The Singular Value Decomposition in Multivariate Statistics

Sven Hammarling, NAG Central Office,
256 Banbury Road, Oxford OX2 7DE, UK

To Gene Golub who has done so much to encourage and advance the use of stable numerical techniques in multivariate statistics.

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 understanding and analysing 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.

This is a review article aimed at the statistician and mathematician who, while not being expert numerical analysts, would like to gain some understanding of why it is better to work with the data matrix, and of the techniques that allow us to avoid the explicit computation of sums of squares and cross products matrices. To give a focus and to keep the article of moderate length, we concentrate in particular on the use of the singular value decomposition and its application to multiple regression problems. In the final two sections we give a brief discussion of principal components, canonical correlations and the generalized singular value decomposition.

2. Notation

Rather than use the standard notation of numerical linear algebra, we use notation that is more akin to that of the statistician and so we let X denote an n by p data matrix (design matrix, matrix of observation), where p is the number of variables and n is the number of data points (objects, individuals, observations). Let x_j denote the j-th column of X, so that

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

and x_j is the n element vector of sample observations for the j-th variable, let $\bar{x}_j$ and s_j be respectively the sample mean and standard deviation for the j-th variable and denote $\bar{x}$, D and e respectively as

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

where

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

Then the matrix

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

is the <u>zero means data matrix</u> and the matrix

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

is the <u>standardized data matrix</u>, because the mean of each column of $\tilde{X}$ is zero and the variance of each column is unity, unless $s_j = 0$ in which case the j-th column is zero.

The normal matrices X^TX and $\hat{X}^T\hat{X}$ are the <u>sums of squares and cross products matrix</u> and the <u>corrected sums of squares and cross products matrix</u> respectively, the matrix

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

is the <u>sample covariance matrix</u> and

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

is the <u>sample correlation matrix</u>, with r_{ij} as the sample correlation coefficient of variables x_i and x_j. Of course

$$R = DCD$$

and both C and R are symmetric non-negative definite.

The notation $||z||$ and $||Z||$ will be used to denote respectively the Euclidean length of the n element vector z and the spectral or two norm of the n by p matrix Z given by

$$||z|| = (\sum_{i=1}^{n} z_i^2)^{\frac{1}{2}}, \quad ||Z|| = \max_{||z||=1} ||Zz|| = \rho^{\frac{1}{2}}(Z^TZ),$$

where $\rho(Z^TZ)$ denotes the spectral radius (largest eigenvalue) of Z^TZ. The reason for our interest in these particular norms is that when Z is orthogonal then

$$||Zz|| = ||z|| \text{ and } ||Z|| = 1 \quad (Z^TZ = I).$$

A detailed knowledge of the spectral norm of a matrix is not important here and to give a feel for its size in relation to the elements of Z we note that

$$||Z|| \leqslant (\sum_{j=1}^{p} \sum_{i=1}^{n} z_{ij}^2)^{\frac{1}{2}} \leqslant p^{\frac{1}{2}} ||Z||.$$

For much of the time we shall use X generically to represent X or $\hat{X}$ or $\tilde{X}$.

3. Instability in Forming Normal Matrices

For numerical stability it is frequently desirable to avoid forming normal matrices, but instead use algorithms that work directly on the data matrices (see for example Golub, 1965). This can be especially important 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 in the data than the data matrix.

A well known example is provided by the matrix

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

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 $|\epsilon|$ 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

$$Xb = y \tag{3.1}$$

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

$$c(X) = ||X||\ ||X^{-1}|| \tag{3.2}$$

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

$$(X+E)\ (b+e) = y \tag{3.3}$$

satisfies

$$\frac{||e||}{||b+e||} \leq c(X)\ \frac{||E||}{||X||} \tag{3.4}$$

For the spectral norm it can be readily be shown that

$$c(X^TX) = c^2(X), \quad \text{(note that } c(X) \geq 1) \tag{3.5}$$

so that unless $c(X) = 1$, which occurs only when X is orthogonal, X^TX is more sensitive to perturbations than X. From (3.5) 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 (3.4) and (3.5) 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 and the residual (error) vector is small relative to the solution; 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, 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.

