CONTENTS

	Page
INTRODUCTION	1
ADDRESS FOR ENQUIRIES	1
NAG LIBRARY SERVICE NEWS	2
DOCUMENTATION NEWS	2
MULTIPLE PRECISION ARITHMETIC REVISITED	3
'How to Live Without Covariance Matrices' *Sven Hammarling*	6
'Reflections on Surfaces' *Geoff Hayes*	32
PARTIAL DIFFERENTIAL EQUATIONS	44
IMPLEMENTATION REVIEW	45
NAG NEWSLETTER CROSSWORD - SOLUTION TO NO. 1	48
- NO. 2	
IN BRIEF	50
- FINITE ELEMENT LIBRARY	
- ICES	
CONFERENCE NOTICE	50
PERSONNEL	51
- NEW STAFF	
- VISITORS	
WHO TO CONTACT AT NAG CENTRAL OFFICE	52
NAG USERS ASSOCIATION	53
1983 MEETING	

Numerical Algorithms Group Limited July 1983

Editor: Miss Janet Bentley

Produced and Printed by NAG Central Office

HOW TO LIVE WITHOUT COVARIANCE MATRICES:

Numerical Stability in Multivariate Statistical Analysis

by Sven Hammarling, NAG Central Office

1. Introduction

Many multivariate techniques in statistics are described in terms of an appropriate sums of squares and cross products matrix, such as a covariance matrix or a correlation matrix, rather than in terms of the original data matrix. While this is frequently the best way of analysing and understanding a technique, it is not necessarily the most satisfactory approach for implementing the technique computationally.

From a numerical point of view, it is usually better to work with the data matrix and avoid the formation of a sums of squares and cross products matrix. In this article we indicate why it is better to work with the data matrix, look at techniques that allow us to avoid the explicit computation of sums of squares and cross products matrices and briefly consider the application of these techniques to three particular multivariate problems.

2. Notation

Let X denote an n by p data matrix (design matrix, matrix of observations), where p is the number of variables and n is the number of data points (objects, individuals, observations), let x_i denote the ith column of X, so that

$$X = [x_1\ x_2\ \ldots\ x_p] \tag{2.1}$$

and x_i is the n element vector of sample observations for the ith variable, let $\bar{x}_i$ and s_i be respectively the sample mean and standard deviation for the ith variable and denote $\bar{x}$, D and e respectively as

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}, \quad D=\mathrm{diag}(d_i)=\begin{bmatrix} d_1 & & 0 \\ & d_2 & \\ & & \ddots \\ 0 & & d_p \end{bmatrix}, \quad e=\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \tag{2.2}$$

where

$$d_i = \begin{cases} 1/s_i, & s_i \neq 0 \\ 0, & s_i = 0 \end{cases}$$

Then the matrix

$$\hat{X} = X - e\bar{x}^T \tag{2.3}$$

we call the *zero means data matrix* or the *matrix of deviations from the mean* and the matrix

$$\tilde{X} = \hat{X}D \tag{2.4}$$

we call the *standardized data matrix*, since the mean of each column of $\tilde{X}$ is zero and the standard deviation of each column is unity (unless $s_i = 0$ in which case the column is zero.)

The normal matrices X^TX and $\hat{X}^T\hat{X}$ are called the *sums of squares and cross products matrix* and the *corrected sums of squares and cross products matrix* respectively, the matrix C given by

$$C = \frac{1}{n-1}\hat{X}^T\hat{X} \tag{2.5}$$

is called the *sample covariance matrix*, because c_{ij} is the sample covariance of variables x_i and x_j and the matrix R given by

$$R = \frac{1}{n-1}\tilde{X}^T\tilde{X} \tag{2.6}$$

is called the *sample correlation matrix*, because r_{ij} is the sample correlation coefficient of variables x_i and x_j. Of course

$$R = DCD \tag{2.7}$$

and both C and R are symmetric positive semi-definite matrices, since for any vector z

$$z^TCz = \frac{1}{n-1}(\hat{X}z)^T(\hat{X}z) \geq 0$$

with a similar result for R.

The notation $\|z\|$ denotes a norm of the n element vector z and throughout this article we shall use only the Euclidean norm (Euclidean length) so that

$$\|z\| = \left(\sum_{i=1}^{n} z_i^2\right)^{\frac{1}{2}} \tag{2.8}$$

Similarly $\|Z\|$ denotes a norm of an n by p matrix Z and throughout this article we shall use only the spectral, or 2-norm given by (Wilkinson, 1965)

$$\|Z\| = \max_{\|z\|=1} \|Zz\| = \sigma^{\frac{1}{2}}(Z^TZ), \tag{2.9}$$

where $\sigma(Z^TZ)$ denotes the spectral radius of Z^TZ, that is the largest eigenvalue of Z^TZ. One reason for our interest in these particular norms is that when Z is orthogonal, so that $Z^TZ = I$, then

$$\|Zz\| = \|z\| \quad \text{and} \quad \|Z\| = 1, \; Z^TZ = I \tag{2.10}$$

It is not necessary for the reader to have a detailed knowledge of the spectral norm of a matrix, and to give a feel for its size in relation to the elements of Z we note that

$$\|Z\| \leqq \Big(\sum_{j=1}^{p}\sum_{i=1}^{n} z_{ij}^2\Big)^{\frac{1}{2}} \leqq p^{\frac{1}{2}}\|Z\| \tag{2.11}$$

3. Instability in Forming Normal Matrices

For numerical stability it is frequently desirable to avoid forming normal matrices such as X^TX or $\hat{X}^T\hat{X}$, but instead use algorithms that work directly on the data matrices X or $\hat{X}$. (See for example Golub, 1965) This can be especially important when the data matrix is close to being rank deficient, or when small perturbations in the data can change, or come close to changing, the rank of the data matrix. In such cases the normal matrix will be much more sensitive to perturbations than the data matrix.

A well known example is provided by the matrix

$$X = \begin{bmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix}, \quad \varepsilon \neq 0, \quad \text{for which } X^TX = \begin{bmatrix} 1+\varepsilon^2 & 1 \\ 1 & 1+\varepsilon^2 \end{bmatrix}$$

Perturbations of order ε are required to change the rank of X, whereas perturbations of only ε^2 are required to change the rank of X^TX. This could be particularly disastrous if $|\varepsilon|$ is above noise level, while ε^2 is close to or below noise level.

