf{x} + ½\mathbf{x}'A'\mathbf{x} = \mathbf{x}'\left(\frac{A+A'}{2}\right)\mathbf{x} \tag{26}$$

and define $A^* = ½(A+A')$, the *symmetric part* of A. Upon replacing A by its symmetric part A^*, we can continue with $g(\mathbf{x}) = \mathbf{x}'A^*\mathbf{x}$, which is the same function as before, except that the coefficient matrix is now symmetric.

The next step is to find an attainable upper bound. Because A is symmetric (or has been replaced by its symmetric part, if needed), it has an eigendecomposition $A=K\Lambda K'$, with K orthonormal and Λ diagonal, with diagonal elements arranged in weakly descending order. The latter detail is often not relevant, but in this case it certainly is. Using the eigendecomposition we can write

$$g(\mathbf{x}) = \mathbf{x}'A\mathbf{x} = \mathbf{x}'K\Lambda K'\mathbf{x} = \mathbf{y}'\Lambda\mathbf{y} = \Sigma_i y_i^2\lambda_i \leq \Sigma_i y_i^2\lambda_1 = \lambda_1 , \tag{27}$$

where $K'\mathbf{x}$ has been denoted by $\mathbf{y}$, and y_i is the i-th element in the vector $\mathbf{y}$. The step $\mathbf{y}'\Lambda\mathbf{y} = \Sigma y_i^2\lambda_i$ follows from the diagonal property of Λ, and the inequality $\lambda_i \leq \lambda_1$ for every i follows from the descending order of the eigenvalues. The constraint $\mathbf{x}'\mathbf{x}=1$ is equivalent to the constraint $\mathbf{y}'\mathbf{y} = \mathbf{x}'KK'\mathbf{x} = \mathbf{x}'I\mathbf{x} = \mathbf{x}'\mathbf{x} = 1$. Because A is constant, so are its eigenvalues. It follows that λ_1 is an upper bound to $g(\mathbf{x})$, see (27). This upper bound can be attained by taking $\mathbf{x}=\mathbf{k}_1$, the first column of K. That yields

$$g(\mathbf{k}_1) = \mathbf{k}_1'A\mathbf{k}_1 = \mathbf{k}_1'\lambda_1\mathbf{k}_1 = \lambda_1\mathbf{k}_1'\mathbf{k}_1 = \lambda_1, \tag{28}$$

owing to the fact that $\mathbf{k}_1$ is an eigenvector of A, that is, $A\mathbf{k}_1=\lambda_1\mathbf{k}_1$. We have thus found the maximum of the quadratic form $\mathbf{x}'A\mathbf{x}$ subject to the constraint $\mathbf{x}'\mathbf{x}=1$: The maximum is the largest eigenvalue of A, and it is attained when $\mathbf{x}$ is the associated unit length eigenvector.

A third type of vector function to be considered is the *regression function*

$$\mathrm{h}(\mathbf{x}) = (A\mathbf{x}-\mathbf{y})'(A\mathbf{x}-\mathbf{y}) = ||A\mathbf{x}-\mathbf{y}||^2 \qquad (29)$$

where the notation $||\ldots||^2$ indicates the sum of squared elements of a vector or matrix. The minimum of h plays a key role in linear regression analysis and a great many other techniques. Specifically, the problem is to find a vector $\mathbf{x}$ of weights that generate a linear combination $A\mathbf{x}$, which is as similar as possible to a given vector $\mathbf{y}$, in terms of the least squares criterion. The vector $(A\mathbf{x}-\mathbf{y})$ contains errors of prediction, and its sum of squares is $(A\mathbf{x}-\mathbf{y})'(A\mathbf{x}-\mathbf{y}) = \mathrm{h}(\mathbf{x})$. Therefore, the minimum of $\mathrm{h}(\mathbf{x})$ corresponds to the smallest possible sum of squared prediction errors. This minimum can be found as follows. From the definition of $\mathrm{h}(\mathbf{x})$ we have

$$\begin{aligned} \mathrm{h}(\mathbf{x}) &= \mathbf{x}'A'A\mathbf{x} - 2\mathbf{x}'A'\mathbf{y} + \mathbf{y}'\mathbf{y} \\ &= \mathbf{x}'A'A\mathbf{x} - 2\mathbf{x}'A'\mathbf{y} + \mathbf{y}'A(A'A)^{-1}A'\mathbf{y} + \mathbf{y}'\mathbf{y} - \mathbf{y}'A(A'A)^{-1}A'\mathbf{y} \\ &= ||A\mathbf{x} - A(A'A)^{-1}A'\mathbf{y}||^2 + \mathbf{y}'\mathbf{y} - \mathbf{y}'A(A'A)^{-1}A'\mathbf{y} \geq \mathbf{y}'\mathbf{y} - \mathbf{y}'A(A'A)^{-1}A'\mathbf{y}, \end{aligned} \qquad (30)$$

which yields the lower bound $(\mathbf{y}'\mathbf{y} - \mathbf{y}'A(A'A)^{-1}A'\mathbf{y})$. It is evident that this lower bound can be attained. That is, if we choose

$$\mathbf{x} = (A'A)^{-1}A'\mathbf{y}, \qquad (31)$$

the term $||A\mathbf{x} - A(A'A)^{-1}A'\mathbf{y}||^2$ is zero, so $\mathrm{h}(\mathbf{x}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'A(A'A)^{-1}A'\mathbf{y}$. This is the minimum of h($\mathbf{x}$). In (31) one may recognize the well-known vector of regression weights, to estimate a vector $\mathbf{y}$ by a linear combination $A\mathbf{x}$ of the columns of a matrix A.

It is important to appreciate the particular way in which a function is specified. The very fact that the regression function (29) has been labeled h($\mathbf{x}$) rather than, for instance, h($\mathbf{y}$), tells us that h is a 'machine', in which the vector $\mathbf{x}$ is inserted as argument, whereas A and $\mathbf{y}$ are fixed but arbitrary constants. The lower bound $\mathbf{y}'\mathbf{y} - \mathbf{y}'A(A'A)^{-1}A'\mathbf{y}$ would not be constant if A or $\mathbf{y}$ would depend on (vary with) $\mathbf{x}$. A completely different situation would arise if we were to minimize, for instance,

$$\mathrm{h}^*(\mathbf{y}) = \|A\mathbf{x}-\mathbf{y}\|^2 . \qquad (32)$$

Now it is $\mathbf{y}$ that varies, while A and $\mathbf{x}$ are constant, and we can use the lower bound $\mathrm{h}^*(\mathbf{y}) \geq 0$. This lower bound follows trivially from the nonnegativity of a sum of squares, and can be attained by simply choosing $\mathbf{y}$ as $\mathbf{y} = A\mathbf{x}$.

Obviously, the nonnegativity of a sum of squares yields a lower bound 0 for h($\mathbf{x}$) as well as it did for h*($\mathbf{y}$). However, for h($\mathbf{x}$) this lower bound is not attainable, because $\mathrm{h}(\mathbf{x}) = \|A\mathbf{x}-\mathbf{y}\|^2$ is zero only if we choose $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{y}$. Such an $\mathbf{x}$ exists only if $\mathbf{y}$ happens to be in the column space of A. Therefore, the lower bound zero is not generally attainable for h($\mathbf{x}$).

Above, three vector functions have been considered. Both for the linear and for the quadratic form, the maximum has been determined, subject to the constraint that the argument vector be of length 1, that is, $\mathbf{x}'\mathbf{x}=1$. For the regression function, the *unconstrained* minimum was determined. If we were to minimize this function subject to, e.g., the constraint $\mathbf{x}'\mathbf{x}=1$, the problem would be far more difficult. The solution to this (complicated) constrained regression problem is not given here, but can be found in Ten Berge and Nevels (1977) or in Golub and Van Loan (1989, § 12.1.2).

2.4. THE BILINEAR FORM, AND A RECAPITULATION

The vector functions that have been treated so far assign a number (function value) to *a single vector* **x** (argument). We now expand the theory by considering a vector function for which the argument (what goes in) is a *pair of vectors* **x** and **y**, rather than a *single* vector **x** or **y**. One such function is the *bilinear form,* defined as

$$g(\mathbf{x},\mathbf{y}) = \mathbf{x}'A\mathbf{y} \tag{33}$$

where A is a constant $p\times q$ matrix. The argument of this function is a pair of vectors **x** and **y**, as can be seen from its specification $g(\mathbf{x},\mathbf{y})$. This function should not be confused with $f^*(\mathbf{x})=\mathbf{x}'A\mathbf{y}$, a linear form with A and **y** constant, or with $f^+(\mathbf{y}) = \mathbf{x}'A\mathbf{y}$, which is another linear form, with **x** and A constant.

We will now determine the maximum of $g(\mathbf{x},\mathbf{y})$, subject to the constraint that both **x** and **y** be unit length vectors. Again, an attainable upper bound will be found, but this is slightly more complicated than in earlier problems.

Let the constraint $\mathbf{x}'\mathbf{x} = \mathbf{y}'\mathbf{y} = 1$ be imposed on the vector pair **x** and **y**. Then the Schwarz inequality, applied to the vectors **x** and $A\mathbf{y}$, yields

$$g(\mathbf{x},\mathbf{y}) = \mathbf{x}'A\mathbf{y} \le (\mathbf{x}'\mathbf{x})^{1/2}(\mathbf{y}'A'A\mathbf{y})^{1/2} = (\mathbf{y}'A'A\mathbf{y})^{1/2}. \tag{34}$$

This inequality is not an upper bound, because $(\mathbf{y}'A'A\mathbf{y})^{1/2}$ depends on which **y** is inserted into the 'machine', and **y** is part of the argument. To get rid of **y** too, we need to take an additional step. In the present case, the step we need is to apply a result that has been obtained above. Specifically, consider the term $\mathbf{y}'A'A\mathbf{y}$. This is a quadratic form, with **y** as argument, and $A'A$ as constant matrix. Because **y** is constrained to be of unit length, this quadratic form has the largest eigenvalue of $A'A$ as its maximum, see (28). It follows from (34) that $g(\mathbf{x},\mathbf{y})$ has the square root of this eigenvalue as an upper bound. This upper bound is also the largest singular value of A. Express the SVD of A as

$$A = PDQ' \tag{35}$$

with $P'P = Q'Q = I_q$ and D diagonal, with nonnegative ordered diagonal elements $d_1 \geq d_2 \geq \ldots \geq d_q \geq 0$. Then $A'A$ has the eigendecomposition

$$A'A = QD^2Q' . \tag{36}$$

So the maximum of $\mathbf{y}'A'A\mathbf{y}$ is d_1^2, whence

$$\mathrm{g}(\mathbf{x},\mathbf{y}) \leq (\mathbf{y}'A'A\mathbf{y})^{½} \leq \sqrt{d_1^2} = d_1. \tag{37}$$

It can be concluded that the bilinear form has as an upper bound the largest singular value of its coefficient matrix (of *A*, that is), if a constraint of unit length for the argument vectors **x** and **y** is imposed. This upper bound is, in fact, the maximum, because it can be attained. That is, if we choose

$$\mathbf{x} = \mathbf{p}_1 \quad \text{and} \quad \mathbf{y} = \mathbf{q}_1, \tag{38}$$

the first columns of *P* and *Q*, respectively, then g attains the value

$$\mathrm{g}(\mathbf{x},\mathbf{y}) = \mathbf{p}_1'A\mathbf{q}_1 = \mathbf{p}_1'PDQ'\mathbf{q}_1 = \mathbf{e}_1'D\mathbf{e}_1 = d_1 \tag{39}$$

where $\mathbf{e}_1$ is the first column of the identity matrix *I*. This completes the solution for the maximum of a bilinear form, subject to unit length constraints.

In the derivation above, we have used a pair of inequalities jointly. The first inequality (34) yielded an expression dependent on **y**, but independent of **x**. In other words, the vector **x** was *eliminated from the argument* in the first step. This principle of *argument reduction* or *elimination of variables* will be used quite often in the sequel.

It should be noted that a quadratic form can be seen as a constrained version of a bilinear form. That is, the function g(**x**) = $\mathbf{x}'A\mathbf{x}$ can be considered as the bilinear form g(**x**,**y**)=$\mathbf{x}'A\mathbf{y}$, subject to the *equality constraint* that **x** and **y** be equal. For this reason, the same symbol (g) has been used to denote either function.

The main results obtained so far can be summarized as follows:

1. The maximum of the linear form f(**u**) = $\mathbf{u}'\mathbf{v}$ subject to the constraint $\mathbf{u}'\mathbf{u} = 1$ is $(\mathbf{v}'\mathbf{v})^{½}$, and is attained for $\mathbf{u} = \mathbf{v}(\mathbf{v}'\mathbf{v})^{-½}$.
2. The maximum of the quadratic form g(**u**) = $\mathbf{u}'B\mathbf{u}$, subject to the constraint $\mathbf{u}'\mathbf{u} = 1$, is the largest eigenvalue of *B*, if *B* is symmetric. If *B* is not symmetric, it should be replaced by its symmetric part ½($B+B'$). The maximum is attained when **u** is the first unit length eigenvector of *B* (or its symmetric part).
3. The maximum of the bilinear form g(**u**,**v**) = $\mathbf{u}'C\mathbf{v}$, subject to the constraint $\mathbf{u}'\mathbf{u} = \mathbf{v}'\mathbf{v} = 1$, is the largest singular value of *C*, and is attained when **u** and **v** are the first left and right hand singular vectors of *C*, respectively. There is no need for *C* to be square, here.
4. The minimum of the (unconstrained) regression function h(**u**) = $\|G\mathbf{u} - \mathbf{z}\|^2$ is $\mathbf{z}'\mathbf{z} - \mathbf{z}'G(G'G)^{-1}G'\mathbf{z}$, and is attained for $\mathbf{u} = (G'G)^{-1}G'\mathbf{z}$. If a length constraint for **u** is imposed, the problem is more complicated, and its solution is to be found in Ten Berge en Nevels (1977) and in Golub and Van Loan (1989).

It should be noted that the symbols used here depart from earlier notation. This illustrates that it is the *form* of the function, rather than the *symbols* used, that determines what the maximum or minimum will be.

It may come as a surprise that most of the optimization problems covered so far deal with maxima rather than minima. As will be explained later, constrained least squares problems can very often be transformed into equivalent maximization problems that are more manageable. In fact, every maximum treated in this book corresponds to a minimum for some least squares function.

CHAPTER 3

FUNCTIONS OF MATRICES

3.1. MATRIX FUNCTIONS AS GENERALIZED VECTOR FUNCTIONS

A matrix function is a rule which assigns a number to a matrix. After the introduction of the concept of a vector function, this does not come as a surprise. A matrix function can be viewed as a machine in which a matrix is inserted as argument, and a number (function value) comes out. In the next sections, attention will first be focused on three matrix functions which can be represented by the trace of a certain matrix. These functions are straightforward generalizations of the linear form, the quadratic form, and the bilinear form, respectively. The previously imposed constraint that the argument vectors have unit lengths now appears in a generalized form as the constraint that the argument matrix be *columnwise orthonormal.* That is, the columns of this matrix must each have unit length, and must be orthogonal (have scalar products zero). The three functions concerned are the *generalized linear form*

$$\mathrm{f}(X) = \mathrm{tr}X'Y, \qquad (40)$$

with Y constant;. the *generalized quadratic form*

$$\mathrm{g}(X) = \mathrm{tr}X'AX \qquad (41)$$

with A constant and square, and the *generalized bilinear form*

$$\mathrm{g}(X_1,X_2) = \mathrm{tr}X_1'AX_2 \qquad (42)$$

with A constant. For $\mathrm{f}(X)$ and $\mathrm{g}(X)$ the constraint $X'X = I$ is imposed, and

$g(X_1,X_2)$ is constrained by both $X_1'X_1 = I$ and $X_2'X_2 = I$.