Of course normal matrices, particularly correlation matrices, provide vital statistical information, but the methods to be discussed provide ready access to the elements of such matrices.

4. The QU Factorization and the Singular Value Decomposition

In this section we briefly introduce and discuss two tools that allow us to avoid forming normal matrices. These tools are the well known factorizations the QU factorization (or QR factorization, but not to be confused with the QR algorithm) and the singular value decomposition (commonly referred to as the SVD). For simplicity of discussion we shall assume that $n \geqslant p$ so that X has at least as many rows as columns. We shall also not discuss the details of the computational 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 and there are a number of sources of quality software that implement these methods (IMSL, NAG, Dongarra et al, 1979; Chan, 1982).

The QU factorization of a matrix X is given by

$$X = Q \begin{bmatrix} U \\ 0 \end{bmatrix} , \qquad (4.1)$$

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, for example, by Householder transformations, plane rotations, or Gram-Schmidt orthogonalization. (Wilkinson, 1965; Golub, 1965; Stewart, 1974; Golub and Van Loan, 1983.)

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 the Cholesky factor of X^TX. Secondly if we perturb U by a matrix F then

$$Q \begin{pmatrix} U+F \\ 0 \end{pmatrix} = X + E, \quad E = Q \begin{pmatrix} F \\ 0 \end{pmatrix} \tag{4.3}$$

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 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

$$X = Q \begin{pmatrix} \Psi \\ 0 \end{pmatrix} P^T, \tag{4.5}$$

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

$$\Psi = \text{diag}(\psi_i) = \begin{pmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \psi_p \end{pmatrix}$$

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

$$\psi_1 \geq \psi_2 \geq \dots \geq \psi_p \geq 0 \tag{4.6}$$

and we shall assume this to be the case. As with the QU factorization the SVD

always exists and is usually obtained by reducing X to bidiagonal form and then applying a variant of the QR algorithm to reduce this to Ψ. (Golub and Kahan, 1965; Golub and Reinsch, 1970; Wilkinson, 1977 and 1978). The ψ_i, i = 1, 2, ..., p are 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. We have adopted here the notation of Stewart (1984) in avoiding the more usual σ_i for the i-th singular value. If we denote the i-th columns of P and Q by p_i and q_i respectively then equation (4.5) implies that

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

For this factorization we have that

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

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

Ψ 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, but having the advantage that the columns of P are orthonormal. Note that (4.8) implies that

$$||X|| = \psi_1. \tag{4.9}$$

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

$$Q \begin{bmatrix} \Psi+F \\ 0 \end{bmatrix} P^T = X + E, \quad E = Q \begin{bmatrix} F \\ 0 \end{bmatrix} P^T \tag{4.10}$$

and

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

so that again perturbations of order ϵ in Ψ 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

$$\psi_{k+1} = \psi_{k+2} = \ldots = \psi_p = 0$$

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

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

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

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

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

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

so that if the elements of E are small then the singular values $\psi_{k+1}, \psi_{k+2}, \ldots, \psi_p$ 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

$$U = \begin{pmatrix} 1 & -1 & -1 & \ldots & -1 & -1 & -1 \\ 0 & 1 & -1 & \ldots & -1 & -1 & -1 \\ 0 & 0 & 1 & \ldots & -1 & -1 & -1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & -1 & -1 \\ 0 & 0 & 0 & \ldots & 0 & 1 & -1 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 1 \end{pmatrix}$$

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

$$E = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ : & : & & : \\ : & : & & : \\ 0 & 0 & \dots & 0 \\ -2^{2-p} & 0 & \dots & 0 \end{pmatrix}$$

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

$$\psi_p \leqslant 2^{2-p}$$

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

When the SVD is computed by numerically stable methods then 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.ψ_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 = $2^{-24} \approx 6 \times 10^{-8}$, the smallest singular value of the above matrix U, as computed by the NAG Library routine F02WDF, was 4.726×10^{-8} instead of $\psi_{32} = 5.280 \times 10^{-10}$.