A second example is provided by the case where X is a square non-singular matrix. The sensitivity of the solution of the equations

$$Xb = y$$

to perturbations in X and y is determined by the size of the condition number of X with respect to inversion, c(X), given by

$$c(X) = \|X\| \ \|X^{-1}\| \tag{3.1}$$

(Wilkinson, 1963; Wilkinson, 1965; Forsythe and Moler, 1967.) Specifically, if we perturb X by a matrix E, the solution of the perturbed equations

$$(X + E)\ (b + e) = y \tag{3.2}$$

satisfies

$$\frac{\|e\|}{\|b + e\|} \leq c(X)\ \frac{\|E\|}{\|X\|}, \tag{3.3}$$

where it should be noted that $c(X) \geq 1$. For the spectral norm it can readily be shown that

$$c(X^TX) = c^2(X) \tag{3.4}$$

so that unless $c(X) = 1$, which occurs only when X is orthogonal, X^TX is more sensitive to perturbations than X. From (3.4) we once again see that perturbations of order ε^2 in X^TX can have the same effect as perturbations of order ε in X.

In terms of solving a system of equations, (3.3) and (3.4) imply that if rounding errors or data perturbations (noise) mean that we might lose t digits accuracy, compared to the accuracy of the data, when solving equations with X as the matrix of coefficients, then we should expect to lose 2t digits accuracy when solving equations with X^TX as the coefficient matrix.

Similar remarks apply to the sensitivity of the solution of linear least squares (multiple regression) problems when X is not square, so long as the residual, or error, vector is small relative to the solution; and once again it is advisable to avoid forming the normal equations in order to solve the least squares problem. (Detailed analyses can be found in Golub and Wilkinson, 1966; Lawson and Hanson, 1974; Stewart, 1977.)

We are not trying to imply that normal matrices should be avoided at all costs. When X is close to being orthogonal then the normal matrix X^TX will be well conditioned. It is not unusual for the data to be given in the form of a normal matrix such as a correlation matrix, in which case one has little option but to work from the normal matrix. Correlations in themselves provide useful statistical information and so one frequently wishes to look at the elements of the correlation matrix, although in this latter case the methods we propose allow us ready access to the elements of X^TX anyway. But the additional sensitivity of X^TX is a real phenomenon, not just a figment of the numerical analyst's imagination and since perturbations in X do not map linearly into perturbations in X^TX, perturbation and rounding error analyses become difficult to interpret when X^TX is used in place of X and decisions about rank and linear dependence (multicollinearity) are harder to make.

4. The QU Factorization and the Singular Value Decomposition

In this section we discuss the tools that allow us to avoid forming normal matrices. These tools are factorizations called the *QU factorization* (or the *QR factorization*) and the *singular value decomposition*. For simplicity of discussion we shall assume throughout that $n \geq p$ so that X has at least as many rows as columns. With one exception we shall also not discuss the details of the algorithms for finding the factorizations, but instead give suitable references for such descriptions. Suffice it to say that both factorizations may be obtained by numerically stable methods using routines in the NAG Fortran Library.

The *QU factorization* of a matrix X is given by

$$X = Q \begin{bmatrix} U \\ O \end{bmatrix}, \tag{4.1}$$

where Q is an n by n orthogonal matrix, so that $Q^TQ = I$, and U is a p by p upper triangular matrix. Of course the rank of U is the same as that of X and when $n = p$ the portion below U does not exist.

The QU factorization of X always exists and may be found by means, for example of Householder transformations, or plane rotations, or Gram-Schmidt orthogonalization. (Wilkinson, 1965; Golub, 1965; Stewart, 1974; Dongarra et al, 1979.) We discuss briefly a method using plane rotations in section 5.

Two features of the QU factorization are important for our purposes. Firstly we see that

$$X^TX = U^TU \tag{4.2}$$

so the elements of X^TX can readily be computed from the inner products of columns of U, which means that U gives a convenient and compact representation of X^TX. In fact, as with X^TX, we need only $\frac{1}{2}p(p+1)$ storage locations for the non-zero elements of U. The matrix U is often called the *Cholesky factor* of X^TX. Secondly if we perturb U by a matrix F then

$$Q\begin{bmatrix} U+F \\ O \end{bmatrix} = X+E, \quad E = Q\begin{bmatrix} F \\ O \end{bmatrix} \tag{4.3}$$

and since Q is orthogonal

$$\|F\| = \|E\| \tag{4.4}$$

so that a perturbation of order ε in U corresponds to a perturbation of the same order of magnitude in X.

Q is an n by n matrix and so it is large if there are a large number of data points, but in fact Q is rarely required explicitly; instead what is usually required is a vector, or part of a vector, of the form Q^Ty, for a given y, and this can be computed at the same time as the QU factorization is computed.

The *singular value decomposition (SVD)* of a matrix X is given by

$$X = Q\begin{bmatrix} \Sigma \\ O \end{bmatrix}P^T, \tag{4.5}$$

where Q is an n by n orthogonal matrix, P is a p by p orthogonal matrix and Σ is a p by p diagonal matrix

$$\Sigma = \mathrm{diag}(\sigma_i) = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_p \end{bmatrix}$$

with non-negative diagonal elements. The factorization can be chosen so that

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0 \tag{4.6}$$

and we shall assume this to be the case. As with the QU factorization the SVD always exists and it may be obtained by first reducing X to bi-diagonal form and then applying a variant of the QR algorithm to reduce this to the diagonal matrix Σ. (Golub and Kahan, 1965; Golub and Reinsch, 1970; Wilkinson, 1977; Wilkinson, 1978.) The σ_i, $i = 1,2,\ldots,p$ are called the *singular values* of X, the columns of P are the *right singular vectors* of X and the first p columns of Q are the *left singular vectors* of X. If we denote the ith columns of P and Q by p_i and q_i respectively then equation (4.5) implies that

$$Xp_i = \sigma_i q_i, \quad i = 1,2,\ldots,p. \tag{4.7}$$

For this factorization we have that

$$X^TX = P\Sigma^2P^T, \tag{4.8}$$

which is the classical spectral factorization of X^TX. Thus the columns of P are the eigenvectors of X^TX and the values σ_i^2, $i = 1,2,\ldots,p$ are the eigenvalues of X^TX.

Σ and P give us an alternative representation for X^TX, although not quite as compact as U since we now need p(p+1) storage locations. Note that (4.8) implies that

$$\|X\| = \sigma_1. \tag{4.9}$$

Analagously to equations (4.3) and (4.4) if we perturb Σ by a matrix F then

$$Q\begin{bmatrix}\Sigma+F\\ O\end{bmatrix}P^T = X+E, \quad E = Q\begin{bmatrix}F\\ O\end{bmatrix}P^T \tag{4.10}$$

and

$$\|F\| = \|E\| \tag{4.11}$$

so that again perturbations of order ε in Σ correspond to perturbations of the same order of magnitude in X.

The SVD is important in multivariate analysis because it provides the most reliable method of determining the numerical rank of a matrix and can be a great aid in analysing near multicollinearities in the data.