In addition to these *trace functions,* attention will also be paid to matrix generalizations of the regression function h(**x**). The most obvious one is the *matrix regression function*

$$h(X) = \|AX - Y\|^2. \tag{43}$$

In certain applications, an even more general regression function is considered, namely the *Penrose regression function*

$$h(X) = \|AXB' - Y\|^2, \tag{44}$$

described by Penrose (1956). The function (43) is the special case of (44) where $B = I$. A function f_1 is a special case of another function f_2 if the two functions are of the same form, except that a certain complication, present in f_2, is missing in f_1. As a result, the solution for a minimum or maximum of f_1 can be obtained at once from the corresponding solution for f_2, but it is likely to have a simpler expression than that of f_2.

A special case should not be confused with a constraint. A constraint implies a limitation of the possible arguments of the function, and may, in fact, complicate the solution for a maximum or minimum. Solutions for constrained and unconstrained optima for the same function are generally unrelated. If they happen to coincide, the constraint is called *inactive.*

Regression problems of various kinds will be treated. First, unconstrained regression problems will be dealt with. Next, regression problems with orthonormality restrictions will be examined. Finally, we shall deal with regression problems with the constraint that the argument matrix be of a certain *low rank.* That restriction has no meaningful counterpart for vector functions, because a vector has rank 1, unless it is the zero vector, which has rank zero. So rank restrictions will enter the treatment as a novel subject matter.

In the sections to follow, it will first be shown how to maximize trace functions of the form (40), (41), or (42), subject to orthonormality constraints. A general theorem will be used to find attainable upper bounds. That theorem will first be treated. The relevance of the trace maximization problems for *least* squares problems is explained in Exercise 39.

3.2. KRISTOF UPPER BOUNDS FOR TRACE FUNCTIONS

Kristof (1970) has given an upper bound for trace functions of the form $f(G_1,...,G_m) = \text{tr}G_1C_1G_2C_2...G_mC_m$, where $C_1,\ldots,C_m$ are given diagonal $n\times n$ matrices, and G_1 through G_m are constrained to be orthonormal. The cases m=1 and m=2 of this upper bound date back to Von Neumann (1937). We shall use only the simplest case, where m=1. For that case Kristof′s upper bound reads

$$f(G) = \text{tr}(GC) \le \text{tr}C \tag{45}$$

if G is constrained to be orthonormal, and C has no negative elements. This simple version of Kristof′s bound can easily be proven. Because C is diagonal, we have $\text{tr}GC = \sum_{i=1}^{n} g_{ii}c_i$. The orthonormality of G implies $g_{ii} \le 1$. This, and the fact that $c_i \ge 0$ for $i = 1,...,n$, yields $\text{tr}GC = \sum_{i=1}^{n} g_{ii}c_i \le \sum_{i=1}^{n} c_i = \text{tr}C$. The upper bound is attained if we choose $G=I_n$.

The range of possible applications is broadened considerably if the constraint is relaxed to the effect that G is merely required to be a *submatrix* of an orthonormal matrix.

Definition. A matrix is *suborthonormal* (s.o.) if it can be completed to an orthonormal matrix, by appending rows or columns, or both. Every s.o. matrix can be viewed as a submatrix of some orthonormal matrix.

Example 1: The matrix $G = \begin{pmatrix} 1.2 & 0 \\ 0 & 1 \end{pmatrix}$ is *not* s.o. because it has an element larger than 1.

Example 2: The matrix $G = \begin{pmatrix} .8 & 0 \\ 0 & 1 \end{pmatrix}$ is s.o. because it is a submatrix of the

orthonormal matrix $G^* = \begin{pmatrix} .8 & 0 & .6 \\ 0 & 1 & 0 \\ .6 & 0 & -.8 \end{pmatrix}$.

Example 3: The matrix $G = \begin{pmatrix} .8 & .6 \\ -.6 & .4 \end{pmatrix}$ is *not* s.o. because, if it were, it would be a submatrix of some orthonormal matrix G^*. Then the first column of G^* would contain elements .8, −.6, and zeroes elsewhere, because the sum of squares of these elements must be 1. It follows that the inner product of the first two columns of G^* would be .24, which would contradict the orthonormality of G^*. Therefore, G is not s.o..

Property 1. Every columnwise or rowwise orthonormal matrix is s.o.. This is because it can readily be completed to be orthonormal.

Property 2. The product of any two s.o. matrices is also s.o.. The proof can be found in Ten Berge (1983).

Using the concept of suborthonormality allows us to generalize Kristof's upper bound (Ten Berge, 1983). Again, only the simplest case ($m = 1$), covered by the following theorem, will be needed.

Theorem: If G is a $n \times n$ s.o. matrix of (limited) rank $r \leq n$, and C is diagonal, nonnegative, with diagonal elements $c_1 \geq c_2 \geq ... \geq c_n \geq 0$, then the trace of GC has the upper bound

$$f(G) = \text{tr}(GC) \leq c_1 + ... + c_r, \qquad (46)$$

which is the sum of the r largest elements in C. This upper bound is usually smaller than tr(C), and is attained for the s.o. matrix

$$G = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}, \qquad (47)$$

which has visibly rank *r*. Two examples will be instructive to illustrate this *generalized Kristof theorem.*

Example 4: If $C = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and *G* is constrained to be s.o. of rank 2 or less, then the maximum of trGC is 7. If, on the other hand, the rank of *G* is to be 1, then the maximum of f(*G*) is 5. In that case 7 is still an upper bound to f, but it is not attainable for a *G* of rank 1.

Example 5: Let *C* be as in the previous example, and let *U* of order 3×2 be constrained to be columnwise orthonormal. Then the function g(*U*) defined as g(*U*)=tr$U'CU$ has an upper bound 7. To verify this, write tr$U'CU$ =tr$UU'C$=trGC, with $G = UU'$. Because *U* must have rank 2, so does *G*. Moreover, the product of s.o. matrices is s.o., hence *G* must be s.o., see properties 1 and 2 above. Applying the generalized Kristof theorem gives the upper bound 7.

3.3. HOW TO MAXIMIZE TRACE FUNCTIONS USING KRISTOF BOUNDS

The generalized Kristof theorem of the previous section is remarkably efficient in determining maxima of generalized linear, quadratic, and bilinear forms, subject to orthonormality constraints.

An upper bound to the generalized linear form

$$f(X) = \text{tr}X'Y \qquad (48)$$

subject to the constraint $X'X = I_q$, where *Y* is a given $p\times q$ matrix, can be obtained from the SVD $Y = PDQ'$. Using this in (48) yields

$$f(X) = \text{tr}X'Y = \text{tr}X'PDQ' = \text{tr}Q'X'PD = \text{tr}GD \leq \text{tr}D, \qquad (49)$$

with G defined as the $q\times q$ matrix $Q'X'P$, with rank(G) $\leq q$. Because G is the product of three s.o. matrices, G is also s.o.. This is why (49) can be derived from (46), with D taking the role of C.

Clearly, tr$D$ does not depend on X. Therefore, (49) gives an upper bound to f(X) subject to $X'X = I$. The upper bound is attained for $G = I$, hence for

$$X = PQ', \tag{50}$$

with P and Q defined by the SVD of Y. We have thus found the maximum of the generalized linear form (48) subject to the constraint $X'X=I$.

Next, we turn to the maximum of the generalized quadratic form

$$\text{g}(X) = \text{tr}X'AX \tag{51}$$

subject to $X'X = I_q$, for a given matrix A of order $p\times p$. When A is asymmetric, it is replaced by its symmetric part ½($A + A'$), see section 2.3. This does not affect the function because tr($X'AX$) = tr($X'A'X$). So A can be taken symmetric, and has an eigendecomposition $A = K\Lambda K'$, with $K'K = KK' = I_p$ and Λ diagonal and ordered. Let it be assumed that A is Gramian, so $\lambda_1 \geq ... \lambda_p \geq 0$. When A is not Gramian, the solution will still be the same, but this is a detail to be ignored here. Next, defining $G = K'XX'K$, which is s.o. and of rank q, we have,

$$\text{g}(X) = \text{tr}X'AX = \text{tr}X'K\Lambda K'X = \text{tr}K'XX'K\Lambda = \text{tr}G\Lambda \leq \lambda_1 + ... \lambda_q, \tag{52}$$

and an upper bound to g(X) has been found. The upper bound is attained for

$$G = \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix}. \tag{53}$$

If we choose

$$X = K_q N, \qquad (54)$$

where K_q contains the first q columns of K, and N is an arbitrary orthonormal $q \times q$ matrix, then $G = K'XX'K = K'K_qNN'K_q'K = K'K_qK_q'K = \begin{pmatrix} I_q \\ 0 \end{pmatrix}(I_q \;\; 0)$, which has the form prescribed by (53). So (54) shows how to maximize the generalized quadratic form (51).

Next, we examine the maximum of the generalized bilinear form

$$\text{g}(X_1, X_2) = \text{tr}X_1'AX_2 \qquad (55)$$

subject to $X_1'X_1 = X_2'X_2 = I_q$, with X_1 of order $p_1 \times q$, and X_2 of order $p_2 \times q$, $p_1 \geq p_2$. Using the SVD $A=PDQ'$ and defining $G=Q'X_2X_1'P$, an s.o. matrix, we have

$$\text{g}(X_1, X_2) = \text{tr}X_1'AX_2 = \text{tr}X_1'PDQ'X_2 = \text{tr}Q'X_2X_1'PD \leq d_1 + \ldots + d_q \qquad (56)$$

due to the fact that rank $(G) \leq q$. Thus, an upper bound to g has been found. Again, the upper bound is attained for a G as in (53), that is, for

$$X_1 = P_q N; \;\; X_2 = Q_q N, \qquad (57)$$

with P_q defined as the matrix of the first q columns of P, N an arbitrary orthonormal $q \times q$ matrix, and Q_q defined analogously to P_q. This settles the maximum of the generalized bilinear form (55), subject to orthonormality constraints.

It is instructive to consider, for each of the three maximum problems dealt with, the special case where the argument matrices have only one column. Then we are back to the corresponding vector function problems. Obviously, the general solutions derived here from the generalized Kristof theorem are also valid for q=1, so in this case they must coincide with the previously derived solutions that were summarized at the end of section 2.4, see Exercise 30.

The trace maximization problems treated in this section are indirectly related to least squares problems (see Exercise 39). The next three sections deal with problems that are least squares problems by definition.

3.4. UNCONSTRAINED MATRIX REGRESSION PROBLEMS

The most general type of matrix regression problem is the Penrose regression problem, already mentioned above. This problem is to minimize

$$\mathrm{h}(X) = \|AXB' - Y\|^2 = \mathrm{tr}(AXB' - Y)'(AXB' - Y) \qquad (58)$$

with A, B, and Y fixed. Only the case where both ($A'A$) and ($B'B$) have an inverse will be treated. The major step toward a solution is to write

$$\mathrm{h}(X) = \|(A'A)^{1/2}X(B'B)^{1/2} - (A'A)^{-1/2}A'YB(B'B)^{-1/2}\|^2 + c \qquad (59)$$

where $c = \mathrm{tr}Y'Y - \mathrm{tr}A(A'A)^{-1}A'YB(B'B)^{-1}B'Y'$, a constant, independent of X. Although (59) is quite difficult to discover, its validity is easy to verify, by expanding (58) and (59) in single terms.

Next, it can be seen from (59) that $\mathrm{h}(X) \geq c$, which is a lower bound to $\mathrm{h}(X)$. It is attained when the part of the function that depends on X vanishes, that is, if $(A'A)^{1/2}X(B'B)^{1/2}=(A'A)^{-1/2}A'YB(B'B)^{-1/2}$. The minimizing X is therefore

$$X = (A'A)^{-1}A'YB(B'B)^{-1} , \qquad (60)$$

as has been shown by Penrose (1956).

Now consider the special case where $B = I$. Because (60) is the solution for *any B,* we are free to substitute $B=I$ in (60), to find

$$X = (A'A)^{-1}A'Y \,. \tag{61}$$

So (61) gives the minimizing X for the function

$$\mathrm{h}^*(X) = \|AX - Y\|^2. \tag{62}$$

It is a direct generalization of (29).

Finally, consider the yet simpler special case where both $A=I$ and $B=I$. That is, consider the function

$$\mathrm{h}^{**}(X) = \|X - Y\|^2. \tag{63}$$

This function, which can be called a *matrix fitting* function, is hardly interesting, because the minimum is 0 and is attained trivially when we take $X = Y$. The function $\mathrm{h}^{**}(X)$ does become useful, however, once it is constrained by restrictions. Constrained regression problems will be taken up in the sections to follow.

3.5. MATRIX REGRESSION SUBJECT TO ORTHONORMALITY CONSTRAINTS

This section deals with h, h* and h** subject to the constraint that the argument X ($p \times q$) satisfies the constraint $X'X = I_q$. This time we *start* with the simple special case of matrix fitting, where $A = I$ and $B = I$. We seek the minimum of

$$h^{**}(X) = \|X - Y\|^2 = \text{tr}X'X - 2\text{tr}X'Y + \text{tr}Y'Y \qquad (64)$$

subject to $X'X=I_q$. It should be noted that, for every X which satisfies the constraint, tr$X'X$ = trI_q = q, a constant. Because tr($Y'Y$) is also constant, we need the minimum of -2tr$X'Y$, hence the maximum of tr$X'Y$, subject to the constraint $X'X = I_q$. This is a trace function problem, that has been solved above, see (49) and (50). The SVD $Y = PDQ'$ yields $X = PQ'$ as the optimizing solution. This solution gives the best columnwise orthonormal approximation (X) to a given matrix (Y).

Next, we consider the slightly more complicated case of matrix regression, and seek the minimum of

$$h^*(X) = \|AX - Y\|^2 = \text{tr}X'A'AX - 2\,\text{tr}X'A'Y + \text{tr}Y'Y, \qquad (65)$$

subject to $X'X = I_q$. This problem is well-known as the orthogonal Procrustes rotation problem. When A has the same number of columns as Y, then X is a $q \times q$ matrix, and hence $XX' = I_q$, whence tr$X'A'AX$ = tr$XX'A'A$ = tr$A'A$, a constant. Clearly, so is tr$Y'Y$. The remaining problem is again of a familiar form. That is, we need the maximum of tr$X'A'Y$ subject to $X'X = I$. Again, the solution is $X = PQ'$, but this time the relevant SVD reads $A'Y = PDQ'$.

When X is a *vertical* matrix, with $p > q$, there is no direct solution for the minimum of $h^*(X)$. An iterative solution for this problem has been given by Green and Gower (1979), and can be found in Gower (1984) and Ten Berge and Knol (1984).

For the minimum of the Penrose function h(X) subject to $X'X = I$ there is no explicit solution either. Iterative solutions for the minimum can be found in Mooijaart and Commandeur (1990) and Koschat and Swayne (1991).