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-emphasized.

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

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

where $\tilde{Q}$ and P^T are orthogonal and Ψ is diagonal. From (4.5) we get that the SVD of X is then given by

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

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 Library routine F02WDF is an example that explicitly allows the user to stop at the QU factorization if U is not too ill-conditioned.

Particularly important in some statistical and real time applications is the fact that the QU factorization may be obtained by processing the matrix X one observation, or a block of observations at a time, so that the complete matrix X need not be held all at once, but can be sequentially processed to give the compact representation U. This can be achieved by well known updating techniques using, for example, plane rotations. (Golub, 1965; Gentleman, 1974a; Gill and Murray, 1977; Dongarra et al, 1979; Cox, 1981.)

In the next section we demonstrate that such techniques can also be used to obtain the QU factorization of $\hat{X}$ and of $\tilde{X}$.

5. 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 (1977) allow one to form covariance and correlation matrices by sequentially processing the data and we now show that we can also obtain the QU factorization of such matrices by a corresponding process.

First we 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 LeVeque, 1982.) If we denote the i-th observation of the j-th 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

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

then it is readily verified that

$$\bar{x}_j^{(r)} = \bar{x}_j^{(r-1)} + (x_j^{(r)} - \bar{x}_j^{(r-1)})/r \tag{5.1}$$

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

Of course

$$\bar{x}_j = \bar{x}_j^{(n)} \quad \text{and} \quad s_j^2 = \gamma_j^{(n)}/(n-1) \tag{5.2}$$

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

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

$$= Q\left\{ \begin{bmatrix} U \\ 0 \end{bmatrix} - f\bar{x}^T \right\}, \quad f = Q^T e \tag{5.3}$$

If we find the QU factorization of the matrix in braces as, say

$$\begin{bmatrix} U \\ 0 \end{bmatrix} - f\bar{x}^T = Q_1 \begin{bmatrix} \hat{U} \\ 0 \end{bmatrix} \tag{5.4}$$

and put

$$\hat{Q} = QQ_1 \tag{5.5}$$

then the QU factorization of $\hat{X}$ is

$$\hat{X} = \hat{Q} \begin{bmatrix} \hat{U} \\ 0 \end{bmatrix}. \qquad (5.6)$$

The factorization of (5.4) is a rank one update problem and there are standard methods by which the factorization can be obtained economically (Gill and Murray, 1977; Golub and Van Loan, 1983, section 12-6). Since we can obtain the QU factorization of X by sequentially processing the data and noting that $\hat{U}D$ is still upper triangular, equations (5.1), (5.2) and (5.6) enable us to find the QU factorizations of $\hat{X}$ and of $\tilde{X}$, and hence the Cholesky factors of the matrices C and R of (2.5) and (2.6), by sequentially processing the data one or more observations at a time.

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, let z_r^T denote the r-th row of X and define $\bar{x}^{(r)}$ as the vector

$$(\bar{x}^{(r)})^T = [\bar{x}_1^{(r)} \; \bar{x}_2^{(r)} \; \ldots \; \bar{x}_p^{(r)}]$$

and take the number of elements in the vector e by context, then using (5.1)

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

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

so that

$$\hat{X}_{r+1} = \begin{bmatrix} \hat{X}_r - \frac{1}{r+1} e \, (z_{r+1} - \bar{x}^{(r)})^T \\ (z_{r+1} - \bar{x}^{(r+1)})^T \end{bmatrix} \qquad (5.8)$$

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 $(z_{r+1} - \bar{x}^{(r+1)})^T$. In either method 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. As described earlier we can readily obtain the SVD of $\hat{X}$ or $\tilde{X}$ via the upper triangular factor.