Of course if X is exactly of rank $k < p$ then from (4.5) and (4.6) we must have

$$\sigma_{k+1} = \sigma_{k+2} = \ldots = \sigma_p = 0$$

and from (4.7)

$$Xp_i = 0, \quad i = k+1,k+2,\ldots,p$$

so that these columns of P form an orthonormal basis for the null space of X. If X is of rank p, but we choose the matrix F in equation (4.10) to be the diagonal matrix

$$F = \text{diag}(f_i), \quad f_i = \begin{cases} 0, & i = 1,2,\ldots,k \\ -\sigma_i, & i = k+1,k+2,\ldots,p \end{cases} \tag{4.12}$$

then (X+E) is of rank k and from (4.11)

$$\|E\| = \sigma_{k+1} \tag{4.13}$$

so that regarding a small singular value of X as zero corresponds to making a perturbation in X whose size is of the same order of magnitude as that of the singular value.

Conversely, if X is of rank p, but E is a matrix such that the perturbed matrix (X+E) is of rank $k < p$ then it can readily be shown (Wilkinson, 1978) that

$$\sum_{j=1}^{p} \sum_{i=1}^{n} e_{ij}^2 \geq \sum_{i=k+1}^{p} \sigma_i^2 \tag{4.14}$$

so that if the elements of E are small then the singular values $\sigma_{k+1},\sigma_{k+2},\ldots,\sigma_p$ must also be small. Thus if X has near multicollinearities, then the appropriate number of singular values of X must be small. To appreciate the strength of this statement consider the p by p matrix

$$U = \begin{bmatrix} 1 & -1 & -1 & \ldots & -1 & -1 & -1 \\ 0 & 1 & -1 & \ldots & -1 & -1 & -1 \\ 0 & 0 & 1 & \ldots & -1 & -1 & -1 \\ \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 & -1 & -1 \\ 0 & 0 & 0 & \ldots & 0 & 1 & -1 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 1 \end{bmatrix}$$

U is clearly of full rank, p, but its appearance belies its closeness to a rank deficient matrix. If we put

$$E = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \dots & 0 \\ -2^{2-p} & 0 & \dots & 0 \end{bmatrix}$$

then the matrix (U+E) has rank (p-1), so that when p is not small U is almost rank deficient. On the other hand (4.14) assures us that

$$\sigma_p \leq 2^{2-p}$$

so that the near rank deficiency will be clearly exposed by the singular values. For instance, when p = 32 so that $2^{2-p} = 2^{-30} \simeq 10^{-9}$ the singular values of U are approximately 20.05, 6.925 ,..., 1.449, 5.280×10^{-10} and σ_{32} is indeed less than 10^{-9}.

The SVD can be computed by numerically very stable methods and the above remarks also hold in the presence of rounding errors, except when the perturbations under consideration are smaller than the machine accuracy, which is not very likely in practice. Even then we only have to allow for the fact that computationally singular values will not usually have values less than about eps.σ_1, where eps is the relative machine precision, because now the machine error dominates the data error. For example on a VAX 11/780 in single precision, for which eps = $2^{-24} \simeq 6 \times 10^{-8}$, the smallest singular value of the above matrix U, as computed by the NAG Library routine F02WDF, was 4.726×10^{-8} instead of $\sigma_{32} = 5.280 \times 10^{-10}$.

The singular value decomposition is of course a more complicated factorization than the QU factorization, it requires more storage and takes longer to compute, although this latter aspect is frequently over-emphasised.

For many applications the QU factorization is quite sufficient and a convenient strategy is to compute this factorization and then test U to see whether or not it is suitable for the particular application.

For example, if U is required to be non-singular then, at a moderate extra expense, we can compute or estimate its condition number c(U) in order to determine whether or not U is sufficiently well-conditioned. If U is not suitable we can then proceed to obtain the SVD of U as, say

$$U = \tilde{Q}\ \Sigma\ P^T, \tag{4.15}$$

where $\tilde{Q}$ and P^T are orthogonal and Σ is diagonal. From

$$X = Q\begin{bmatrix} U \\ 0 \end{bmatrix}$$

we get that the SVD of X is then given by

$$X = \hat{Q}\begin{bmatrix} \Sigma \\ 0 \end{bmatrix}P^T, \text{ where } \hat{Q} = Q\begin{bmatrix} \tilde{Q} & 0 \\ 0 & I \end{bmatrix} \tag{4.16}$$

and thus the singular values and right singular vectors of U and X are identical. We can take advantage of the upper triangular form of U in computing its SVD and for typical statistical data where n is considerably larger than p the time taken will be dominated by the QU factorization of X. The NAG Fortran routines for computing the SVD in the case where $n \geq p$ all compute the QU factorisation first and routine F02WDF explicitly allows the user to stop at the QU factorization if U is not too ill-conditioned.

5. Plane Rotations and the QU Factorization

In this section we give just a brief discussion of plane rotations and their use in obtaining the QU factorization since the ideas will be useful in section 6.

A *plane rotation* is a transformation of the form

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix}\begin{bmatrix} x^T \\ y^T \end{bmatrix} = \begin{bmatrix} cx^T+sy^T \\ -sx^T+cy^T \end{bmatrix}, \quad c = \cos\theta,\ s = \sin\theta \tag{5.1}$$

where x and y are column vectors. The matrix

$$R = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$$

is called a *plane rotation matrix* and because $c^2 + s^2 = 1$ we have $R^TR = I$ so that R is orthogonal. If x^T and y^T are rows i and j of some n by p matrix X

then the transformation is said to be a *plane rotation for the (i,j) plane* and the plane rotation matrix with the (i,i), (i,j), (j,i) and (j,j) positions of the unit matrix replaced by the four elements of R is denoted by R_{ij}. For example when n = 7, i = 2 and j = 5 then

$$R_{25} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & c & 0 & 0 & s & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -s & 0 & 0 & c & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Of course only rows i and j are affected by such a transformation and R_{ij} is orthogonal.

By appropriate choice of the angle θ we can use plane rotations to perform elimination of elements. For instance if we take

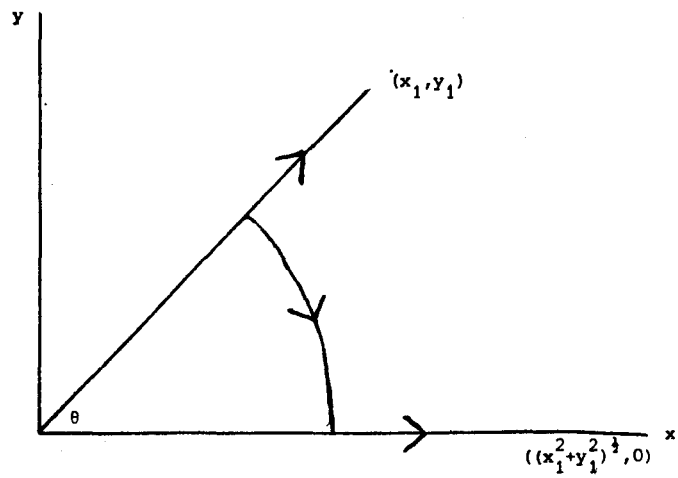
$$c = x_1/(x_1^2+y_1^2)^{\frac{1}{2}}, \quad s = y_1/(x_1^2+y_1^2)^{\frac{1}{2}}, \quad x_1^2 + y_1^2 \neq 0, \tag{5.2}$$

which corresponds to the choice $\theta = \tan^{-1}(y_1/x_1)$, then

$$R\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} (x_1^2+y_1^2)^{\frac{1}{2}} \\ 0 \end{bmatrix}. \tag{5.3}$$

This is illustrated geometrically in Figure 5.1.