3.6. MATRIX REGRESSION SUBJECT TO RANK CONSTRAINTS

The constraints imposed so far have been concerned with columnwise orthonormality of the argument matrix. In the present section, regression problems will be treated where the argument matrix is constrained to have a (low) rank r. This constraint is first applied to the matrix fitting problem, where we seek the minimum of

$$\mathrm{h}^{**}(X) = \|X - Y\|^2 \qquad (66)$$

subject to the constraint that X ($p\times q$) be of rank r or less, with $p\geq q>r$. When X has rank r, there is an r-dimensional orthonormal basis for the column space of X. Let this basis be represented by the columns of a $p\times r$ matrix U, with $U'U = I_r$. Every column of X is a linear combination of the columns of U. Hence, there is an $r\times q$ matrix V such that

$$X = UV. \qquad (67)$$

Because every X of rank r can be written as UV for some V and some V, the function h^{**} can be transformed into the equivalent function

$$\tilde{\mathrm{h}}(U,V) = \|UV - Y\|^2. \qquad (68)$$

This function has to be minimized over pairs of matrices U and V, where U is constrained by $U'U = I_r$, and V is unconstrained. The minimizing pair U, V for (68) will have a product of rank r or less that is the minimizing X for (66). We now have to find the minimizing pair U, V for (68). This will be done by means of argument reduction.

No matter what the optimal U is, the optimal V for $\tilde{\mathrm{h}}$ can be expressed in terms of this U, and Y, as

$$V = (U'U)^{-1}U'Y = U'Y, \tag{69}$$

see (61). Using this expression in (68) to *eliminate* V, all that remains to be done is finding the minimizing U for

$$\tilde{\tilde{\mathrm{h}}}(U) = \|UU'Y - Y\|^2 = \mathrm{tr}UU'YY'UU' - 2\mathrm{tr}Y'UU'Y + \mathrm{tr}Y'Y$$

$$= \mathrm{tr}U'YY'U - 2\mathrm{tr}U'YY'U + \mathrm{tr}Y'Y = \mathrm{tr}Y'Y - \mathrm{tr}U'YY'U\,. \tag{70}$$

In (70), tr$Y'Y$ is constant. Therefore, we seek the U which maximizes tr$U'YY'U$ subject to the constraint $U'U = I_r$. This is a generalized quadratic form problem, see (51), that was solved above, see (54). Accordingly, the optimal U is

$$U = P_rN \tag{71}$$

where P_r contains the first r columns of the left hand singular vectors matrix of $Y = PDQ'$ (SVD), and N is an arbitrary orthonormal $r\times r$ matrix. From (69) it appears that $V=N'P_r'Y$. It follows that the minimum of (66) is attained for

$$X=UV=P_rNN'P_r'Y=P_rP_r'PDQ'=P_r(I_r|0)DQ'=P_r(D_r|0)Q'=P_rD_rQ_r', \tag{72}$$

a result that goes back to Eckart and Young (1936).

It should be noted that P_r contains the first r columns of P, Q_r contains the first r columns of Q, and that D_r is the upper left $r\times r$ submatrix of D, but nevertheless it is not allowed to write Y as $Y=P_rD_rQ_r'$, as in (4). This is because r referred to the rank of Y in (4), whereas it now stands for the rank of the matrix that is the best rank-r approximation to Y.

The solution (72) for the minimum of h^{**} subject to a rank-r constraint can be used to minimize the Penrose regression function

$$\text{h}(X) = \|AXB' - Y\|^2 \tag{73}$$

subject to the same constraint. Assuming that both ($A'A$) and ($B'B$) have an inverse, we can write

$$\begin{aligned}\text{h}(X) &= \|(A'A)^{-½}A'YB(B'B)^{-½} - (A'A)^{½}X(B'B)^{½}\|^2 + c \\ &= \|W - (A'A)^{½}X(B'B)^{½}\|^2 + c\,,\end{aligned} \tag{74}$$

where $W \equiv (A'A)^{-½}A'YB(B'B)^{-½}$ and c = tr $Y'Y$ – tr $W'W$, both constant. Using the SVD $W = PDQ'$, we know that the best rank r approximation to W is $P_rD_rQ_r'$, see (72), and that the rank of the matrix $(A'A)^{½}X(B'B)^{½}$ is at most r. Combining these two facts yields the lower bound

$$\text{h}(X) \geq \|W - P_rD_rQ_r'\|^2 + c\,. \tag{75}$$

This lower bound is attained for the rank r matrix

$$X = (A'A)^{-1/2}P_rD_rQ_r'(B'B)^{-1/2} \tag{76}$$

which gives the minimum of h(X) subject to a rank r constraint. The same solution has been found by different means by Takane and Shibayama (1991). They also discuss a practical application.

Once a general solution has been obtained, it can be used to handle all special cases of it. In the present context, we have not yet discussed the regression problem of minimizing the function $\text{h}^*(X) = \|AX - Y\|^2$, subject to the constraint that X be of rank r at most. It is the special case of (73) where $B=I$. The solution can be found from (76) as

$$X = (A'A)^{-1/2}P_rD_rQ_r', \tag{77}$$

where P_r, D_r and Q_r' are defined by the SVD $(A'A)^{-1/2}A'Y = PDQ'$.

The regression problems and their solutions are summarized in Table 1. A review of some of these problems can also be found in Rao (1980). The problems for which references are given have no closed-form solutions and require an iterative procedure. As before, a matrix is called *vertical* when it has more rows than columns.

Table 1: *Matrix regression problems and their solutions for X of order* $p\times q$*;* $p\geq q$*.*

Problem	Unconstrained (see 3.4)	$X'X=I_q$ (see 3.5)	rank $(X)\leq r$ (see 3.6)
Minimize $h(X)=\|\|AXB'-Y\|\|^2$	$X=(A'A)^{-1}A'YB(B'B)^{-1}$	Koschat & Swayne or Mooijaart & Commandeur	$X=(A'A)^{-½}P_rD_rQ_r'(B'B)^{-½}$ using the SVD $(A'A)^{-½}A'YB(B'B)^{-½}=PDQ'$
↓ The special case $B = I$			
Minimize $h^*(X)=\|\|AX-Y\|\|^2$ (Regression	$X=(A'A)^{-1}A'Y$	Green & Gower (X vertical) or by $X=PQ'$ using the SVD $A'Y=PDQ'$ (X square)	$X=(A'A)^{-½}P_rD_rQ_r'$ using the SVD $(A'A)^{-½}A'Y=PDQ'$
↓ The special case $A=I$ and $B=I$			
Minimize $h^{**}(X)=\|\|X-Y\|\|^2$	$X=Y$	$X=PQ'$ using the SVD $Y=PDQ'$	$X=P_rD_rQ_r'$ using the SVD $Y=PDQ'$

Table 1 constitutes, together with the three trace maximizing solutions of 3.3, the framework of this book. In Chapter 4, the solutions will be applied without any further concern with the proofs of Chapter 3. Instead, all efforts of Chapter 4 are directed at reformulating the least squares problems, inherent to nine methods of Multivariate Analysis, in a form that can be handled by means of the solutions from the framework.

CHAPTER 4

APPLICATIONS

4.1. MULTIPLE REGRESSION ANALYSIS

Multiple linear regression analysis is concerned with estimating a given vector **y** by a linear combination $A\mathbf{x}$ of the columns of a given matrix A. The standard approach, which has already been discussed in previous sections, is based on minimizing the sum of squared errors

$$\mathrm{h}(\mathbf{x}) = ||\mathbf{y} - A||^2 = (\mathbf{y} - A\mathbf{x})'(\mathbf{y} - A\mathbf{x}) \qquad (78)$$

as a function of **x**. The solution (see section 2.3) reads

$$\mathbf{x} = (A'A)^{-1}A'\mathbf{y}. \qquad (79)$$

Multiple regression is often applied to variables (columns of A, and **y**) with means zero, and afterwards the correlation between **y** and $A\mathbf{x}$, known as the *multiple correlation*, is inspected, as a measure of success of the estimation of **y** from A. This policy can be justified by the fact that the *minimizing* solution (79) for the least squares problem (78) also *maximizes* the multiple correlation. This can be proved as follows.

First, express the correlation $r(\mathbf{y}, A\mathbf{x})$, with both **y** and the columns of A in deviations from their means, as a function of **x**:

$$\mathrm{f}(\mathbf{x}) = n^{-1}\mathbf{y}'A\mathbf{x}(n^{-1}\mathbf{y}'\mathbf{y})^{-1/2}(n^{-1}\mathbf{x}'A'A\mathbf{x})^{-1/2} = \frac{\mathbf{x}'A'\mathbf{y}}{(\mathbf{y}'\mathbf{y})^{1/2}(\mathbf{x}'A'A\mathbf{x})^{1/2}} \qquad (80)$$

Next, an upper bound to f can be found. The numerator can be written as

$$\mathbf{x}'A'\mathbf{y} = (\mathbf{x}'A)'(A(A'A)^{-1}A'\mathbf{y}) \leq (\mathbf{x}'A'A\mathbf{x})^{1/2}(\mathbf{y}'A(A'A)^{-1}A'\mathbf{y})^{1/2} \tag{81}$$

using Schwarz. Combining (80) and (81) yields

$$\begin{aligned} \mathrm{f}(\mathbf{x}) = \frac{\mathbf{x}'A'\mathbf{y}}{(\mathbf{x}'A'A\mathbf{x})^{1/2}(\mathbf{y}'\mathbf{y})^{1/2}} &\leq \frac{(\mathbf{x}'A'A\mathbf{x})^{1/2}(\mathbf{y}'A(A'A)^{-1}A'\mathbf{y})^{1/2}}{(\mathbf{x}'A'A\mathbf{x})^{1/2}(\mathbf{y}'\mathbf{y})^{1/2}} \\ &= \frac{(\mathbf{y}'A(A'A)^{-1}A'\mathbf{y})^{1/2}}{(\mathbf{y}'\mathbf{y})^{1/2}} \end{aligned} \tag{82}$$

which is an upper bound to f(**x**). The upper bound is attained when $A\mathbf{x}$ is proportional to $A(A'A)^{-1}A'\mathbf{y}$, that is, for instance, when

$$\mathbf{x} = (A'A)^{-1}A'\mathbf{y} . \tag{83}$$

This **x**, accordingly, does not only minimize the function h(**x**) of (78), but also maximizes the correlation f(**x**) between $A\mathbf{x}$ and **y**. This justifies the use of the multiple correlation as a measure of success of the estimation of **y** by $A\mathbf{x}$.

It is a well-known limitation of linear regression analysis that it is useless unless the number of persons (n) is far larger than the number of predictors (p), when A is a $n\times p$ matrix. Otherwise, using the optimal **x** from a first sample in another sample is likely to reveal *shrinkage* of the multiple correlation in the second sample. In other words, the multiple correlation has a positive sampling bias, and can be far too high when the sample size n is not much larger than p. The extreme case of $n=p$ shows this quite clearly. When $n=p$, A is a square matrix, and will have an inverse, if the assumption of zero means is ignored for the moment. Therefore, the function (78) will have a trivial minimum of zero, which is attained for $\mathbf{x} = A^{-1}\mathbf{y}$. Also, the multiple correlation will be 1, in that case. Compared to other techniques to be discussed in the sequel, multiple linear regression analysis is excessively sensitive to chance capitalization. That is, the technique is highly liable to yield unduly favorable results when applied to small samples.

So far, the multiple linear regression problem has been treated for the case of a *single* criterion variable **y**. The same principles, and, in fact, the same solution, apply when it is a *matrix* Y that is to be estimated by linear combinations of the columns of a matrix A. Every column of Y is independently estimated by the corresponding column of AX. For each column of X, we use (79) to find the optimal solution. As a result, the complete solution for X can be written as $X = (A'A)^{-1}A'Y$, see section 3.4. This solution maximizes a whole set of multiple correlations independently.

4.2. PRINCIPAL COMPONENTS ANALYSIS (PCA)

Principal Components Analysis (PCA) is a popular technique. For a given $n\times k$ matrix of standard scores Z, with correlation matrix $R = n^{-1}Z'Z$, it is the purpose of PCA to find a $k\times q$ matrix of weights B ($q<k$), yielding linear combinations (columns of ZB) that optimally summarize the information contained in Z. Specifically, we search for those q linear combinations from which Z can be reconstructed as accurately as possible in terms of the least squares criterion. To refer to ZB, we use the symbol F ($n\times q$). It is desired to minimize the function

$$\ell(B,A') = \|Z - ZBA'\|^2 = \|Z - FA'\|^2, \tag{84}$$

where B defines the linear combinations, and A' provides the optimal weights for estimating Z from ZB. Without loss of optimality, the constraint is imposed that F has standardized and uncorrelated columns. That is, F must satisfy the constraint $n^{-1}F'F = B'RB = I_q$. This is a constraint on B rather than on F, because constraints always pertain to the *argument* of a function.

To minimize ℓ, we use argument reduction, by eliminating A'. Regardless of B, the associated A' must be optimal for (84), and therefore must be of the form $A'=(F'F)^{-1}F'Z$, see (61). Using this expression for A leaves us with minimizing

$$\hat{\ell}(B) = ||Z-ZB(F'F)^{-1}F'Z||^2 = ||Z-ZB(nB'RB)^{-1}B'Z'Z||^2 \tag{85}$$
$$= ||Z-ZBn^{-1}B'Z'Z||^2 = ||Z-ZBB'R||^2$$

where the constraint $B'RB = I_q$ and the definition $R = n^{-1}Z'Z$ have been used. This function can be expanded as

$$\hat{\ell}(B) = \text{tr}Z'Z - 2\text{tr}Z'ZBB'R + \text{tr}RBB'Z'ZBB'R$$
$$= \text{tr}Z'Z - 2n\text{tr}B'R^2B + n\text{tr}B'R^2B = nk - n\text{tr}B'R^2B. \tag{86}$$

Clearly, the minimum of $\hat{\ell}(B)$ subject to $B'RB = I_q$ coincides with the maximum of tr$B'R^2B$, subject to the same constraint. Let $X=R^{½}B$. Then we need the maximum of $g(X) = \text{tr}X'RX$ subject to $X'X=B'RB=I_q$. We have thus arrived at a well-known problem, with solution $X = K_qN$, where K_q contains the first q columns of K, defined by the eigendecomposition $R=K\Lambda K'$, and N is an arbitrary orthonormal $q\times q$ matrix (see section 3.3). So the B wanted is $R^{-1/2}K_qN = K_q\Lambda_q^{-1/2}N$, where Λ_q is the upper left submatrix ($q\times q$) of Λ, and N an arbitrary rotation matrix. Here and elsewhere, the symbol $R^{-1/2}$ reveals that R is assumed to be nonsingular.

In addition to this derivation of principal components as those columns of F for which the optimal (least squares) reconstruction of Z is optimal, a second approach can be encountered in the literature. In that approach, the orthonormality constraint is not imposed on $n^{-½}F$, with $n^{-1}F'F = I_q$, but on B directly, and it is not the residual variance $||Z - FA'||^2$ that is minimized, but the sum of variances of the columns of F itself is maximized. Specifically, this approach is aimed at maximizing the function

$$g(B) = \text{tr}n^{-1}F'F = n^{-1}\text{tr}B'Z'ZB = \text{tr}B'RB \tag{87}$$

subject to the constraint $B'B = I_q$.

This is again a well-known problem, with $B = K_qN$ as the optimal solution. If we disregard the rotation N, we arrive at components that are proportional to those obtained earlier from the first approach to PCA. This means that no harm is done by using the second approach. Nevertheless, it is undesirable to maximize the variance of the components rather than the variance explained by the components, because only the latter is relevant for the purpose of finding components that summarize the information contained in the variables.

Above, we have derived the PCA solution by defining the components as linear combinations of the variables, see (84). In fact, this constraint is inactive, as can be seen from Exercise 43.