6. Solving Multiple Regression Problems

In this section we consider the application of the QU factorization and the SVD to multiple regression, or linear least squares, and 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 the other.

We wish to determine the vector b (the regression coefficients) to

$$\text{minimize } r^T r, \quad \text{where } y = Xb + r, \tag{6.1}$$

where y is a vector of dependent observations 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 (6.1) is equivalent to

$$\text{minimize } \tilde{r}^T \tilde{r}, \quad \text{where } Q^T y = Q^T X b + \tilde{r}, \quad \tilde{r} = Q^T r. \tag{6.2}$$

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

$$Q^T y = \begin{pmatrix} \tilde{y} \\ w \end{pmatrix}, \quad \tilde{y} \text{ a p element vector} \tag{6.3}$$

then

$$\tilde{r} = \begin{pmatrix} \tilde{y} - Ub \\ w \end{pmatrix} \tag{6.4}$$

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

$$Ub = \tilde{y}. \tag{6.5}$$

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

$$\tilde{r} = \begin{bmatrix} 0 \\ w \end{bmatrix} \quad \text{so that } r^Tr = w^Tw \tag{6.6}$$

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^Tw on completion of the QU factorization.

As the discussion in section 3 indicates, the sensitivity of the solution of (6.5) 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 (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 (6.4) becomes

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

so that (6.1) is equivalent to

$$\text{minimize } \hat{r}^T\hat{r}, \quad \text{where } \hat{r} = \tilde{Q}^T\tilde{y} - \Psi P^Tb. \tag{6.7}$$

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

$$\Psi = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}, \quad S - k \text{ by } k \text{ and non-singular} \tag{6.8}$$

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

$$\tilde{Q}^T\tilde{y} = \begin{bmatrix} \hat{y} \\ v \end{bmatrix}, \quad P^Tb = \begin{bmatrix} \hat{c} \\ \tilde{c} \end{bmatrix}, \quad \hat{y}, \hat{c} - k \text{ element vectors.} \tag{6.9}$$

Then

$$\hat{r} = \begin{bmatrix} \hat{y}-S\hat{c} \\ v \end{bmatrix} \tag{6.10}$$

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

$$S\hat{c} = \hat{y}. \tag{6.11}$$

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

$$r^Tr = w^Tw + \hat{r}^T\hat{r} = w^Tw + v^Tv. \tag{6.12}$$

As $\tilde{c}$ is not determined by (6.11) we can see that the solution is not unique. The particular solution for which b^Tb is also a minimum is called the minimal length solution and from (6.9) we see that this is given by taking

$$\tilde{c} = 0 \text{ in which case } b = P \begin{bmatrix} \hat{c} \\ 0 \end{bmatrix} \tag{6.13}$$

(Golub and Reinsch, 1970; Peters and Wilkinson, 1970).

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; Stewart, 1984) equations (4.10) - (4.13) tell us about the effects of neglecting small singular values. Furthermore (4.7) gives

$$||Xp_i|| = \psi_i \tag{6.14}$$

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

$$V = s^2(X^TX)^{-1}, \tag{6.15}$$

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

$$V = s^2(U^TU)^{-1} = s^2U^{-1}U^{-T} \tag{6.16}$$

and the element v_{ij} is given by

$$v_{ij} = s^2 e_i^T U^{-1} U^{-T} e_j = s^2 f_i^T f_j, \quad U^T f_j = e_j \tag{6.17}$$

where e_i is the i-th 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

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

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, then in place of (6.15) we must take

$$V = s^2 (X^T X)^\dagger \tag{6.18}$$

where $(X^T X)^\dagger$ is the pseudo-inverse of $X^T X$, and from (4.8) this becomes