Fig 5.1



Obviously if x_1 and y_1 are both zero then they remain zero for any choice of θ.

The QU factorization of X can be obtained by performing a sequence of plane rotations on X. Various sequences will achieve this, for example, the process analogous to the usual description of Gaussian elimination is to perform the elimination column by column. The elements below the diagonal element in the rth column being eliminated by a sequence of rotations in the (r,r+1) plane, the (r,r+2) plane ,..., the (r,n) plane. Thus if we define the orthogonal matrix Q_r by

$$Q_r^T = R_{rn} \cdots R_{r,r+2} R_{r,r+1} \tag{5.4}$$

then we have that (if n = p take $Q_p = I$)

$$Q_p^T Q_{p-1}^T \cdots Q_2^T Q_1^T X = \begin{bmatrix} U \\ 0 \end{bmatrix} \tag{5.5}$$

so that

$$X = Q \begin{bmatrix} U \\ 0 \end{bmatrix}, \quad Q = Q_1 Q_2 \cdots Q_{p-1} Q_p \tag{5.6}$$

(Givens, 1958.) Note that if we require a vector or part of a vector of the form $Q^T y$, then we can apply the plane rotations to y at the same time as they are applied to X, thus avoiding the need to store the rotations.

An alternative scheme, which can be very useful in statistical applications, is to process X a row at a time instead of a column at a time. In statistical terms this corresponds to processing one observation at a time as opposed to the above case where X is processed one variable at a time. Thus in this scheme, in place of (5.4), we define

$$Q_r^T = \begin{cases} R_{r-1,r} \cdots R_{2r} R_{1r}, & 1 < r \leq p \\ R_{pr} \cdots R_{2r} R_{1r}, & r > p \end{cases} \tag{5.7}$$

and then in place of (5.6)

$$X = Q \begin{bmatrix} U \\ 0 \end{bmatrix}, \quad Q = Q_2 Q_3 \cdots Q_n. \tag{5.8}$$

The importance of this scheme is that it enables us to process the data sequentially, an observation at a time, without having to store the data matrix X. If we let X_r denote the data matrix for the first r observations

then, after processing the rth observation, we have a QU factorization

$$\begin{bmatrix} X_r \\ 0 \end{bmatrix} = Q^{(r)} \begin{bmatrix} U_r \\ 0 \end{bmatrix}; \quad Q^{(r)} = Q_2 Q_3 \cdots Q_r, \; Q^{(1)} = I, \tag{5.9}$$

where U_r is a p by p upper triangular matrix whose last (p-r) rows are zero if $r < p$. If z_{r+1}^T denotes the (r+1)th row of X then, since $Q^{(r)}$ affects only the first r rows of X,

$$\begin{bmatrix} X_{r+1} \\ 0 \end{bmatrix} = \begin{bmatrix} X_r \\ z_{r+1}^T \\ 0 \end{bmatrix} = Q^{(r)} \begin{bmatrix} U_r \\ 0 \\ z_{r+1}^T \\ 0 \end{bmatrix} = Q^{(r+1)} \begin{bmatrix} U_{r+1} \\ 0 \end{bmatrix}$$

so that

$$Q_{r+1}^T \begin{bmatrix} U_r \\ 0 \\ z_{r+1}^T \\ 0 \end{bmatrix} = \begin{bmatrix} U_{r+1} \\ 0 \end{bmatrix}$$

The essential part of this transformation is of the form

$$P_{r+1}^T \begin{bmatrix} U_r \\ \\ z_{r+1}^T \end{bmatrix} = \begin{bmatrix} U_{r+1} \\ 0 \end{bmatrix}, \tag{5.10}$$

where in place of (5.7)

$$P_r^T = \begin{cases} R_{r-1,p+1}^{(r)} \cdots R_{2,p+1}^{(r)} R_{1,p+1}^{(r)}, & 1 < r \leq p \\ R_{p,p+1}^{(r)} \cdots R_{2,p+1}^{(r)} R_{1,p+1}^{(r)}, & r < p \end{cases} \tag{5.11}$$

and the (p+1) by (p+1) matrix $R_{i,p+1}^{(r)}$ is defined by the same angle θ as that defining the n by n matrix R_{ir}.

Thus we can think of this as an updating process and indeed, whenever we have a QU factorization, we can use this technique to update the factorication with new observations. (Golub, 1965; Gentleman, 1974a; Gill and Murray, 1977; Dongarra et al, 1979; Cox, 1981.)

6. The QU Factorization of Corrected Sums of Squares and Cross Product Matrices

As mentioned in the previous section there are many applications where it is desirable to process the data sequentially without storing the data matrix X. Statistical packages such as BMDP allow one to form covariance and correlation matrices by sequentially processing the data (BMDP, 1977, section A.2) and we now show that we can also obtain the QU factorization of such matrices by a corresponding process. As far as we are aware this has not been described elsewhere in this context, although it is a straightforward application of a standard rank one update.

First we must note that sample means and variances can be computed sequentially and, indeed, there are good numerical reasons for preferring to compute means and variances this way, rather than by the traditional formulae. (Chan and Lewis, 1979; West, 1979; Chan, Golub and Le Veque, 1982). If we denote the ith observation of the jth variable as $x_j^{(i)}$ and let $\gamma_j^{(r)}$ and $\bar{x}_j^{(r)}$ denote, respectively, the estimated sums of squares of deviations from the mean and the estimated mean of the first r observations, so that

$$\bar{x}_j^{(r)} = \Big(\sum_{i=1}^{r} x_j^{(i)}\Big)/r, \qquad \gamma_j^{(r)} = \sum_{i=1}^{r} (\bar{x}_j^{(r)} - x_j^{(i)})^2,$$

then it is readily verified that

$$\bar{x}_j^{(r)} = \bar{x}_j^{(r-1)} + (x_j^{(r)} - \bar{x}_j^{(r-1)})/r \tag{6.1}$$

$$\gamma_j^{(r)} = \gamma_j^{(r-1)} + (r-1)(x_j^{(r)} - \bar{x}_j^{(r-1)})^2/r$$

Of course

$$\bar{x}_j = \bar{x}_j^{(n)} \text{ and } s_j^2 = \gamma_j^{(n)}/(n-1) \tag{6.2}$$

and so we can obtain $\bar{x}$ and D of (2.2) with one pass through the data. Given the QU factorization of X, (2.3) gives

$$\hat{X} = Q\begin{bmatrix} U \\ 0 \end{bmatrix} - e\bar{x}^T$$

$$= Q\left\{\begin{bmatrix} U \\ 0 \end{bmatrix} - f\bar{x}^T\right\}, \qquad f = Q^T e \tag{6.3}$$