4.3. SIMULTANEOUS COMPONENTS ANALYSIS IN TWO OR MORE POPULATIONS

When the same variables have been administered in two or more groups of respondents from, for instance, different populations, the question may rise of how to generalize PCA. An elegant generalization of PCA is Simultaneous Components Analysis (SCA). The SCA method is aimed at finding one common $k \times q$ matrix B of weights, defining a matrix of components $Z_iB = F_i$ in each group, $i=1,...,m$. Because both variables and weights per component are the same across groups, so are the interpretations of the components. This leaves the importance of the components still free to vary across groups. In SCA, the optimal B is defined by the criterion that the sum of explained variances over groups should be as high as possible. In algebraic terms, SCA is aimed at minimizing the function

$$\ell(B, P_1, ..., P_m) = \sum_{i=1}^{m} ||Z_i - F_iP_i'||^2 = \sum_{i=1}^{m} ||Z_i - Z_iBP_i'||^2 \tag{88}$$

where P_i' is the *pattern* matrix of regression weights that yields the best least squares reconstruction of Z_i from $F_i = Z_iB$.

If we were to follow the same approach as in the derivation of PCA, we would start by eliminating P_i, by expressing it in terms of B and R_i as

$$P_i' = (F_i'F_i)^{-1}F_i'Z_i = (B'R_iB)^{-1}B'R_i. \qquad (89)$$

Substituting this expression for P_i' in (88) would reduce the problem to that of maximizing $\sum_{i=1}^{m} \text{tr}B'R_i^2B(B'R_iB)^{-1}$, a function of B only. Unfortunately, the maximum of this function is unknown. Therefore, eliminating $P_1,...,P_m$ is of no avail here. Eliminating B does not help either.

We are now facing a problem that has not been encountered so far. It is not possible to find an explicit minimum for the function ℓ. 'To be sure, we do know how to choose P_i when B is known. In fact, it even appeared possible to express that choice explicitly in terms of B and R. But it is not clear how to minimize the resulting function of B. We shall have to settle for something less than an explicit globally minimal solution, resulting from an attainable lower bound. In the present case, we shall settle for an *alternating least squares* (ALS) algorithm.

An ALS algorithm is an iterative procedure that starts by choosing *arbitrary* initial values for the complete set of argument parameters of the function. In the SCA context, we choose arbitrary values for B and $P_1,...,P_m$. The next step is to treat *a subset of the argument variables* temporarily as constant, for instance, the elements of B, and minimize the function by choosing the optimal values for the remaining argument variables, i.e., the elements of P_1 through P_m, given the current B. As a result, the function value will usually be lower than it was before. It will never be higher than before, because the *optimal* values for $P_1,...,P_m$ cannot do worse than their *previous values.*

In the next step, the roles are reversed. The just updated $P_1,...,P_m$ are now treated as fixed, and B is replaced by the best B, given these $P_1,...,P_m$. This step will also decrease the function value. This procedure can be repeated as often as we like, and will decrease the function value by alternately optimizing B for fixed $P_1,...,P_m$ and vice versa. How long we continue this procedure depends on the resulting function values. Usually, the first updates give major improvements, after which the improvements gradually diminish. When the improvements of the function value become smaller than a previously determined threshold value, e.g., .0001, the iterations are terminated. More often than not, we will end up close to the (global) minimum. But this is not guaranteed. An ALS algorithm may get caught in a *local minimum.* This danger can be countered in two ways. When the initial estimates of the argument values are cleverly chosen, to the effect that the function value starts already close to the global minimum, then it is impossible to converge to those local minima for which the function value is *above* the current one. This is because an ALS algorithm always moves in the *right direction,* and can only *decrease* the function value.

The second remedy against local minima is to try *several* (random) starting points, successively. The ALS procedure will yield as many solutions as starting points used, and the resulting function values (after convergence) can be compared. Only the solution associated with the lowest function value is maintained. The more starting points are used, the higher is the probability of finding the global minimum in this way.

An ALS algorithm is comparable to a procedure one might use to choose a sofa and a carpet, to optimize the atmosphere of a living room. That is, we first buy a sofa. Next, we acquire the best carpet to go with this particular sofa. Having replaced the carpet, we buy the best sofa to go with this particular carpet, and so on. Each purchase improves the atmosphere. Although we cannot be sure to attain the optimal sofa-carpet combination, we are likely to do well if the very first sofa is a clever choice, or if we repeat the entire procedure for a variety of randomly chosen initial sofas. It may seem wasteful to buy so many carpets and sofas where only one pair is needed, but in computational problems such wasteful procedures are well worth their cost, being only a matter of computer time.

At this point we return to the SCA problem, for which an ALS algorithm can be derived. We have already seen in (89) how to optimize $P_1,...,P_m$ for an arbitrary fixed *B*. The remaining question is how to optimize *B*, for fixed values of $P_1,...,P_m$. The answer comes from Vec notation (see section 1.4).

For fixed $P_1,...,P_m$, the problem is to choose *B* such that

$$\hat{\ell}(B) = \sum_{i=1}^{m} \left\| Z_i - Z_i B P_i' \right\|^2 \tag{90}$$

is a minimum. This function can also be written as

$$\hat{\ell}(B) = \sum_{i=1}^{m} \left\| \mathrm{Vec}(Z_i - Z_i B P_i') \right\|^2$$

$$= \sum_{i=1}^{m} \left\| \mathrm{Vec}(Z_i) - \mathrm{Vec}(Z_i B P_i') \right\|^2$$

$$= \sum_{i=1}^{m} \left\| \mathrm{Vec}(Z_i) - (P_i \otimes Z_i)\,\mathrm{Vec}(B) \right\|^2$$

$$= \left\| \begin{pmatrix} \mathrm{Vec}(Z_1) \\ . \\ . \\ . \\ \mathrm{Vec}(Z_m) \end{pmatrix} - \begin{pmatrix} P_1 \otimes Z_1 \\ . \\ . \\ . \\ P_m \otimes Z_m \end{pmatrix} \mathrm{Vec}(B) \right\|^2 . \tag{91}$$

From this, the optimal Vec(*B*) can be found at once as a solution for a simple regression problem (see section 2.3). Hence it is clear how to find the optimal *B* for fixed $P_1,...,P_m$. That is, if Vec(*B*) is known, then so is the $k \times q$ matrix *B*. All it takes to get *B* is to rearrange the elements of Vec(*B*) into a matrix of the proper order.

Iteratively updating *B* by (91) and $P_1,\ldots,P_m$ by (89) in an Alternating Least Squares algorithm reduces $\ell(B,P_1,...,P_m)$ monotonically. The algorithm converges to a stable function value because the function is bounded from below, for instance, by zero. Details of the present and other ALS algorithms for SCA can be found in Kiers and Ten Berge (1989). The idea of SCA was introduced by Millsap and Meredith (1988). Sampling properties of SCA have been examined by Ten Berge, Kiers, and Van der Stel (1992).

It is important to note that the resulting B and $P_1,...,P_m$ for SCA are not uniquely determined. That is, taking BT and $P_1(T^{-1})',...,P_m(T^{-1})'$, where T is an arbitrary nonsingular matrix, yields, as an estimate of Z_i in the i-th group, $Z_iBTT^{-1}P_i' = Z_iBP_i'$. This shows that the estimates, hence, the explained variance, remain the same, regardless of what transformation T is used. The freedom of transformation can be used in various ways. One way is to ensure that, in the i-th group, F_i is orthogonal and standardized. Alternatively, these properties can be implemented in

$$F = \begin{pmatrix} F_1 \\ \cdot \\ \cdot \\ \cdot \\ F_m \end{pmatrix}. \tag{92}$$

A third option is to transform B so as to satisfy certain ‘simple structure criteria’. A computer program for SCA, designed by Kiers (1990), allows for all of these transformations.

Finally, it should be pointed out that the loss incurred for the matrices $Z_1\ldots,Z_m$, respectively, can be weighted differentially by inserting weights into (88). For instance, one might weight the loss pro rato to the number of respondents involved. In the limiting case where B is determined on the basis of one group only, by giving zero weights to the residual variances for the other groups, SCA amounts to doing a PCA in one group, with cross-validation of the resulting components in the remaining groups, see Ten Berge (1986).

4.4. MINRES FACTOR ANALYSIS

The classical formulation of factor analysis is as follows. Given a $k\times k$ correlation matrix R, a $k\times q$ matrix A is wanted such that the off-diagonal elements of AA' resemble those of R as much as possible in the least squares sense. Accordingly, it is desired to minimize the function

$$\ell(A) = \sum_{i>j} (r_{ij} - \mathbf{a}_i'\mathbf{a}_j)^2, \qquad (93)$$

where $\mathbf{a}_i'$ is the i-th row of A. An ALS algorithm to minimize ℓ has been designed by Harman and Jones (1966). It starts with an arbitrary $k\times q$ matrix A, and alternately improves each row $\mathbf{a}_i'$, for fixed values of the other rows. To do this, one needs to know how to update any single row of A. As a matter of convenience, the discussion will be limited to $\mathbf{a}_1'$, the first row of A. We slice $\ell(A)$ up into a part that depends on $\mathbf{a}_1'$, and a part that does not vary with $\mathbf{a}_1$ and can be treated as fixed, for the time being. Specifically, define

$$\ell_1(\mathbf{a}_1) = (r_{21} - \mathbf{a}_2'\mathbf{a}_1)^2 + (r_{31} - \mathbf{a}_3'\mathbf{a}_1)^2 + ... + (r_{k1} - \mathbf{a}_k'\mathbf{a}_1)^2 + c_1, \qquad (94)$$

where c_1 is the sum of all terms of ℓ(A) in which $\mathbf{a}_1$ does not occur. Let $\mathbf{r}_1$ be the first column of R, the diagonal element excluded, so $\mathbf{r}_1' = (r_{21},...,r_{k1})$. Define A_1 as the $(k–1)\times q$ matrix that is left when the first row of A ($k\times q$) is deleted. Then $\ell_1(\mathbf{a}_1)$ can also be written as

$$\ell_1(\mathbf{a}_1) = \|\mathbf{r}_1 - A_1\mathbf{a}_1\|^2 + c_1. \qquad (95)$$

Minimizing this is a standard regression problem, and has as solution

$$\mathbf{a}_1 = (A_1'A_1)^{-1}A_1'\mathbf{r}_1. \qquad (96)$$

This solution yields the best possible first row of A, given the other rows of A. When the previous version of that row is updated according to (96), $\ell(A)$ will decrease.

Similarly, for any arbitrary row $\mathbf{a}_i'$ of A, the best update, given the other rows, is

$$\mathbf{a}_i = (A_i'A_i)^{-1}A_i'\mathbf{r}_i, \tag{97}$$

where A_i is what is left of the current A upon deleting its i-th row, and $\mathbf{r}_i$ is the i-th column of R, the diagonal element excluded. It has thus been shown how to decrease $\ell(A)$ iteratively, by updating each row of A in turn. This ALS algorithm is called MINRES. A different algorithm, which also decreases (93) iteratively, is the ULS (unweighted least squares) algorithm, proposed by Jöreskog (1977). It will not be treated here.

MINRES factor analysis has lost the popularity it once had. This is not only because many factor analysts nowadays prefer maximum likelihood methods, but also due to some technical deficiencies of MINRES itself. The first of these is that one or more rows of the matrix A that results from MINRES may have a sum of squares larger than one (a so-called Heywood case), which would imply that more than 100% of variance has been explained for the corresponding variable. This deficiency can easily be overcome by replacing the update (97) for $\mathbf{a}_i$, whenever $\mathbf{a}_i'\mathbf{a}_i > 1$, by that vector $\mathbf{a}_i$ that minimizes a function of the form (95) subject to the constraint $\mathbf{a}_i'\mathbf{a}_i = 1$, see Harman and Fukuda (1966). The solution for this problem has been discussed in section 2.4.

However, even when Heywood cases are absent, a more general problem with MINRES remains. It has to do with the separation of common variance in explained common variance, associated with those common factors that are maintained, and unexplained common variance, associated with those common factors that are discarded. When there is perfect fit, that is, when AA' equals the correlation matrix R (except for its diagonal elements), there is 100% explained common variance. In cases of imperfect fit, however, there is no meaningful way of evaluating the percentage of common variance that is explained by the factors which are maintained. This problem can also be overcome, see Ten Berge and Kiers (1991), but the solution is too involved to be treated here.

4.5. CANONICAL CORRELATION ANALYSIS

The data to which SCA can be applied can be referred to as *vertical* data. That is, SCA is applied to scores of different groups of respondents (rows) on the same variables (columns), as a result of which these data matrices can be arranged *one below the other* in one supermatrix. When, on the other hand, scores are available from one group of respondents on different sets of variables, the data can be called *horizontal:* The data matrices for different sets of variables can now be arranged *one next to the other* in one supermatrix, for instance, $(Z_1|Z_2)$ in the case of two sets of variables.

Canonical Correlation Analysis (CCA) requires *horizontal* data. Only the case of two data sets will be treated here, although more general approaches are widely available.

The purpose of CCA is to answer this question: Is it possible to construct certain linear combinations of the columns of Z_1 (of order $n \times k_1$) and of Z_2 ($n \times k_2$), such that they correlate pair wise as highly as possible? Let B_1 ($k_1 \times q$) be the matrix of weights which defines q linear combinations of Z_1, and let B_2 ($k_2 \times q$) define q linear combinations of Z_2, where q is the dimensionality of the solution, $q_1 \leq k_1$ and $q_2 \leq k_2$. Then we seek the maximum of the sum of the q correlations (canonical correlations) between the corresponding columns of Z_1B_1 and Z_2B_2. This is what CCA is about. The columns of Z_1B_1 and Z_2B_2 are called canonical variates of Z_1 and Z_2, respectively.

To avoid trivial results, the constraint is imposed that both the canonical variates of Z_1 and those of Z_2 are uncorrelated. For convenience, they are also standardized. Both constraints can be captured in the constraint $B_1'R_1B_1=B_2'R_2B_2=I_q$, where $R_1 = n^{-1}Z_1'Z_1$ and $R_2 = n^{-1}Z_2'Z_2$. Now the purpose of CCA can be expressed as maximizing the sum of the q canonical correlations, written as

$$\ell(B_1,B_2) = \text{tr}(n^{-1}B_1'Z_1'Z_2B_2) = \text{tr}B_1'R_{12}B_2, \tag{98}$$

subject to the constraint $B_1'R_1B_1=B_2'R_2B_2=I_q$, where $R_{I2}=n^{-1}Z_1'Z_2$ This maximum problem will now be transformed into an equivalent problem, with different argument matrices. Specifically, define $X_1 = R_1^{1/2}B_1$ and $X_2 = R_2^{1/2}B_2$. Then we want the maximum of

$$\text{g}(X_1, X_2) = \text{tr}X_1^{'}R_1^{-1/2}R_{12}R_2^{-1/2}X_2 \tag{99}$$

subject to the constraint $X_1'X_1 = X_2'X_2 = I_q$. The solution can be obtained at once from section 3.3. Assume that $k_1 \geq k_2$ (otherwise, switch Z_1 and Z_2) and define the SVD

$$R_1^{-1/2}R_{12}R_2^{-1/2} = PDQ'. \tag{100}$$

Then g attains its maximum subject to the constraint when we take

$$X_1 = P_qN \text{ and } X_2 = Q_qN \tag{101}$$

where P_q and Q_q contain the first q columns of P and Q, respectively, and N is an arbitrary orthonormal $q \times q$ matrix, see (57). Therefore, taking

$$B_1 = R_1^{-1/2}P_qN \text{ and } B_2 = R_2^{-1/2}Q_qN \tag{102}$$

solves the CCA problem.