$$V = s^2 (P \Psi^2 P^T)^\dagger = s^2 P (\Psi^2)^\dagger P^T$$

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

and if we partition P as

$$P = [P_1 \quad P_2], \quad P_1 \text{ - p by k}$$

then corresponding to (6.17), here we have

$$v_{ij} = s^2 f_i^T f_j, \quad S f_j = P_1^T e_j \tag{6.19}$$

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, ..., x_p, then the estimated variance of the estimate of the dependent variable $x^T b$ is given by

$$\alpha^2 = s^2 x^T V x \tag{6.20}$$

and this can be computed from

$$\alpha^2 = s^2 f^T f, \quad U^T f = x \tag{6.21}$$

for the QU factorization and

$$\alpha^2 = s^2 f^T f, \quad Sf = P_1^T x \tag{6.22}$$

for the SVD.

If we relax the assumption that $E(rr^T) = \sigma^2 I$ and instead have

$$E(rr^T) = \sigma^2 W,$$

where W is non-negative definite, then the usual approach is to obtain the regression coefficients as the solution of the weighted least squares problem

$$\text{minimize } r^T W^{-1} r, \quad \text{where } y = Xb + r \tag{6.23}$$

because $E(W^{-1} rr^T) = \sigma^2 I$. Unless W is well-conditioned, solving (6.23) explicitly is numerically unstable and the problem is not even defined when W is singular. If we let F be any matrix such that

$$W = FF^T$$

and let e be an error vector satisfying

$$Fe = r,$$

then (6.23) is equivalent to the generalized linear least squares problem

$$\begin{aligned} &\text{minimize } e^T e \\ &\text{subject to } y = Xb + Fe \end{aligned} \tag{6.24}$$

and now W is not required to be non-singular. Methods for the solution of (6.24) based on the QU factorization and on the SVD have been discussed by Paige (1978, 1979a, 1979b) and by Korouklis and Paige (1981). To briefly indicate how the SVD may be used, partition Q as

$$Q = [Q_1 \; Q_2]$$

Then multiplying the linear constraints in (6.24) by Q^T and using the notation of (6.8) and (6.9) we find that

$$S\hat{c} = Q_1^T y - Q_1^T B e,$$

from which $\hat{c}$ is determined from e, $\tilde{c}$ is arbitrary and e must satisfy

$$Q_2^T y = Q_2^T B e.$$

An SVD of $Q_2^T B$ either allows e to be determined, or shows that the equations are inconsistent (Hammarling, Long and Martin, 1983).

When X, as well as y, contains experimental error then in place of (6.1) it may be more appropriate to find the regression coefficients as the solution of the <u>total least squares</u> problem (Golub and Van Loan, 1980)

$$\text{minimize } ||(E,r)||_F^2, \quad \text{where } y = (X+E)b + r \tag{6.25}$$

and $||X||_F^2 = \sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2$. If we assume that X is of full rank and put

$$Z = (X,-y), \quad F = (E,r)$$

then (6.25) becomes

$$\text{minimize } ||F||_F^2, \quad \text{where } (Z+F) \begin{pmatrix} b \\ 1 \end{pmatrix} = 0 \tag{6.26}$$

and so we require the minimum perturbation that makes Z rank deficient with $[b^T \; 1]^T$ in the null space of (Z+F). If we let the SVD of Z be

$$Z = Q \begin{pmatrix} \Psi & 0 \\ 0 & \psi_{p+1} \\ 0 & 0 \end{pmatrix} P^T \quad \text{and put } F = Q \begin{pmatrix} 0 & 0 \\ 0 & -\psi_{p+1} \\ 0 & 0 \end{pmatrix} P^T$$

then F makes (Z+F) rank deficient and is of minimum norm (Golub and Van Loan, 1983, corollary 2.3-3). If the (p+1)th right singular vector (last column of P) is denoted by $[p^T \; \rho]^T$, and if $\rho \neq 0$ then it is readily verified that the regression coefficients are given by

$$b = (1/\rho)p. \tag{6.27}$$

For further details see Golub and Van Loan (1980, 1983) and for discussion of the case where $\rho = 0$ and a comparison with standard regression see Van Huffel, Vandewalle and Staar (1984).