Now f can be reduced to a scalar multiple of the first column of the unit matrix, e_1, by various sequences of plane rotations. Of particular interest here is the sequence defined by

$$P_1^T f = n^{\frac{1}{2}} e_1, \quad P_1^T = R_{12}^{(1)} R_{23}^{(1)} \ldots R_{n-2,n-1}^{(1)} R_{n-1,n}^{(1)} \tag{6.4}$$

so that

$$f = n^{\frac{1}{2}} P_1 e_1$$

and

$$\hat{X} = QP_1 \left\{ P_1^T \begin{bmatrix} U \\ 0 \end{bmatrix} - n^{\frac{1}{2}} e_1 \bar{x}^T \right\} \tag{6.5}$$

The matrix $e_1 \bar{x}^T$ is zero everywhere except the first row and rotations defined by P_1^T introduce an extra sub-diagonal below U so that the matrix in the braces has the form illustrated, when p = 5 and n = 8, by

$$P_1^T U - n^{\frac{1}{2}} e_1 \bar{x}^T = \begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \\ & & & & \\ & & & & \end{bmatrix}$$

This can be transformed back to upper traingular form by a sequence of plane rotations

$$P_2^T \left\{ P_1^T \begin{bmatrix} U \\ 0 \end{bmatrix} - n^{\frac{1}{2}} e_1 \bar{x}^T \right\} = \begin{bmatrix} \hat{U} \\ 0 \end{bmatrix}, \quad P_2^T = R_{p,p+1}^{(2)} \ldots R_{23}^{(2)} R_{12}^{(2)}, \tag{6.6}$$

where $\hat{U}$ is a p by p upper triangular matrix, so that the QU factorization of $\hat{X}$ is given by

$$\hat{X} = \hat{Q} \begin{bmatrix} \hat{U} \\ 0 \end{bmatrix}, \quad \hat{Q} = QP_1P_2 \tag{6.7}$$

Since we can obtain the QU factorization of X without storing X and noting that $\hat{U}D$ is still upper triangular, this process, together with (6.1) and (6.2) enables us to find the QU factorizations of $\hat{X}$ and $\tilde{X}$ of

of (2.3) and (2.4) and hence the Cholesky factors of the matrices C and R of (2.5) and (2.6), by sequentially processing the data an observation at a time.

This method requires storage of the n element vector f.

As an alternative we can obtain the QU factorization of $\hat{X}$ by an updating process. If we let $\hat{X}_r$ denote the zero means data matrix for the first r observations, define $\bar{x}^{(r)}$ as the vector

$$(\bar{x}^{(r)})^T = \begin{bmatrix} \bar{x}_1^{(r)} & \bar{x}_2^{(r)} & \dots & \bar{x}_p^{(r)} \end{bmatrix}$$

and take the number of elements in the vector e by context, then using (6.1)

$$\hat{X}_{r+1} = X_{r+1} - e(\bar{x}^{(r+1)})^T$$

$$= \begin{bmatrix} X_r \\ z_{r+1}^T \end{bmatrix} - \begin{bmatrix} e \\ 1 \end{bmatrix} \left[\bar{x}^{(r)} + (z_{r+1} - \bar{x}^{(r)})/(r+1) \right]^T$$

so that

$$\hat{X}_{r+1} = \begin{bmatrix} \hat{X}_r - \frac{1}{r+1} e (z_{r+1} - \bar{x}^{(r)})^T \\ (z_{r+1} - \bar{x}^{(r+1)})^T \end{bmatrix} \tag{6.8}$$

We can obtain the QU factorization of $\{\hat{X}_r - \frac{1}{r+1} e (z_{r+1} - \bar{x}^{(r)})^T\}$ from that of $\hat{X}_r$ by the method described above and we can then update this QU factorization by the additional row $(z_{r+1} - \bar{x}^{(r+1)})^T$. Again it does not seem to be possible to avoid storing the n element vector $Q^T e$.

A method requiring storage only of additional p element vectors would be useful.

7. Solving Multiple Regression Problems

In this section we consider the application of the QU factorization and the singular value decomposition to multiple regression, or linear least squares and we shall take X to denote either the data matrix, or the standardized data matrix since the solution of a regression with one matrix can be deduced from the solution with either of the others.

We wish to determine the vector b to

$$\text{minimize } r^T r, \quad \text{where } y = Xb + r \tag{7.1}$$

and y is a vector of dependent observations. The elements of b are called the *regression coefficients* and r is the residual vector usually assumed to come from a normal distribution with

$$E(r) = 0 \text{ and } E(rr^T) = \sigma^2 I$$

If Q is orthogonal then

$$r^T r = r^T Q^T Q r = (Qr)^T (Qr)$$

and (7.1) is equivalent to

$$\text{minimize } \tilde{r}^T \tilde{r}, \quad \text{where } Q^T y = Q^T Xb + \tilde{r}, \quad \tilde{r} = Q^T r \tag{7.2}$$

If we choose Q as the orthogonal matrix of the QU factorization of X and partition $Q^T y$ as

$$Q^T y = \begin{bmatrix} \tilde{y} \\ w \end{bmatrix}, \quad \tilde{y} \text{ a p element vector} \tag{7.3}$$

then

$$\tilde{r} = \begin{bmatrix} \tilde{y} - Ub \\ w \end{bmatrix} \tag{7.4}$$

If X has linearly independent columns then U will be non-singular and we can choose b so that

$$Ub = \tilde{y} \tag{7.5}$$

Since w is independent of b, this must be the choice of b that minimizes $\tilde{r}^T \tilde{r}$ and hence $r^T r$ (Golub, 1965; Gentleman, 1974b). For this choice

$$\tilde{r} = \begin{bmatrix} 0 \\ w \end{bmatrix} \text{ so that } r^T r = w^T w \tag{7.6}$$

which is information that is lost when the normal equations are formed. We need not retain w during the factorization, but we can instead just update the sum of squares so that we have the single value $w^T w$ on

completion of the QU factorization. As the discussion in section 3 indicates, the sensitivity of the solution of (7.5) is determined by the closeness of X to rank deficiency, whereas the sensitivity of the solution of the normal equations is determined by the closeness of X^TX to rank deficiency and we have seen that, if perturbations of order ε change the rank of X, then perturbations of order ε^2 change the rank of X^TX. (Wilkinson, 1974).