Often, the orthonormal matrix N is chosen as I_q . Then the matrix of correlations between the canonical variates of Z_1 and Z_2 is diagonal. This follows from the expression

$$n^{-1}B_1'Z_1'Z_2B_2 = B_1'R_{12}B_2 = P_q'R_1^{-1/2}R_{12}R_2^{-1/2}Q_q = P_q'PDQ'Q_q = (I_q\,|\,0)D\left(\frac{I_q}{0}\right) = D_q \tag{103}$$

the diagonal $q \times q$ matrix containing the q largest singular values from D on the diagonal. Hence, each column of Z_1B_1 correlates zero with each column of Z_2B_2, the corresponding column excluded. The diagonal elements of D_q are the q canonical correlations.

In quite a number of textbooks, the solution for the CCA problem is derived in terms of eigenvectors. This is rather inconvenient, because canonical correlations then may turn out negative, and reflections of certain columns of B_1 or B_2 are needed to remedy this.

An intriguing question is what use CCA has. It was introduced as a method to find the most valid predictor (column of Z_1B_1) for the most valid criterion (column of Z_2B_2). This is indeed what CCA does. It is, however, far from clear what purpose is served by this. Although taking linear combinations of predictors is absolutely legitimate and useful, combining criterion variables linearly (or otherwise) is suspicious. Criterion variables should be determined by the investigator, rather than by a method of data analysis. Replacing them by linear combinations reflects doubt as to what should be predicted. In practical applications, it is unnatural to allow for linear combinations of the criterion variables, which leaves no place for CCA. Once one or more criteria have been selected, multiple linear regression analysis is wanted. The matrix $B_1 = R_1^{-1}R_{12}$ is computed as the matrix of weights that maximizes the sum of the (multiple, instead of canonical) correlations between the corresponding columns of Z_1B_1 and Z_2 (instead of Z_2B_2).

Having discarded CCA as a prediction method, it may still seem possible to salvage it as a form of Components Analysis, yielding summarizers of two sets of variables. This is because the columns of Z_1B_1 and of Z_2B_2 are components of Z_1 and Z_2, respectively. Clearly, CCA produces maximally correlated summarizers of Z_1 and Z_2. The problem with this interpretation of CCA is, that the summarizers obtained from CCA *can be very poor summarizers* of Z_1 and Z_2, respectively. They were derived as *highly correlated* components, which is totally different from *variance explaining* components. It can be concluded that, even from the Components Analysis point of view, it is not clear what the use of CCA is. This point will be revisited in section 4.6.

A related issue is that CCA cannot be viewed as a generalization of PCA. That is, when Z_1 coincides with Z_2, CCA does not necessarily yield principal components as canonical variates. In fact, in this case there is an infinite number of solutions for CCA, and the PCA solution is just one of them.

Although CCA has no direct relevance for practical data analysis, it can sometimes be encountered as a subroutine of a different method of data analysis. In such cases, the legitimacy of using CCA stems from the latter method rather than from CCA itself.

4.6. REDUNDANCY ANALYSIS

The liability of CCA to explain only little variance has led to the construction of an alternative technique, called Redundancy Analysis (RA), see, for instance, Van den Wollenberg (1977). This technique is aimed at finding those linear combinations of the columns of Z_1, that yield the best prediction of Z_2, by means of multiple linear regression analysis. Accordingly, RA is aimed at minimizing the function

$$\ell(B,W) = \|Z_1BW - Z_2\|^2. \tag{104}$$

Without loss of optimality, the constraint is imposed that the columns of Z_1B are orthogonal and standardized, that is, $B'R_1B = I_q$. Regardless of the choice of B, the optimal W depends on it by the relation

$$W = (B'R_1B)^{-1}B'R_{12} = B'R_{12}. \tag{105}$$

So the minimum of $\ell(B,W)$ can be found as the minimum of

$$\tilde{\ell}(B) = \left\| Z_1 BB' R_{12} - Z_2 \right\|^2$$

$$= n\mathrm{tr}R_{21}BB'R_1BB'R_{12} - 2n\mathrm{tr}R_{21}BB'R_{12} + n\mathrm{tr}R_2$$

$$= n(\mathrm{tr}R_2 - \mathrm{tr}B'R_{12}R_{21}B)\,, \qquad (106)$$

subject to the constraint $B'R_1B=I_q$. Clearly, this problem has the same solution as the problem of maximizing

$$\tilde{\mathrm{g}}(B) = \mathrm{tr}B'R_{12}R_{21}B\,, \qquad (107)$$

subject to $B'R_1B=I_q$. This constraint can be reformulated as the constraint that $R_1^{1/2}B = X$ has to be columnwise orthonormal. Therefore, we may equivalently maximize the function

$$\mathrm{g}(X) = \mathrm{tr}X'R_1^{-1/2}R_{12}R_{21}R_1^{-1/2}X \qquad (108)$$

subject to $X'X = I_q$. This is a generalized quadratic form problem. Define the SVD $R_1^{-1/2}R_{12} = PDQ'$. Then

$$R_1^{-1/2}R_{12}R_{21}R_1^{-1/2} = PD^2P' \qquad (109)$$

is an eigendecomposition, and the solution is $X = P_qN$, see (54). Hence

$$B = R_1^{-1/2}P_qN \qquad (110)$$

is the solution for the RA problem. The resulting estimate of Z_2 is

$$Z_1BW = Z_1R_1^{-1/2}P_qNN'P_q'R_1^{-1/2}R_{12} = Z_1R_1^{-1/2}P_qP_q'PDQ' = Z_1R_1^{-1/2}P_qD_qQ_q' \tag{111}$$

an expression to be discussed later.

It should be noted that the columns of Z_1B, as defined in RA, are standardized linear combinations, and hence components of the columns of Z_1. However, they should not be mistaken for the principal components of Z_1. This is because they explain as much variance as possible in the *external* matrix Z_2, rather than in the *internal* matrix Z_1. For this reason, RA is sometimes called 'External PCA'.

In the special case where Z_2=Z_1, RA and PCA coincide. Therefore, contrary to CCA, one may consider RA as a generalization of PCA.

Interpreting RA as external PCA can also be justified in a different manner. Specifically, if we were to find a matrix B yielding the highest sum of squared correlations between all columns of Z_1B and all columns of Z_2, subject to the constraint $B'R_1B=I_q$, we would end up with the same solution. To verify this, write the sum of squared correlations as

$$\left\| n^{-1}B'Z_1'Z_2 \right\|^2 = \left\| B'R_{12} \right\|^2 = \mathrm{tr}B'R_{12}R_{21}B = \tilde{g}(B)\,, \tag{112}$$

and note the equality to (107).

There are still other interpretations of RA. One of these will be explained here, namely, RA as a rank constrained version of regression analysis. It has been seen before that the best possible unconstrained (least squares) estimate of Z_2 by a set of linear combinations of the columns of Z_1 is obtained when Z_1 is postmultiplied by the matrix of regression weights $(Z_1'Z_1)^{-1}Z_1'Z_2 = R_1^{-1}R_{12}$. Suppose that the matrix of weights is constrained to have rank q ($q < k_1$; $q < k_2$). In that case the solution is different. Using the SVD $R_1^{-1/2}R_{12} = PDQ'$, as we did above (109), it follows from (77) that the optimal rank-q regression weights matrix is

$$R_1^{-1/2} P_q D_q Q_q^{'} . \qquad (113)$$

If Z_1 is postmultiplied by this matrix, the estimate of Z_2 will be exactly the same as the RA estimate, see (111). As a result, RA can also be seen as rank constrained linear regression analysis.

The interpretation of RA as rank constrained regression analysis reveals a severe limitation of the practical usefulness of RA. When, for instance, peer ratings of subjects on a set of traits (columns of Z_2) are to be predicted on the basis of certain scales of a personality questionnaire (Z_1), there is no point in adopting a rank constraint. It would result in a loss of predictive power, without any reason. However, there is a place for RA in prediction contexts when ordinary linear regression is very successful. In that case one may desire to replace the predictors (the k_1 columns of Z_1) by a parsimonious set of q linear combinations of them, contained in the matrix Z_1B. It is conceivable that only a small loss of predictive power will be incurred when this procedure is adopted, and RA provides the best set of the linear combinations to replace the full set of predictors.

Another application may arise when it is desired to understand what the major dimensions are in the predictable part of Z_2. This is because RA provides a low rank approximation to the regression of Z_2 on Z_1.

4.7. PARAFAC

Above, it has been seen that SCA is a method for vertical data, and that Multiple Regression Analysis, CCA, and RA require horizontal data. We now turn to data that are both horizontal and vertical, for instance, scores of n persons on m variables, on p occasions. Such data can be represented in a *three-way data array X*, similar to a loaf of bread, so to speak, consisting of p slices $X_1,...,X_p$, each of order $n \times m$.

Tucker (1966) has laid the foundations for three-way components analysis. Although efficient algorithms for this components analysis have been developed since, e.g., see Kroonenberg and De Leeuw (1980) or Kroonenberg (1983), the method has not become popular. The problem is that the method yields matrices of coefficients for persons, variables, and occasions, and a so-called (three-way) core array, containing weights for joint contributions of person components, variable components, and occasion components. For the practitioner, this is often too much information. For this reason, three-way components analysis in the Tucker tradition will not be treated here. Instead, only a simple variant of three-way PCA, called PARAFAC, will be considered here.

PARAFAC is an acronym for PARAllel FACtor Analysis, and has been designed by Harshman (1970). Let x_{ijk} be the score of person i on variable j at occasion k. Then the PARAFAC model reads

$$x_{ijk} = \sum_{h=1}^{r} a_{ih} b_{jh} c_{kh} + e_{ijk} \; , \tag{114}$$

where e_{ijk} is a residual term, a_{ih} is an element of an $n{\times}r$ matrix A, b_{jh} is an element of an $m{\times}r$ matrix B, and c_{kh} is an element of a $p{\times}r$ matrix C. Fitting the PARAFAC model amounts to minimizing the sum of squared residuals $\sum e_{ijk}^2$. Define D_k as the diagonal $r{\times}r$ matrix containing the elements of row k of C in the diagonal, (k=1,...,p), and define X_k as the k-th frontal slice of X, of order $n{\times}m$. Then the PARAFAC model can be written as

$$X_k = AD_kB' + E_k \quad (k=1,\ldots,p). \tag{115}$$

The notation used may conceal the fact that (115) can be interpreted as a generalized form of PCA. However, upon writing F instead of A and P_k' instead of D_kB', it appears that every X_k is decomposed as the product of a common component scores matrix F and a specific pattern matrix P_k', with E_k as a residual matrix. In the case p=1 we may set $D_1 = I$, whence PARAFAC yields an approximation of X_1 of the form $AIB'=AB'$, a job that might as well be done by PCA. In the case $p>1$, PARAFAC differs from PCA on $X_1,...,X_p$ separately, in two respects. First, the components matrix A in PARAFAC is the same across occasions. Second, the pattern matrices $BD_1,...,BD_p$ of PARAFAC are essentially the same. Basically, there is only one pattern matrix B, the columns of which are scaled differentially across the occasions, by $D_1,...,D_p$, respectively. This means that the overall contributions of components to the explained variance differs between and within occasions, but that the relative contributions of a specific component to the variables are the same across the occasions.

As always, it is up to the user to determine the number r of components. The residual sum of squares will, obviously, decrease as r increases. On the other hand, any increase of r implies a loss of parsimony of PARAFAC.

A computational solution for the matrices A, B, and $D_1,...,D_p$ in PARAFAC is obtained from least squares fitting. Specifically, PARAFAC is aimed at minimizing the function

$$\sum_{k=1}^{p}\|E_k\|^2 = \ell(A,B,D_1,...,D_p) = \sum_{k=1}^{p}\|X_k - AD_kB'\|^2 . \tag{116}$$

An ALS algorithm, called CANDECOMP (Carroll & Chang, 1970) is available for this purpose. It is based on alternately updating A, B, and $D_1,...,D_p$, keeping the other matrices fixed. First, initial values for A and B are chosen, and $D_1,...,D_p$ are updated independently. Updating D_k requires the minimum of that part of (116) that depends on it, that is, the minimum of

$$\ell_k(D_k) = \|X_k - AD_kB'\|^2$$

(117)

subject to the constraint that D_k be a diagonal matrix. This constraint implies that the Penrose solution of section 3.4 cannot be used. To find the minimum of ℓ_k, we write

$$\ell(D) = \| X - \sum_{i=1}^{r} \mathbf{a}_i d_i \mathbf{b}_i^{'} \|^2 , \tag{118}$$

where the subscript k has been omitted for simplicity. In (118), $\mathbf{a}_i$ is the i-th column of A, $\mathbf{b}_i$ the i-th column of B and d_i the i-th diagonal element of D. A sum of squared elements of a matrix does not change when these elements are rearranged in a vector. It follows that

$$\ell(D) = \| \mathrm{Vec}(X) - \mathrm{Vec}\sum_{i} (\mathbf{a}_i d_i \mathbf{b}_i^{'}) \|^2 \tag{119}$$

$$= \| \mathrm{Vec}(X) - \sum_{i} \mathrm{Vec}(\mathbf{a}_i d_i \mathbf{b}_i^{'}) \|^2$$

$$= \| \mathrm{Vec}(X) - \sum_{i} \mathbf{b} \otimes \mathbf{a}_i d_i \|^2 ,$$

where (19) and (23) have been used, as well as the fact that Vec(d_i) = d_i. Let U be the $nm \times r$ matrix, having the Kronecker products of the corresponding columns of B and A as columns, and collect $d_1,...,d_r$ in a vector $\mathbf{d}$. Then we have

$$\ell(D) = \| \mathrm{Vec}(X) - U\mathbf{d} \|^2 . \tag{120}$$

Clearly, searching for the minimizing D of (118) is equivalent to searching for the minimizing $\mathbf{d}$ of (120), because the diagonal elements of D are the very elements of $\mathbf{d}$. The optimal $\mathbf{d}$ can be found as $\mathbf{d} = (U'U)^{-1}U'\mathrm{Vec}(X)$, whence the optimal D has been determined. By subsequently using $X_1,...,X_p$ in (120), we obtain the optimal updates for $D_1,...,D_p$, respectively, keeping A and B fixed.

A next step in CANDECOMP is to update B', for fixed A and $D_1,...,D_p$. This is a straightforward application of multiple regression analysis. Upon writing

$$\mathrm{h}(B) = \sum_{k=1}^{p} \| X_k - AD_k B^{'} \|^2 = \left\| \begin{pmatrix} X_1 \\ . \\ . \\ . \\ X_p \end{pmatrix} - \begin{pmatrix} AD_1 \\ . \\ . \\ . \\ AD_p \end{pmatrix} B^{'} \right\|^2 \qquad (121)$$

it is clear that B' is a regression weights matrix, and must be taken as

$$B^{'} = (\sum_{k=1}^{P} D_k A^{'} A D_k)^{-1} (\sum_{l=1}^{p} D_l A^{'} X_l) . \qquad (122)$$

Analogously, it can be shown that, for fixed B and $D_1,...,D_p$, the best choice for A is obtained from

$$A^{'} = (\sum_{k=1}^{p} D_k B^{'} B D_k)^{-1} (\sum_{l=1}^{p} D_l B^{'} X_l^{'}) . \qquad (123)$$

This completes the basic ingredients for the CANDECOMP algorithm. Alternately, A, B, and $D_1,...,D_p$ are updated, and each update decreases (116). Clearly, CANDECOMP is an ALS algorithm, which decreases the function ℓ monotonically.