If X is rank deficient, so that its columns are not linearly independent then U will be singular. Using the notation of (4.14), if we then obtain the SVD of U (7.4) becomes

$$\tilde{r} = \begin{bmatrix} \tilde{y} - \tilde{Q}\,\Sigma\,P^Tb \\ w \end{bmatrix} = \begin{bmatrix} \tilde{Q} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{Q}^T\tilde{y} - \Sigma\,P^Tb \\ w \end{bmatrix}$$

so that (7.1) is equivalent to

$$\text{minimize } \hat{r}^T\hat{r}, \quad \text{where } \hat{r} = \tilde{Q}^T\tilde{y} - \Sigma\,P^Tb \tag{7.7}$$

If X has rank k then only the first k singular values will be non-zero, so let us put

$$\Sigma = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}, \quad S \text{ - k by k and non-singular} \tag{7.8}$$

and correspondingly partition $\tilde{Q}^T\tilde{y}$ and P^Tb as

$$\tilde{Q}^T\tilde{y} = \begin{bmatrix} \hat{y} \\ v \end{bmatrix}, \quad P^Tb = \begin{bmatrix} \hat{c} \\ \tilde{c} \end{bmatrix}, \quad \hat{y}, \hat{c} \text{ - k element vectors} \tag{7.9}$$

Then

$$\hat{r} = \begin{bmatrix} \hat{y} - S\hat{c} \\ v \end{bmatrix} \tag{7.10}$$

and $\hat{r}^T\hat{r}$ is minimized by choosing $\hat{c}$ so that

$$S\hat{c} = \hat{y} \tag{7.11}$$

Since S is diagonal $\hat{c}_i = \hat{y}_i/\sigma_i$, $i = 1,2,\ldots,k$. We also have

$$r^Tr = w^Tw + \hat{r}^T\hat{r} = w^Tw + v^Tv \tag{7.12}$$

As $\tilde{c}$ is not determined by (7.11) we can see that the solution is not unique. The particular solution for which $b^T b$ is also a minimum is called the *minimal length solution* and from (7.9) we see that this is given by taking

$$\tilde{c} = 0 \text{ in which case } b = P \begin{bmatrix} \hat{c} \\ 0 \end{bmatrix} \tag{7.13}$$

(Golub and Reinsch, 1970). It is of interest to note that this solution can formally be written as

$$b = X^{+} y, \tag{7.14}$$

where X^{+} is the Moore-Penrose pseudo inverse of X. (Peters and Wilkinson, 1970.) Using (4.15) it is readily verified that

$$X^{+} = P \begin{bmatrix} \Sigma^{+} & 0 \end{bmatrix} Q^T \qquad \Sigma^{+} = \begin{bmatrix} S^{-1} & 0 \\ 0 & 0 \end{bmatrix} \tag{7.15}$$

In practice X will not be exactly rank deficient and the computed singular values will not be exactly zero and while it is not always easy to decide upon the numerical rank of X, (Golub, Klema and Stewart, 1976; Stewart, 1979; Klema and Laub, 1980) equations (4.10) - (4.13) tell us about the effects of neglecting small singular values. Furthermore (4.7) gives

$$\| X p_i \| = \sigma_i \tag{7.16}$$

and so the columns of P corresponding to small singular values give valuable information on the near multicollinearities in X. We can also readily assess the affects of different decisions on the rank of X on the solution and on the residual sum of squares from a knowledge of the singular values and right singular vectors (Lawson and Hanson, 1974, chapter 25, section 6.)

The NAG Fortran routine F04JGF for computing least squares solutions allows the solution to be computed from the QU factorization when U is sufficiently well-conditioned, but proceedes on to the SVD if U is nearly singular and in this case, although a particular solution is computed, the singular values and right singular vectors are also returned so that alternative possibilities can readily be assessed.

The usual additional statistical information can efficiently be computed from either of these factorizations. For example, when X is of full rank then the estimated variance-covariance matrix of the sample regression, V, is defined as

$$V = s^2(X^TX)^{-1}, \tag{7.18}$$

where s^2 is the estimated variance of the residual and from (4.2) this becomes

$$V = s^2(U^TU)^{-1} = s^2U^{-1}U^{-T} \tag{7.19}$$

and the element v_{ij} is given by

$$v_{ij} = s^2e_i^TU^{-1}U^{-T}e_j = s^2f_i^Tf_j, \quad U^Tf_j = e_j \tag{7.20}$$

where e_i is the ith column of the unit matrix. In particular the diagonal elements v_{ii} are the estimated variances of the sample regression coefficients and these can efficiently be computed from

$$v_{ii} = s^2f_i^Tf_i, \quad U^Tf_i = e_i,$$

this requiring just one forward substitution for each f_i.

When X is not of full rank and a minimal length solution has been obtained, so that b is given by (7.14) then in place of (7.18) we must take

$$V = s^2(X^TX)^{\dagger} \tag{7.21}$$

From (4.8) this becomes

$$V = s^2(P\Sigma^2P^T)^{\dagger} = s^2P(\Sigma^2)^{\dagger}P^T$$

$$= s^2P\begin{bmatrix} S^{-2} & 0 \\ 0 & 0 \end{bmatrix}P^T$$

and if we partition P as

$$P = \begin{bmatrix} P_1 & P_2 \end{bmatrix}, \quad P_1 \text{ - p by k}$$

then corresponding to (7.20), here we have

$$v_{ij} = s^2f_i^Tf_j, \quad Sf_j = P_1^Te_j \tag{7.22}$$

and once again the elements of V can be computed efficiently.

As a second example, if x is a p element vector of values of the variables $x_1, x_2, \ldots, x_p$, then the estimated variance of the estimate of the dependent variable $x^T b$ is given by

$$\alpha^2 = s^2 x^T V x \tag{7.23}$$

and this can be computed from

$$\alpha^2 = s^2 f^T f, \quad U^T f = x \tag{7.24}$$

for the QU factorization and

$$\alpha^2 = s^2 f^T f, \quad Sf = P_1^T x \tag{7.25}$$

for the SVD.

Reliable methods for solving regression problems for which $E(rr^T) = \sigma^2 W$ when $W \neq I$ are discussed in Paige, 1978 and 1979 and Kourouklis and Paige, 1981.

Despite the fact that some authors do not approve of the SVD - for instance to quote from Beale (1982) "...the square roots of the pivot elements $\bar{a}_{rr}$ arising in the Gauss-Jordan process, may all be reasonably large and yet A [= $X^T X$] may be effectively singular. This sad fact has led to proposals for the use of Singular Value Decomposition to determine the true number of effectively linearly independent regressor variables. These proposals should be resisted." - when X has near multicollinearities the SVD provides such useful information that its use really should not be resisted, but should instead be encouraged.

8. Principal Components and Canonical Correlations

In this section we give just a brief mention of two further applications of the singular value decomposition in multivariate analysis

Given a zero means data matrix $\hat{X}$, possibly standardized, the aim of a principal component analysis is to determine an orthogonal transformation of the columns of $\hat{X}$ to a data matrix $\hat{Y}$ whose columns have non-increasing variance, each column of $\hat{Y}$ having as large a variance as possible. If we let B be the transformation matrix so that,

$$\hat{Y} = \hat{X}B \tag{8.1}$$

and denote the columns of $\hat{Y}$ and B by $\hat{y}_i$ and b_i respectively then we wish to determine the first principal component $\hat{y}_1$

$$\hat{y}_1 = \hat{X} b_1$$

so that $\hat{y}_1$ has maximum variance subject to $b_1^T b_1 = 1$.

The second principal component $\hat{y}_2$

$$\hat{y}_2 = \hat{X}b_2$$

is determined so that $\hat{y}_2$ has maximum variance subject to $b_2^T b_2 = 1$ and $b_1^T b_2 = 0$ and so on for the remaining components.

Since $\hat{X}$ has zero means so does $\hat{Y}$ and hence maximizing the variance of $\hat{y}_i$ is equivalent to maximizing $\hat{y}_i^T \hat{y}_i$. Thus $\hat{y}_1$ is determined by

$$\text{maximize } \hat{y}_1^T \hat{y}_1 \text{ subject to } b_1^T b_1 = 1 \tag{8.2}$$

If Q is an orthogonal matrix this is equivalent to

$$\text{maximize } z_1^T z_1 \text{ subject to } b_1^T b_1 = 1, \quad z_1 = Q^T \hat{X} b_1 \tag{8.3}$$

If Q is chosen as the left-hand orthogonal matrix of the SVD of $\hat{X}$ then

$$z_1 = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} P^T b_1 \tag{8.4}$$

and if we put

$$a_i = P^T b_i, \quad a_i^T = \begin{bmatrix} a_{1i} & a_{2i} & \cdots & a_{pi} \end{bmatrix}$$

then

$$z_1^T z_1 = a_1^T \begin{bmatrix} S^2 & 0 \\ 0 & 0 \end{bmatrix} a_1 = \sigma_1^2 a_{11}^2 + \sigma_2^2 a_{21}^2 + \cdots + \sigma_k^2 a_{k1}^2$$

$$\leq \sigma_1^2 (a_{11}^2 + a_{21}^2 + \cdots + a_{k1}^2) \leq \sigma_1^2$$

Equality occurs with the choice $a_{11} = 1$, $a_{j1} = 0$, $j > 1$ which gives

$$b_1 = p_1 \text{ and } \hat{y}_1^T \hat{y}_1 = \sigma_1^2 \tag{8.5}$$

It is a straightforward matter to show that

$$b_i = p_i \text{ so that } B = P \tag{8.6}$$

and hence P is the matrix that transforms $\hat{X}$ into its principal components. Notice that

$$\hat{Y}^T \hat{Y} = P^T (\hat{X}^T \hat{X}) P = P^T (P \Sigma^2 P^T) P \tag{8.7}$$

$$= \Sigma^2$$

so that the columns of $\hat{Y}$ are uncorrelated and the estimated variance of $\hat{y}_i$ is $\sigma_i^2/(n-1)$. Notice also that

$$\hat{Y} = \hat{X}P = Q\begin{bmatrix}\Sigma\\0\end{bmatrix}$$

and hence

$$\hat{y}_i = \sigma_i q_i \tag{8.8}$$

Once again the SVD allows us to properly use our judgement as to which components are significant.

Given two zero mean data matrices $\hat{X}$ and $\hat{Y}$ the canonical correlation problem is to find a transformation, A, of the columns of $\hat{Y}$

$$Z = \hat{Y}A \tag{8.9}$$

such that the columns of Z are orthonormal and such that regression of a column of Z on $\hat{X}$ maximizes the multiple correlation coefficient. If we denote the ith columns of A and Z by a_i and z_i respectively and denote the vector of regression coefficients of z_i on $\hat{X}$ by b_i then, since $\hat{X}$ and $\hat{Y}$ have zero column means the multiple correlation coefficient is

$$r_i^2 = (\hat{X}b_i)^T(\hat{X}b_i)/z_i^Tz_i$$

$$= (\hat{X}b_i)^T(\hat{X}b_i), \quad \text{because } z_i^Tz_i = 1$$

Thus to find the first canonical correlation we wish to determine a_1 to

$$\text{maximize: } r_1^2 = (\hat{X}b_1)^T(\hat{X}b_1) \tag{8.10}$$

$$\text{subject to: } (z_1-\hat{X}b_1)^T(z_1-\hat{X}b_1) \text{ being minimum and } z_1^Tz_1 = 1$$

Let the singular value decompositions of $\hat{X}$ and $\hat{Y}$ be given by

$$\hat{X} = Q_X\begin{bmatrix}S & 0\\0 & 0\end{bmatrix}P_X^T, \quad \hat{Y} = Q_Y\begin{bmatrix}\Sigma & 0\\0 & 0\end{bmatrix}P_Y^T \tag{8.11}$$

and partition Q_X, Q_Y and P_Y as

$$Q_X = \begin{bmatrix}\tilde{Q}_X & \hat{Q}_X\end{bmatrix}, \quad Q_Y = \begin{bmatrix}\tilde{Q}_Y & \hat{Q}_Y\end{bmatrix}, \quad P_Y = \begin{bmatrix}\tilde{P}_Y & \hat{P}_Y\end{bmatrix} \tag{8.12}$$

If we take the minimal length solution to the regression problem, this of course being the standard solution when $\hat{X}$ is of full rank, then from (7.14)

$$b_1 = \hat{X}^{\dagger} z_1 \tag{8.11}$$

and (8.10) becomes

$$\text{maximize:} \quad r_1^2 = (\hat{X}\hat{X}^{\dagger} z_1)^T (\hat{X}\hat{X}^{\dagger} z_1) \tag{8.12}$$

$$\text{subject to:} \quad z_1^T z_1 = 1$$

Now put

$$c_i = \Sigma \, \tilde{P}_Y^T a_i \tag{8.13}$$

Bearing in mind that $z_i = \hat{Y} a_i$, it is readily verified that

$$c_i^T c_i = z_i^T z_i \text{ and } c_i^T c_j = z_i^T z_j \tag{8.14}$$

and

$$\hat{X} \hat{X}^{\dagger} z_i = Q_X \begin{bmatrix} \tilde{Q}_X^T \tilde{Q}_Y c_i \\ 0 \end{bmatrix} \tag{8.15}$$

and hence (8.12) becomes

$$\text{maximize:} \quad (\tilde{Q}_X^T \tilde{Q}_Y c_1)^T (\tilde{Q}_X^T \tilde{Q}_Y c_1) \tag{8.16}$$

$$\text{subject to:} \quad c_1^T c_1 = 1$$

Comparison with (8.2) shows that this is equivalent to the problem of finding the first principal component of $\tilde{Q}_X^T \tilde{Q}_Y$. From (8.14) we can see that if we solve the principal component problem for $\tilde{Q}_X^T \tilde{Q}_Y$ then we have solved the canonical correlation problem for the pair $(\hat{X}, \hat{Y})$.