The component matrix A in PARAFAC is not constrained to be orthogonal. When orthogonal components are desired for PARAFAC, then CANDECOMP must be adjusted to satisfy this constraint. It will now be shown how to handle this. Only the update for A needs to be adjusted, by imposing the constraint $A'A = I$. This constraint renders the regression solution (123) invalid. To derive the constrained solution, write the part of (116) that depends on A as

$$\mathrm{h}(A) = \sum_{k=1}^{p} \left\| X_k - AD_k B^{'} \right\|^2 = \sum_{k=1}^{p} \left\| X_k^{'} - BD_k A^{'} \right\|^2$$

$$= \left\| \begin{pmatrix} X_1' \\ . \\ . \\ . \\ X_p' \end{pmatrix} - \begin{pmatrix} BD_1 \\ . \\ . \\ . \\ BD_p \end{pmatrix} A' \right\|^2 = \left\| G - HA' \right\|^2, \tag{124}$$

where G and H denote the two supermatrices pictured. It is desired to find the minimizing A for h(A), subject to $A'A=I$. This problem departs from the problem of section 3.5, in that the regression weights matrix A' is now constrained to be rowwise rather than columnwise orthonormal. Nevertheless, the solution can readily be found. Expanding h(A) as

$$\begin{aligned} \text{h}(A) &= \text{tr}(G - HA')'(G - HA') = \text{tr}GG' - 2\text{tr}G'HA' + \text{tr}AH'HA' \\ &= \text{tr}GG' - 2\text{tr}A'G'H + \text{tr}H'H = -2\text{tr}A'G'H' + c, \end{aligned} \tag{125}$$

with c constant, shows that we need to maximize f(A) = tr$A'G'H$ subject to $A'A=I$. This is a familiar problem, see section 3.3. The solution is $A = PQ'$, defined by the SVD $G'H = PDQ'$. It is clear from (124) that $G'H$ can be computed as $G'H = \sum_{k=1}^{p} X_k BD_k$. The SVD of this known matrix provides us with the optimal columnwise orthonormal update for A, when B and $D_1,...,D_p$ are kept constant.

PARAFAC is appropriate in cases where components are the same across occasions, but differ in importance. PARAFAC is a demanding model, that will more often than not be too restrictive for real life data. However, in those cases where it does fit nicely, it offers a highly elegant method of analyzing a three-way array.

4.8. INDSCAL

When a three-way array X consists of *symmetric* slices $X_1,\ldots,X_p$, that can be interpreted as *similarity matrices*, the INDSCAL model is often used. This model decomposes the slices as

$$X_k = AD_kA' + E_k \tag{126}$$

with $D_1,...,D_p$ diagonal and nonnegative. A situation where INDSCAL applies is, for instance, that where X_k is a symmetric matrix of similarities between n stimuli, as viewed by a subject k, k=l,...,p. The INDSCAL model is based on the premise that the subjects use the same matrix A of coordinates of the stimuli on q dimensions, but differ in the idiosyncratic weights (saliences) they attach to these dimensions. These weights are the diagonal elements of $D_1,...,D_p$. They should be nonnegative in most applications.

The problem of how to fit the INDSCAL model in the least squares sense amounts to minimizing $\sum_{k=1}^{p} \text{tr}E_k'E_k$, as defined implicitly in (126), as a function of A and $D_1,...,D_p$. This problem has not been solved directly. The most popular approach, due to Carroll and Chang (1970), is based on a technique called *splitting* (De Leeuw & Heiser, 1982). That is, the two appearances of A in (126) are represented by different matrices, A and B, say, which are optimized independently, along with $D_1,...,D_p$. Next, the CANDECOMP algorithm is applied to obtain estimates for the matrices A, B, and $D_1,...,D_p$. After convergence of CANDECOMP, it is *hoped* that the A and B obtained happen to be equal. In practice, this always appears to be the case. Also, one must *hope* that the solutions obtained for $D_1,...,D_p$ have nonnegative diagonal elements throughout. That is not always the case in practice. It is possible to avoid negative saliences by imposing nonnegativity constraints, but the technical implementation of this is rather laborious, see Ten Berge, Kiers and Krijnen (1993).

4.9. HOMOGENEITY ANALYSIS

Homogeneity Analysis, also known as Multiple Correspondence Analysis, is a generalization of PCA for qualitative variables. A qualitative variable can, in the present context, conveniently be represented by a *matrix* G_j, of order $n \times k_j$, where k_j is the number of categories of the variable j. Each of the n respondents has a score 1 for the category to which he/she belongs, and a score of zero otherwise. For instance, when five respondents have scores *a, b, a, b, c* on a qualitative variable j, this can be portrayed in a 5×3 matrix G_j as

$$G_j = \begin{matrix} a & b & c \\ \end{matrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{127}$$

A matrix of this kind is called an *indicator matrix.* In the case of m qualitative variables, there will be m indicator matrices involved. They can be collected in a supermatrix

$$G = (G_1|G_2|...|G_m) \tag{128}$$

of order $n \times k$, where $k = k_1 + k_2 + ... + k_m$, the total number of categories of the m variables.

Homogeneity Analysis is a technique which yields weights, collected in vectors $\mathbf{y}_1,...,\mathbf{y}_m$, to quantify the categories of the variables. As a result, quantified variables are constructed, of the form $G_1\mathbf{y}_1$, $G_2\mathbf{y}_2,...,G_m\mathbf{y}_m$. The weights are chosen so as to render the quantified variables as homogeneous as possible. That means that the quantified variables deviate as little as possible from a certain vector $\mathbf{x}$, in the least squares sense. Specifically, the weights and $\mathbf{x}$ are chosen to minimize

$$\ell(\mathbf{y}_1,...,\mathbf{y}_m,\mathbf{x}) = \sum_{j=1}^{m} \left\| G_j\mathbf{y}_j - \mathbf{x} \right\|^2 , \tag{129}$$

subject to the constraints $\mathbf{x}'\mathbf{x} = n$ and $\mathbf{1}'\mathbf{x} = 0$. Constraints have to be introduced, because without them, there would be *trivial* solutions $\mathbf{x}$=0 and $\mathbf{y}_j$=0, or $\mathbf{x}=\mathbf{1}_n$ and $\mathbf{y}_j=\mathbf{1}_{k_j}$, respectively, j=1,...,m.

One way to find the minimum of (129) subject to the constraint is as follows. Regardless of $\mathbf{x}$, the associated $\mathbf{y}_j$ (j=1,...,m) must satisfy

$$\mathbf{y}_j = (G_j'G_j)^{-1}G_j'\mathbf{x} = D_j^{-1}G_j'\mathbf{x}, \tag{130}$$

where D_j is defined as the diagonal matrix $G_j'G_j$, containing the category frequencies for the j-th variable on the diagonal. Using (130), the problem of minimizing (129) can be simplified to that of minimizing

$$\tilde{\ell}(\mathbf{x}) = \sum_{j=1}^{m} \left\| G_j D_j^{-1} G_j' \mathbf{x} - \mathbf{x} \right\|^2$$

$$= \mathbf{x}'(\sum_{j=1}^{m} G_j D_j^{-1} G_j')\mathbf{x} - 2\mathbf{x}'(\sum_{j=1}^{m} G_j D_j^{-1} G_j')\mathbf{x} + nm$$

$$= nm - \mathbf{x}'(\sum_{j=1}^{m} G_j D_j^{-1} G_j')\mathbf{x}\,. \tag{131}$$

The problem that remains is to maximize the quadratic form

$$\mathrm{g}(\mathbf{x}) = \mathbf{x}'(\sum_{j=1}^{m} G_j D_j^{-1} G_j')\mathbf{x} \tag{132}$$

subject to the constraints $\mathbf{x}'\mathbf{x} = n$ and $\mathbf{1}'\mathbf{x} = 0$. Clearly, the constraint $\mathbf{l}'\mathbf{x} = 0$ is equivalent to the constraint $J\mathbf{x}=\mathbf{x}$, where $J = (I - \frac{\mathbf{11}'}{n})$. Accordingly, we have

$$\mathrm{g}(\mathbf{x}) = \mathrm{g}(J\mathbf{x}) = \mathbf{x}'J(\sum_{j=1}^{m} G_j D_j^{-1} G_j')J\mathbf{x} = \mathbf{x}'(\sum_{j=1}^{m} JG_j D_j^{-1} G_j' J)\mathbf{x} \equiv \mathbf{x}'W\mathbf{x}$$

$$= n(\mathbf{x}/\sqrt{n})'W(\mathbf{x}/\sqrt{n}) \le n\lambda_1(W), \tag{133}$$

where W is defined as $W = \sum_{j=1}^{m} JG_jD_j^{-1}G_j^{'}J$, and $\lambda_1(W)$ is the largest eigenvalue of W. The upper bound of (133) is attained when we choose $\mathbf{x}$ as the first eigenvector of W, scaled to a sum of squares n. That eigenvector also satisfies both constraints. Clearly, we have thus found the minimum of (129).

The explicit approach above, based on eliminating $\mathbf{y}_j$ and then optimizing $\mathbf{x}$, is not the most attractive from a computational point of view. The best known computer program for Homogeneity Analysis, called HOMALS (Gifi, 1990), is an ALS algorithm. It proceeds by alternately updating $\mathbf{y}_1,...,\mathbf{y}_m$ for fixed $\mathbf{x}$ using (130), and updating $\mathbf{x}$ for fixed $\mathbf{y}_1,...,\mathbf{y}_m$ by minimizing the function

$$\mathrm{h}(\mathbf{x}) = \sum_{j=1}^{m} \|G_j\mathbf{y}_j - \mathbf{x}\|^2 \tag{134}$$

subject to the constraints $\mathbf{x}'\mathbf{x} = n$ and $J\mathbf{x} = \mathbf{x}$. Expanding (134) yields

$$\mathrm{h}(\mathbf{x}) = \sum_{j=1}^{m} \mathbf{y}_j^{'}G_j^{'}G_j\mathbf{y}_j - 2\mathbf{x}^{'}\sum_{j=1}^{m} G_j\mathbf{y}_j + mn, \tag{135}$$

which shows that we need to maximize the linear form

$$\mathrm{f}(\mathbf{x}) = \mathbf{x}^{'}\sum_{j=1}^{m} G_j\mathbf{y}_j = \mathbf{x}^{'}\sum_{j=1}^{m} JG_j\mathbf{y}_j , \tag{136}$$

subject to the constraints $J\mathbf{x}=\mathbf{x}$ and $\mathbf{x}'\mathbf{x}=n$. The solution for $\mathbf{x}$ can be obtained by rescaling $\sum_{j=1}^{m} JG_j\mathbf{y}_j$ to a sum of squares n. A proof, based on the Schwarz inequality, will not be given here.

Once a solution for $\mathbf{y}_1,...,\mathbf{y}_m$ and $\mathbf{x}$ has been obtained, another set of quantifications can be determined, with a second solution for $\mathbf{x}$, orthogonal to the first. The resulting quantifications will of necessity differ from the first. This can be meaningful, when the categories differ in more than one respect. For instance, when four political parties, arranged on a left-right dimension, have been given quantifications 1, 2, 3, 4, expressing this dimension, it may be meaningful to introduce another set of quantifications like 1, 3, 1, 3, expressing the degree of religious affiliation of the parties.

Algebraically, Homogeneity Analysis in two or more dimensions requires the minimum of the function

$$\mathrm{h}(Y_1,...,Y_m, X) = \sum_{j=1}^{m} \| G_j Y_j - X \|^2 \qquad (137)$$

subject to the constraints $X'X = nI$ and $JX = X$. The solution can be obtained explicitly in terms of eigenvectors of $\sum_j JG_j D_j^{-1} G_j' J$, or implicitly by an ALS procedure, see Gifi (1990).

EPILOGUE

Alternating least squares methods have performed a key role in the treatment of Simultaneous Components Analysis, MINRES factor analysis, PARAFAC, INDSCAL, and in many other techniques of Multivariate Analysis not treated here. They are based on the assumption that the variables of the argument of the function can be split into subsets that can be optimized conditionally, for fixed values of the remaining variables in the argument. By definition, such procedures converge monotonically to a stable function value (De Leeuw, Young & Takane, 1976, p.475). However, this property of monotonic convergence can also be obtained with less than conditional *optimization:* Conditional *improvement* of subsets of variables is already enough to have monotonic convergence. In other words, we may consider Alternating Lower Squares as a viable alternative when Alternating Least Squares is not feasible.

Alternating Lower Squares methods would make a suitable topic for a Chapter 5 of this book, but they will not be treated here. Much of the theory of Alternating Lower Squares can be found in the literature on a method called Majorization. Reviews of this method can be found in De Leeuw (1984; 1988), Meulman (1986), Kiers (1990), Heiser (1991) and Kiers and Ten Berge (1992).

REFERENCES

Carroll, J.D. & Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an *N-way* generalization of Eckart-Young decomposition. *Psychometrika,* 35, 283–319.

De Leeuw, J. (1984). Convergence of the majorization algorithm for multi-dimensional scaling. *Research Report RR* - 84 - 07. Leiden: Department of Data Theory.

De Leeuw, J. (1988). Convergence of the majorization method for multi dimensional scaling. *Journal of Classification,* 5, 163–180.

De Leeuw, J. & Heiser, W.J. (1982). Theory of multidimensional scaling. In P.R. Krishnaiah and L.N. Kanai (Eds.), *Handbook of Statistics (Vol.* 2), [pp. 285–316]. Amsterdam: North-Holland.

De Leeuw, J., Young, F.W. & Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika,* 41, 471–503.

Eckart, G., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika,* 1, 211–218.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester: Wiley.

Golub, G.H. & Van Loan, C.F. (1989). *Matrix Computations (2nd Ed.)* Baltimore: The John Hopkins University Press.

Gower, J.C. (1984). Multivariate analysis: Ordination, multidimensional scaling and allied topics. In E. Lloyd (Ed.), *Handbook of Applicable Mathematics VI, Statistics, B.* New York: Wiley, 727–781.

Green, B.F. & Gower, J.C. (1979). *A problem with congruence.* Paper presented at the Annual Meeting of the Psychometric Society, Monterey (Calif.).

Harman, H.H. & Fukuda, Y. (1966). Resolution of the Heywood case in the minres solution. *Psychometrika,* 31, 563–571.

Harman, H.H. & Jones, W.H. (1966). Factor analysis by minimizing residuals (Minres). *Psychometrika*, 31, 351–369

Harshman, R.A. (1970). *Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-model factor analysis.* University of California at Los Angeles. UCLA Working papers in Phonetics, 22, 111–117.

Heiser, W.J. (1991). A generalized majorization method for least squares multidimensional scaling of pseudodistances that may be negative.

Psychometrika, 56, 7–27.

Jöreskog, K.G. (1977). Methods. In K. Enslein, A. Ralston and H.S. Wilf (Eds.). *Mathematical methods for digital computers (Vol.* 3). New York: Wiley.

Kiers, H.A.L. (1990). Majorization as a tool for optimizing a class of matrix functions. *Psychometrika,* 55, 417–428.

Kiers, H.A.L. (1990). *SCA: A program for simultaneous components analysis.* University of Groningen: lEC Progamma.

Kiers, H.A.L. & Ten Berge, J.M.F. (1989). Alternating least squares algorithms for simultaneous components analysis with equal component weight matrices in two or more populations. *Psychometrika*, 54, 467–473.

Kiers, H.A.L. & Ten Berge, J.M.F. (1992). Minimization of a class of matrix trace functions by means of refined majorization. *Psychometrika*, 57, 371–382.

Koschat, M.A. & Swayne, D.F. (1991). A weighted Procrustes criterion. *Psychometrika*, 56, 229–239.

Kristof, W. (1970). A theorem on the trace of certain matrix products and some applications. *Journal of Mathematical Psychology*, 7, 515–530.

Kroonenberg, P.M. (1983). *Three-mode Principal Component Analysis.* Leiden: DSWO Press

Kroonenberg, P.M. & De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45, 69–97.

Magnus, J.R. & Neudecker, H. (1991). *Matrix differential calculus with applications in Statistics and Econometrics.* New York: Wiley.

Meulman, J.J. (1986). *A distance approach to nonlinear multivariate analysis.* Leiden, DSWO Press.

Millsap, R.E. & Meredith, W. (1988). Component analysis in cross-sectional and longitudinal data. *Psychometrika*, 53, 123–134.

Mooijaart, A. & Commandeur, J.J.F. (1990). A general solution of the weighted orthonormal Procrustes problem. *Psychometrika*, 55, 657–663.

Penrose, R. (1956). On the best approximate solutions of linear matrix equations. *Proc. Cambridge Phil. Soc.*, 52, 17–19.

Rao, C.R. (1980). Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In P.R. Krishnaiah (Ed.), *Multivariate Analysis-V,* [3–22]. Amsterdam: North-Holland.

Takane, Y. & Shibayama, T. (1991). Principal components analysis with external information on both subjects and variables. *Psychometrika,* 56, 97–120.

Ten Berge, J.M.F. (1983). A generalization of Kristof's theorem on the trace of certain matrix products. *Psychometrika,* 48, 519–523.

Ten Berge, J.M.F. (1986). Rotation to perfect congruence and the cross-validation of component weights across populations. *Multivariate Behavioral Research,* 21, 41–64 & 262–266.

Ten Berge, J.M.F & Kiers, H.A.L. (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika,* 56, 309–315.

Ten Berge, J.M.F., Kiers, H.A.L. & Krijnen, W.P. (1993). Computational solutions for the problem of negative saliences and nonsymmetry in INDSCAL. *Journal of Classification,* 10, 115-124.

Ten Berge, J.M.F., Kiers, H.A.L. & Van der Stel, V. (1992). Simultaneous Components Analysis. *Statistica Applicata* 4, 377-392.

Ten Berge, J.M.F. & Knol, D.L. (1984). Orthogonal rotations to maximal agreement for two or more matrices of different column orders. *Psychometrika,* 49, 49–55.

Ten Berge, J.M.F. & Nevels, K. (1977). A general solution to Mosier's oblique Procrustes problem. *Psychometrika,* 42, 593–600.

Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika,* 31, 279–311.

Van den Wollenberg, A.L. (1977). Redundancy Analysis: An alternative for canonical correlation analysis. *Psychometrika,* 42, 207–219.

Von Neumann, J. (1937). Some matrix inequalities and metrization of matrix space. *Tomsk University Review,* 1, 286–300. Reprinted in A.H. Taub (Ed.), (1962). *John von Neumann collected works (Vol. IV).* New York: Pergamon.

EXERCISES AND ANSWERS

EXERCISES CHAPTER 1

1. Let X=2**aa**′, with **a**′=(.8 .6). Does the expression $X=K\Lambda K'$ satisfy the definition of an eigendecomposition, when $K=\begin{pmatrix} .8 & 0 \\ .6 & 0 \end{pmatrix}$ and $\Lambda=\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$?

2. Let $S=K\Lambda K'$, with $K=\begin{pmatrix} .8 & .6 \\ .6 & -.8 \end{pmatrix}$ and $\Lambda=\begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$.

   a. Compute $S^{1/2}$ and verify that $S^{1/2}S^{1/2}= S$.
   b. Compute $S^{-1/2}$ and verify that $S^{-1/2}S^{-1/2}= S^{-1}$.
   c. Simplify the product $S^{-1/2}SS^{-1/2}$.
   d. Simplify the product $S^{-1/2}S^2S^{-1/2}$.

3. Let D be a diagonal 3×3 matrix with diagonal elements 4, –2, and 1. Find a matrix K and a diagonal matrix Λ, with $\lambda_1 \geq \lambda_2 \geq \lambda_3$, such that the expression $D=K\Lambda K'$ is an eigendecomposition of D.

4. Construct an example (a 2×2 matrix) of a Gramian matrix that has no inverse, and of a nonsingular matrix that is not Gramian.

5. Give the SVD for an arbitrary vector **x**.

6. Let A be a matrix such that $A'A=C$, a diagonal matrix with strictly positive elements in weakly descending order. Compute an SVD for A.

7. Compute an SVD for $X=\begin{pmatrix} 4 & 0 \\ 3 & 0 \\ 0 & 1 \end{pmatrix}$.

8. Give an eigendecomposition of $A'A$ and of AA', when A has the SVD $A=PDQ'$.

9. Let X have the SVD $X=PDQ'$, with D nonsingular. Prove that $X(X'X)^{-1/2}=PQ'$.

10. Let $X=ADA'$, with $A=\begin{pmatrix} .8 & .6 \\ .6 & -.8 \end{pmatrix}$ and $D=\begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix}$. Give an eigendecomposition

and an SVD of X.

11. Give the singular values of a columnwise orthonormal matrix *H*.

12. Does the Schwarz inequality imply that $\mathbf{u}'\mathbf{v} \leq 1$ if both **u** and **v** are vectors of unit length?

13. Prove that $(\mathbf{x}'R^2\mathbf{x})(\mathbf{x}'R\mathbf{x})^{-1} \geq (\mathbf{x}'R\mathbf{x})(\mathbf{x}'\mathbf{x})^{-1}$, for *R* Gramian and nonsingular, by applying the Schwarz inequality to the vectors **x** and $R\mathbf{x}$.

14. Let *Y* be a 3×2 matrix, with Vec(*Y*)=[2 1 1 0 3 2]′. Determine *Y*.

15. Let $\|X\|^2$ denote the sum of squared elements of *X*. Which of the following statements are correct?
   a. $\|\text{Vec}(A)\|^2 = \text{tr}A'A$
   b. $\text{tr}ABCC'B'A' = \|(C'\otimes A)\text{Vec}(B)\|^2$
   c. $\text{Vec}(AB) = B'\text{Vec}(A)$
   d. $(D'\otimes A)\text{Vec}(BC) = (D'C'\otimes A)\text{Vec}(B)$
   e. $\|Y-ABC'\|^2 = \|\text{Vec}(Y)-(C'\otimes A)\text{Vec}(B)\|^2$

ANSWERS CHAPTER 1

1. No, this is not an eigendecomposition, because *K* is not columnwise orthonormal.

2. a. $S^{1/2} = \begin{pmatrix} 1.64 & .48 \\ .48 & 1.36 \end{pmatrix}$ and $S = \begin{pmatrix} 2.92 & 1.44 \\ 1.44 & 2.08 \end{pmatrix}$.

   b. $S^{-1/2} = \begin{pmatrix} .68 & -.24 \\ -.24 & .82 \end{pmatrix}$ and $S^{-1} = \begin{pmatrix} .52 & -.36 \\ -.36 & .73 \end{pmatrix}$.

   c. The product is S^0 which is the identity matrix.
   d. The product is S^1 which equals *S*.

3. If we take I_3 as eigenvector matrix *K*, and set $\Lambda=D$, the order of the second and third eigenvalues still needs to be reversed in Λ. So the last two columns in *K* need to be reversed. Accordingly, the solution becomes

$$K = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \text{ and } \Lambda = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

4. The matrix $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ is Gramian but has no inverse. The matrix $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ has an inverse (itself) but is not Gramian.

5. Write $\mathbf{x}$ as a unit length vector times the scalar $(\mathbf{x}'\mathbf{x})^{1/2}$. This yields the expression $\mathbf{x}=PDQ'$, with P the vector $\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1/2}$, D the 1×l matrix $(\mathbf{x}'\mathbf{x})^{1/2}$, and Q the l×l identity matrix.

6. Note that $AC^{-1/2}$ is columnwise orthonormal, define P as this matrix, and define D as $C^{1/2}$. Taking $Q=I$, we have $A=PDQ'$ as SVD of A.

7. Take $P = \begin{pmatrix} .8 & 0 \\ .6 & 0 \\ 0 & 1 \end{pmatrix}$, $D = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$, and $Q=I_2$, to obtain the SVD $A=PDQ'$.

8. Saying that the expression $A=PDQ'$ is an SVD means that P is columnwise orthonormal, D is diagonal, nonnegative and ordered, and that Q is at least columnwise orthonormal. It follows that the expression $A'A=QDP'PDQ'=QD^2Q'$ is an eigendecomposition, and so is the expression $AA'=PDQ'QDP'=PD^2P'$.

9. From the eigendecomposition $X'X=QD^2Q'$, we have $(X'X)^{-1/2}=QD^{-1}Q'$. It follows that $X(X'X)^{-1/2}= PDQ'(QD^{-1}Q') = PDD^{-1}Q' = PQ'$.

10. Because A is orthonormal and D is diagonal and ordered, the expression $X=ADA'$ is already an eigendecomposition. It is not an SVD, because singular values cannot be negative. Define $T = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. Then the expression $X=(AT')(TD)(A')$ does satisfy the definition of an SVD.

11. Because the expression $H=HII'$ satisfies the definition of an SVD, it is

clear that all singular values are 1.

12. Yes, as follows from writing $\mathbf{u}'\mathbf{v} \leq (\mathbf{u}'\mathbf{u})^{1/2}(\mathbf{v}'\mathbf{v})^{1/2}$ and noting that $\mathbf{u}'\mathbf{u} = \mathbf{v}'\mathbf{v} = 1$.

13. Applying Schwarz to the vectors $\mathbf{x}$ and $R\mathbf{x}$ yields $(\mathbf{x}'R\mathbf{x})^2 \leq (\mathbf{x}'\mathbf{x})(\mathbf{x}'R^2\mathbf{x})$. Next, divide both sides by $(\mathbf{x}'R\mathbf{x})(\mathbf{x}'\mathbf{x})$ and the inequality is obtained.

14. $Y = \begin{pmatrix} 2 & 0 \\ 1 & 3 \\ 1 & 2 \end{pmatrix}$.

15. Statements a, b and d are correct. In d both sides equal Vec($ABCD$). Statement c is false because a Kronecker product is missing in the right hand side; the trick is to write Vec(AB) as Vec(IAB) or as Vec(AIB) or as Vec(ABI). Statement e has one prime too many.

EXERCISES CHAPTER 2

16. Find an attainable lower bound to the function $f(x)= x^2 - 4x$ and specify for what value of x this bound is attained.

17. Does the inequality $f(x) = 4x^2+2x-5 = (2x+1)^2 - 2x - 6 \geq -2x-6$ represent a lower bound to the function f?

18. Determine an upper bound and the maximum for the function $f(x) = x^2-4x+4$ subject to the constraint $x^2 \leq 1$.

19. Determine, for a fixed matrix A and a fixed vector $\mathbf{v}$, the unit-length vector $\mathbf{u}$ which maximizes the linear form $\mathbf{u}'A\mathbf{v}$.

20. Consider the function $w(\mathbf{x})=\mathbf{x}'AA'\mathbf{y}$, where A is columnwise orthonormal.
    a. Which *form* do we have here?
    b. Find an attainable upper bound to w subject to the constraint that **x** have length one.
    c. Determine the **x** for which the upper bound (maximum) is attained.

21. Consider the 4×3 matrix A of section 1.2, and its SVD. Determine the maximum of the quadratic form $g(\mathbf{y})=\mathbf{y}'A'A\mathbf{y}$ subject to the constraint $\mathbf{y}'\mathbf{y}=1$. Compute the optimal **y**.

22. Determine, for the same A as in the previous case, the maximum of the bilinear form $h(\mathbf{x},\mathbf{y}) = \mathbf{x}'A\mathbf{y}$, subject to the constraints $\mathbf{x}'\mathbf{x} = \mathbf{y}'\mathbf{y} = 1$. Compute the optimal **x** and **y**.

23. Which vector **y** minimizes the function $h(\mathbf{y})=\|\mathbf{x}-GH\mathbf{y}\|^2$, given that GH is of full column rank?

24. Which vector **x** minimizes the function $h(\mathbf{x})=\|\mathbf{x}-GH\mathbf{y}\|^2$?

25. Which vector **z** minimizes the function $w(\mathbf{z})=\|\mathbf{y}' - \mathbf{z}'X\|^2$, given that (XX') has an inverse?

26. Let $A=\begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix}$. Which vector **x** of unit length maximizes the function $w(\mathbf{x}) = \mathbf{x}'A\mathbf{x}$?

27. Which vector **x** minimizes the function $h(\mathbf{x}) = \|A\mathbf{x}-\mathbf{b}\|^2 + \mathbf{x}'Y'Y\mathbf{x}$? Note that $\mathbf{x}'Y'Y\mathbf{x} = \|Y\mathbf{x}-0\|^2$.

ANSWERS CHAPTER 2

16. Writing $f(x) = (x-2)^2-4 \geq -4$ shows that -4 is a lower bound. It is attained

for x=2.

17. No, because $-2x-6$ still depends on x, whence it is not constant.

18. Writing $f(x)= x^2 -4x + 4 \leq 1-4x+4 = -4x+5 \leq 9$ shows that 9 is an upper bound. It is attained for $x = -1$. So 9 is in fact the maximum.

19. Apply Schwarz to the vectors **u** and $A\mathbf{v}$. This yields $\mathbf{u}'A\mathbf{v} \leq (\mathbf{u}'\mathbf{u})^{1/2}(\mathbf{v}'A'A\mathbf{v})^{1/2}$ which equals $(\mathbf{v}'A'A\mathbf{v})^{1/2}$ because $\mathbf{u}'\mathbf{u}$=1. So $(\mathbf{v}'A'A\mathbf{v})^{1/2}$ is an upper bound to $\mathbf{u}'A\mathbf{v}$, considered as a function of **u**. The bound is the maximum because it can be attained, namely for the unit length vector $\mathbf{u}=A\mathbf{v}(\mathbf{v}'A'A\mathbf{v})^{-1/2}$.

20. a. This is a linear form.
    b. Apply Schwarz to the vectors **x** and $AA'\mathbf{y}$. This yields the upper bound $(\mathbf{y}'AA'\mathbf{y})^{1/2}$.
    c. The bound is attained when **x** is taken as $AA'\mathbf{y}$, rescaled to unit length. That is, when $\mathbf{x} = AA'\mathbf{y}(\mathbf{y}'AA'\mathbf{y})^{-1/2}$.

21. The maximum is 25, and is attained when **y**= [.64 .60 .48]′, the first unit length eigenvector of $A'A$.

22. The maximum is 5, attained for **x** = [.8 0 0 .6]′ and **y** = [.64 .60 .48]′, the first left and right hand singular vector of A, respectively.

23. The optimal **y** is $(H'G'GH)^{-1}H'G'\mathbf{x}$.

24. The optimal **x** is simply $GH\mathbf{y}$.

25. The sum of squares of a vector does not change when the vector is transposed. The vector in point is $(\mathbf{y}-X'\mathbf{z})'$. The optimal **z** is now easily found to be $\mathbf{z}=(XX')^{-1}X\mathbf{y}$.

26. First, A is replaced by its symmetric part $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$. The largest eigenvalue

of this matrix is 2, and the associated eigenvector is $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ or minus this vector. The upper bound 2 is attained for **x**=[0 1]′ or **x**=[0 –1]′.