The singular values of $\tilde{Q}_X^T \tilde{Q}_Y$ are the multiple correlation coefficients r_i^2 and these are called the *canonical correlation coefficients*. A full discussion of this and related topics can be found in Björck and Golub, 1973.

The above discussion can be considerably simplified if $\hat{X}$ and $\hat{Y}$ are assumed to be of full rank and we use QU factorizations in place of the SVD (Golub, 1969), but the above discussion is included to indicate the potential power of the SVD as an aid to the solution of difficult multivariate statistical problems.

Many other applications of the singular value decomposition in multivariate analysis are discussed in Chambers, 1977 and Banfield, 1978.

References

[1] BANFIELD C.F. (1978) Singular value decomposition in multivariate Analysis. In "Numerical Software - Needs and Availability." Ed. JACOBS D.A.H., Academic Press, London.

[2] BEALE E.M.L. (1982) Avoiding nonsense in multiple regression computations. The Professional Statistician, 1, 12-13.

[3] BJÖRCK A. and GOLUB G.H. (1973) Numerical methods and computing angles between linear subspaces. Maths. of Computation, 27, 579-594.

[4] BMDP (1977) Biomedical Computer Programs. University of California Press, London.

[5] CHAMBERS J.M. (1977) Computational Methods for Data Analysis. Wiley, New York.

[6] CHAN T.F., GOLUB G.H. and LEVEQUE R.J. (1982) Updating formulae and a pairwise algorithm for computing sample variances. In "COMPSTAT '82, Part I: Proceedings in Computational Statistics." Ed. CAUSSINUS H., ETTINGER P. and TOMASSONE R. Physica-Verlag, Vienna.

[7] CHAN T.F. and LEWIS J.G. (1979) Computing standard deviations: Accuracy. Comm. ACM, 22, 526-531.

[8] COX M.G. (1981) The least squares solution of overdetermined linear equations having band or augmented band structure. IBM J. Numer. Anal., 1, 3-22.

[9] DONGARRA J.J., MOLER C.B., BUNCH J.R. and STEWART G.W. (1979) Linpack Users' Guide. SIAM, Philadelphia.

[10] FORSYTHE G.E. and MOLER C.B. (1967) Computer Solution of Linear Algebraic Equations. Prentice-Hall, New Jersey.

[11] GENTLEMAN W.M. (1974) Basic procedures for large, sparse or weighted linear least squares problems. Appl. Statist., 23, 448-454.

[12] GENTLEMAN W.M. (1974b) Regression problems and the QU decomposition. Bulletin IMA, 10, 195-197.

[13] GILL P.E. and MURRAY W. (1977) Modification of matrix factorizations after a rank-one change. In "The State of the Art of Numerical Analysis." Ed. JACOBS D.A.H. Academic Press, London.

[14] GIVENS J.W. (1958) Computation of plane unitary rotations transforming a general matrix to triangular form. J. SIAM 6, 26-50.

[15] GOLUB G.H. (1965) Numerical methods for solving linear least squares problems. Num. Math., 7, 206-216.

[16] GOLUB G.H. (1969) Matrix decompositions and statistical calculations. In "Statitical Computatation." Eds. MILTON R.C. and NELDER J.A., Academic Press, London.

[17] GOLUB G.H. and KAHAN W. (1965) Calculating the singular values and pseudo-inverse of a matrix. SIAM J. Num. Analysis, 2, 202-224.

[18] GOLUB G.H., KLEMA V.C. and STEWART G.W. (1976) Rank degeneracy and least squares problems. Technical Report STAN-CS-76-559, Stanford University, Stanford, CA94305, USA.

[19] GOLUB G.H. and REINSCH C. (1970) Singular value decomposition and least squares solutions. Num. Math., 14, 403-420.

[20] GOLUB G.H. and WILKINSON J.H. (1966) Note on iterative refinement of least squares solutions. Numer. Math., 9, 139-148.

[21] KLEMA V.C. and LAUB A.J. (1980) The singular value decomposition: its computation and some applications. IEEE Trans. Automat. Control, AC-25, 164-176.

[22] KOUROUKLIS S. and PAIGE C.C. (1981) A constrained least squares approach to the general Gauss-Markov Linear model. J. American Statistical Assoc., 76, 620-625.

[23] LAWSON C.L. and HANSON R.J. (1974) Solving Least Squares Problems. Prentice-Hall, New Jersey.

[24] PAIGE C.C. (1978) Numerically stable computations for general univariate linear models. Commun. Statist.-Simula. Computa., B7(5), 437-453.

[25] PAIGE C.C. (1979) Fast numerically stable computations for generalized linear least squares problems. SIAM J. Numer. Anal., 16, 165-171.

[26] PETERS G. and WILKINSON J.H. (1970) The least squares problem and pseudo-inverses. Computer J., 309-316.

[27] STEWART G.W. (1974) Introduction to Matrix Computations. Academic Press, New York.

[28] STEWART G.W. (1977) On the perturbation of pseudo-inverses, projections and linear least squares problems. SIAM Rev., 4, 634-662.

[29] STEWART G.W. (1979) Assessing the effects of variable error in linear regression. Technical Report 818, University of Maryland, College Park, Maryland 20742, USA.

[30] WEST D.H.D. (1979) Updating mean and variance estimates: An improved method. Comm. ACM, 22, 532-535.

[31] WILKINSON J.H. (1963) Rounding Errors in Algebraic Processes. Notes on Applied Science No. 32, HMSO, London and Prentice-Hall, New Jersey.

[32] WILKINSON J.H. (1965) The Algebraic Eigenvalue Problem. Oxford University Press, London.

[33] WILKINSON J.H. (1974) The classical error analyses for the solution of linear systems. Bulletin IMA, 10, 175-180.

[34] WILKINSON J.H. (1977) Some recent advances in numerical linear algebra. In "The State of the Art in Numerical Analysis." Ed. JACOBS D.A.H. Academic Press, London.

[35] WILKINSON J.H. (1978) Singular-value decomposition - Basic aspects. In "Numerical Software - Needs and Availability." Ed. JACOBS D.A.H. Academic Press, London.