27. Stack Y below A in a supermatrix G and stack a zero vector below **b** in a vector **w**, such that h(**x**) is the sum of squared elements in (G**x**–**w**). The minimizing **x** is $(G'G)^{-1}G'$**w**. Noting that $(G'G)=(A'A+Y'Y)$ we arrive at the solution **x**=$(A'A+Y'Y)^{-1}A'$**b**.

EXERCISES CHAPTER 3

28. Which of the following statements are true?
    a. Every columnwise orthogonal matrix is suborthonormal (s.o.).
    b. The product of any three s.o. matrices is s.o..
    c. If a matrix contains no elements above 1, it is s.o..

29. Consider the trace tr$X'AB'Y$.
    a. Maximize this trace as a function of X subject to the constraint $X'X=I$.
    b. Maximize this trace as a function of X and Y subject to the constraint $X'X=Y'Y=I_q$, where q is a fixed number of dimensions.
    c. Maximize this trace as a function of X and Y subject to the constraints $X=Y$ and $X'X=I_q$.

30. Verify that the solutions to the three problems of the previous question coincide with those of section 2.4 in case X and Y are vectors (q=1).

31. Let A be a symmetric 3×3 matrix, with eigenvalues 4, 1, and –6.
    a. What is the maximum of the generalized quadratic form $g(X)$=tr$X'AX$ subject to the constraint $X'X=I_2$?
    b. What is the maximum of the generalized bilinear form g(X,Y) = tr$X'AY$ subject to the constraint $X'X = Y'Y = I_2$?

32. Determine the minimizing Y for the function $||Z–ZYP'||^2$, when $Z'Z$ and $P'P$ are nonsingular matrices.

33. Let

$$A=\begin{pmatrix} 2 & 3 \\ 1 & 2 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \text{ with } (A'A)^{-1}=\begin{pmatrix} 3 & -2 \\ -2 & 1.4 \end{pmatrix}, \text{ and let } Y=\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ -1 & 0 \end{pmatrix}.$$

Find the minimum and the minimizing X for $h(X) = ||AX – Y||^2$. Verify that the residual variables are orthogonal to the predictor variables, that is, $A'(Y–AX) = 0$.

34. (deleted)

35. Minimize the function $h(V)= ||Z–UVW'||^2$, for U columnwise orthonormal and $W'W$ nonsingular, subject to the constraint that V has rank 2 or less. Find the minimizing V.

36. Let $A=UVW'$, with

$$U=\begin{pmatrix} .8 & 0 \\ .6 & 0 \\ 0 & 1 \end{pmatrix}, V=\begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix}, \text{ and } W=\begin{pmatrix} .8 & .6 \\ -.6 & .8 \end{pmatrix}.$$

a. Find the minimum of $h(X)=||A–X||^2$ subject to $X'X=I_2$.

b. Find the minimum of the same function, subject to the constraint that X have rank 1.

c. Find the maximizing **x** and **y** for the function $g(\mathbf{x},\mathbf{y}) = \mathbf{x}'A\mathbf{y}$ subject to the constraint that **x** and **y** be of unit length.

37. Consider the least squares problem of minimizing $\ell(X) = ||A–XX'||^2$ subject to the constraint $X'X=I_q$, when A is a symmetric matrix. Find the minimizing X by expanding the function in terms of traces.

38. Consider the least squares problem of minimizing the function $\ell(X,Y)=\|A-XY'\|^2$, subject to $X'X=Y'Y=I_q$. Find the minimizing X and Y.

39. The generalized linear, quadratic and bilinear forms are defined as trace functions. The maximum of each of these functions coincides with a minimum of a least squares function. Show this.

ANSWERS CHAPTER 3

28. Only b is true. Example 1 of section 3.2 is a counterexample to a, and Example 3 is a counterexample to c.

29. a. Compute the SVD $AB'Y=PDQ'$ and take $X=PQ'$.
    b. Compute the SVD $AB'=PDQ'$ and let X and Y contain the first q columns of P and Q, respectively. A joint rotation of this X and Y is also allowed.
    c. Compute the first q unit length eigenvectors of the symmetric part $.5(AB'+BA')$ to obtain X and Y. A joint rotation is also permitted. The idea is to treat this function as a generalized quadratic form in X or Y.

30. We now have $\mathbf{x}'AB'\mathbf{y}$ to consider.
    a. Define $\mathbf{v}=AB'\mathbf{y}$, with SVD $\mathbf{v}=(\mathbf{v}(\mathbf{v}'\mathbf{v})^{-1/2})(\mathbf{v}'\mathbf{v})^{1/2}(1)'$. Then we must take $\mathbf{x}=\mathbf{v}(\mathbf{v}'\mathbf{v})^{-1/2}$. This is in agreement with section 2.4.
    b. For this bilinear form the agreement is obvious. The freedom of rotation is now the freedom to change the signs in $\mathbf{x}$ and $\mathbf{y}$ jointly.
    c. Obvious.

31. a. The maximum is 5, the sum of the largest two eigenvalues.
    b. The maximum is 10, the sum of the largest two singular values.

32. This is a Penrose problem, with solution $Y=(Z'Z)^{-1}Z'ZP(P'P)^{-1}=P(P'P)^{-1}$.

33. The minimizing X is $\begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}$, and the minimum is 2. The residual matrix

is $\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ -1 & 0 \end{pmatrix}$, with columns orthogonal to those of A.

34. (deleted)

35. The solution is $V=P_2D_2Q_2'(W'W)^{-1/2}$, where P_2, D_2, and Q_2 are submatrices of P, D and Q, obtained from the SVD $U'ZW(W'W)^{-1/2}=PDQ'$. Specifically, P_2 and Q_2 contain the first two columns of P and Q, respectively, and D_2 is the upper left 2×2 submatrix of D.

36. a. The minimizing X is $UW' = \begin{pmatrix} .64 & -.48 \\ .48 & -.36 \\ .60 & .80 \end{pmatrix}$ .The minimum can be written as

$||A–X||^2 = ||UVW'–UW'||^2 = ||U(V–I)W'||^2= ||V–I||^2 = 16+1 = 17$.

b. The minimizing X is $v_1\mathbf{u}_1\mathbf{w}_1' = \begin{pmatrix} 3.2 & -2.4 \\ 2.4 & -1.8 \\ 0 & 0 \end{pmatrix}$. The minimum can be written as

$||A–X||^2 = ||UVW'–v_1\mathbf{u}_1\mathbf{w}_1'||^2 = ||v_1\mathbf{u}_1\mathbf{w}_1' + v_2\mathbf{u}_2\mathbf{w}_2' – v_1\mathbf{u}_1\mathbf{w}_1'||^2 = ||v_2\mathbf{u}_2\mathbf{w}_2'||^2 = v_2^2= 4$.

c. The maximum is 5, and is attained for $\mathbf{x}=\mathbf{u}_1$ and $\mathbf{y}=\mathbf{w}_1$.

37. Writing ℓ as $\text{tr}A^2–2\text{tr}X'AX+\text{tr}XX'XX'$ and using the constraint shows that the minimizing X for ℓ is the maximizing X for the quadratic form tr($X'AX$). This X is found as the matrix of the first q eigenvectors of A, or a rotation thereof, see section 3.3.

38. Writing ℓ as $\text{tr}A'A–2\text{tr}X'AY+\text{tr}XY'YX'$ and using the constraints shows that the minimizing X and Y for ℓ are the maximizing X and Y for the bilinear

form tr$X'AY$. The X and Y are found from the SVD of A, see section 3.3.

39. The quadratic form is related to the least squares problem of 37. The bilinear form is related to the least squares problem of 38. Finally, the linear form is related to the least squares problem of minimizing $h(X)=||A–X||^2$ subject to $X'X=I$, see 36a.

EXERCISES CHAPTER 4

40. Let A be a nonsingular $p \times p$ matrix, and $\mathbf{y}$ a p-vector. Prove that there is a vector $\mathbf{x}$ such that $A\mathbf{x}=\mathbf{y}$. What practical implication does this result have for multiple regression analysis in relatively small samples?

41. Compute the multiple correlation coefficient for the predictor matrix $A=\begin{pmatrix} 2 & 1 \\ 0 & -1 \\ -2 & 0 \end{pmatrix}$ and the criterion vector $\mathbf{y}=\begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}$.

42. Consider the constraint $X'GX=I$ on X, where G is a Gramian matrix. For which matrix can this constraint be seen as a columnwise orthonormality constraint?

43. In the derivation of PCA (section 4.2) a matrix F was determined from which Z could be optimally reconstructed, subject to the constraint that each column of F be in the column space of Z. This constraint was implemented by writing $F=ZB$, for some matrix B. Show that this constraint is *inactive,* which means that it can be omitted without changing the solution. Specifically, express the F of rank q, that minimizes the function $\ell(F,A)=||Z–FA'||^2$ for a fixed A of rank q, in terms of Z and A, and show that all columns of the resulting F must be in the column space of Z.

44. Find the minimum of $h(X)=||A–BX||^2$ subject to the constraint that each row

of X be in the column space of some matrix C, when it is given that both $(B'B)$ and $(C'C)$ are nonsingular.

45. In SCA, the function $\ell(B_1,...,B_m,P_1,...,P_m) = \sum_{i=1}^{m} \|Z_i - Z_iB_iP_i'\|^2$ is minimized subject to the constraint that $B_1=B_2=...=B_m$, which means that the weight matrices B_i must be the same for all groups, and can be set equal to one matrix B. An alternative approach would be to minimize the same function subject to the constraint that $P_1=P_2=...=P_m$. Develop an explicit solution for the latter approach.
   a. Replace P_i in ℓ by P, and show that the optimal B_i, for fixed P, is $(Z_i'Z_i)^{-1}(Z_i'Z_i)P(P'P)^{-1} = P(P'P)^{-1}$, $i=1,...,m$.
   b. Next, eliminate B_i by writing $P(P'P)^{-1}$ for B_i in ℓ. The remaining problem is to minimize $\sum_i \|Z_i - Z_iP(P'P)^{-1}P'\|^2$ as a function of P only. Define the columnwise orthonormal matrix U as $U=P(P'P)^{-1/2}$ and find the minimizing U for $\sum_i \|Z_i - Z_iUU'\|^2$ subject to $U'U=I_q$. Show that the minimizing U must contain the first q eigenvectors of $\sum_i^m Z_i'Z_i$, or any orthogonal rotation of those.
   c. Suppose that one would wish to minimize ℓ subject to the constraint that both the B_i and the P_i be equal across groups. Show that the optimal B and P for this case are the same as in the case where only the P_i are required to be equal.

46. Determine the maximum of the function $g(X,Y)=\text{tr}X'AY$ subject to the constraint that all columns of X and Y have unit lengths.

47. Explain why CCA would yield trivial results if the orthogonality constraints for the canonical variates were omitted.

48. Prove that the minimum of $\sum_{k=1}^{p} \|X_k - AD_kB'\|^2$ as a function of A is attained for $A = (\sum_{k=1}^{p} X_kBD_k)(\sum_{l=1}^{p} D_lB'BD_l)^{-1}$, when D_k is diagonal, $k=1,...,p$.

49. Let **1** be an n-vector of elements 1 and define $J=(I_n-n^{-1}\mathbf{11}')$, an idempotent matrix (that is, $J=JJ=JJJ=...$).

a. Prove that, for any matrix X of n rows, $JX=X$ if and only if $\mathbf{1}'X=0'$.

b. Let A be a vertical matrix of rank r, with SVD $A=P_rD_rQ_r'$. Prove that $JA=A$ if and only if $JP_r=P_r$.

c. When Y is a vertical matrix of full column rank, which columnwise orthonormal matrix X with zero column means maximizes the function $f(X)=\text{tr}X'Y$?

ANSWERS CHAPTER 4

40. Take $\mathbf{x}=A^{-1}\mathbf{y}$. It follows that multiple regression analysis is trivial when the number of predictor variables equals (or is close to) the number of observation units (subjects).

41. It is evident that $\mathbf{y}$ is the difference between column 1 and column 2 of A. Hence, $\mathbf{y}=A\mathbf{x}$ if we take $\mathbf{x}=[1 \quad -1]'$. Because $\mathbf{y}$ is in the column space of A, the multiple correlation is 1.

42. For the matrix $G^{1/2}X$. The square root $G^{1/2}$ exists because G is Gramian.

43. Regardless of A, the optimal F must minimize ℓ, which can also be written as $\|Z'-AF'\|^2$. It follows that $F'=(A'A)^{-1}A'Z'$, hence $F=ZA(A'A)^{-1}$ and each column of F is in the column space of Z.

44. The constraint can be taken care of by writing $X=EC'$, for some matrix E. The problem is to find the minimizing E for $\ell(E)=\|A-BEC'\|^2$. This is a Penrose problem, with solution $E=(B'B)^{-1}B'AC(C'C)^{-1}$. The X wanted is therefore $X=(B'B)^{-1}B'AC(C'C)^{-1}C'$.

45. a. This follows at once from Penrose, applied to each B_i separately.

b. Upon expanding $\sum_i \|Z_i-Z_iUU'\|^2$ as $\sum_i \text{tr}Z_i'Z_i-2\sum_i \text{tr}U'Z_i'Z_iU+\sum_i \text{tr}U'Z_i'Z_iU$ it is clear that we need the maximizing U for $\text{tr}U'\sum_i Z_i'Z_iU$, subject to $U'U=I_q$.

This is a generalized quadratic form problem, see section 3.3. Having determined U, we can take $P=U$, or any orthogonal rotation of U.

c. If only the P_i are constrained to be equal, as has been done above, it appears that the optimal B_i are also equal, because each of these matrices equals $P(P'P)^{-1}$. In other words, once the constraint $P_1=P_2=\ldots=P_m$ is adopted, the further constraint $B_1=B_2=...=B_m$ will be *inactive*.

46. Let $A=PDQ'$ be a SVD of A, and write $g(X,Y)=\mathbf{x}_1'A\mathbf{y}_1+...+\mathbf{x}_q'A\mathbf{y}_q$. Then each term can be maximized independently by taking $\mathbf{x}_1=\mathbf{x}_2=...=\mathbf{x}_q=\mathbf{p}_1$, the first column of P, and $\mathbf{y}_1=\mathbf{y}_2=...=\mathbf{y}_q=\mathbf{q}_1$, the first column of Q.

47. Without orthogonality constraints, all canonical variates would be the same for the first set of variables, and also for the second set, see the previous exercise.

48. This function can be written as a regression function, see (124), with A' as matrix of regression weights, and the solution is immediate from section 3.4.

49. a. Write $JX=X-n^{-1}\mathbf{1}\mathbf{1}'X$. The second term vanishes if $\mathbf{1}'X=0'$, and vice versa: If $n^{-1}\mathbf{1}\mathbf{1}'X=0$ then premultiplying by $\mathbf{1}'$ gives $\mathbf{1}'X=0'$.

    b. We have $\mathbf{1}'A=0' \Leftrightarrow \mathbf{1}'P_rD_rQ_r'=0' \Leftrightarrow \mathbf{1}'P_r=0'$ because postmultiplying by $Q_rD_r^{-1}$ and by D_rQ_r', respectively, is permitted. Finally, use a.

    c. Let JY have the SVD $JY=PDQ'$. Considering only those X for which $X=JX$ we may write $\text{tr}X'Y=\text{tr}X'JY=\text{tr}X'PDQ' \leq \text{tr}D$. This upper bound can be attained by taking $X=PQ'$. This satisfies the orthonormality constraint but also the constraint that $JX=X$, because the SVD of JY is involved, see b.