7. Other Applications in Multivariate Analysis

In this section we give a very brief mention of two further applications of the SVD 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.

It is well known, and in any case readily established from the Courant-Fischer theorem (Wilkinson, 1965, chapter 2, section 43), that $\hat{Y}$ is given by

$$\hat{Y} = \hat{X}P, \tag{7.1}$$

where P is the matrix of eigenvectors of $\hat{X}^T\hat{X}$. From (4.8) and (4.5) we therefore see that P is the matrix of right singular vectors of $\hat{X}$ and that

$$\hat{Y} = Q \begin{bmatrix} \Psi \\ 0 \end{bmatrix} \tag{7.2}$$

so that the j-th principal component of $\hat{X}$ is given by

$$\hat{y}_j = \psi_j q_j, \tag{7.3}$$

where q_j is the j-th left singular vector of $\hat{X}$, and the estimated variance of $\hat{y}_j$ is $\psi_j^2/(n-1)$. Again, we can avoid the formation of $\hat{X}^T\hat{X}$ and 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 $\hat{Y}$

$$Z = \hat{Y}A \tag{7.4}$$

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.

Let the SVD's of $\hat{X}$ and $\hat{Y}$ be

$$\hat{X} = Q_x \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} P_x^T, \quad \hat{Y} = Q_y \begin{bmatrix} \Psi & 0 \\ 0 & 0 \end{bmatrix} P_y^T, \tag{7.5}$$

partition Q_x, Q_y and P_y as

$$Q_x = [\tilde{Q}_x \; \hat{Q}_x], \quad Q_y = [\tilde{Q}_y \; \hat{Q}_y], \quad P_y = [\tilde{P}_y \; \hat{P}_y]$$

and let

$$C = \Psi \tilde{P}_y^T A. \tag{7.6}$$

This gives

$$Z = Q_y \begin{bmatrix} \Psi \tilde{P}_y^T A \\ 0 \end{bmatrix} \quad \text{so that} \quad Z^T Z = C^T C \tag{7.7}$$

and hence if the columns of Z are orthonormal then so are the columns of C. Noting that

$$\hat{X}\hat{X}^{\dagger} Z = Q_x \begin{bmatrix} \tilde{Q}_x^T \tilde{Q}_y \\ 0 \end{bmatrix} C, \tag{7.8}$$

it is now a straightforward matter to show that the canonical correlation problem for the pair $(\hat{X},\hat{Y})$ can be solved from the solution to the principal component problem for the matrix $\tilde{Q}_x^T \tilde{Q}_y$. The multiple correlation coefficients, or the canonical correlation coefficients, are the singular values of $\tilde{Q}_x^T \tilde{Q}_y$. The canonical correlation example is included to indicate the potential power of the SVD as an aid to the solution of difficult multivariate statistical problems.

A full discussion of this and related topics is given by Björck and Golub (1973). See also Golub and Van Loan (1983, section 12.4). When $\hat{X}$ and $\hat{Y}$ are both of full rank then we can use the QU factorization of $\hat{X}$ and $\hat{Y}$ in place of their SVD's (Golub, 1969).

A number of other applications of the SVD in multivariate analysis are discussed by Chambers (1977) and by Banfield (1978).

8. The Generalized Singular Value Decomposition

Here we briefly mention an important generalization of the SVD that is relevant to a pair of data matrices (X,Y) of dimension n by p and m by p. To simplify the discussion we assume that $m \geq p$ and that Y is of full rank. In this case the generalized singular value decomposition (GSVD) is given by

$$X = Q_X \begin{pmatrix} \Psi_X & 0 \\ 0 & 0 \end{pmatrix} Z^{-1}, \quad Y = Q_Y \begin{pmatrix} \Psi_Y & 0 \\ 0 & I \end{pmatrix} Z^{-1}, \tag{8.1}$$

where Q_X and Q_Y are orthogonal and Ψ_X and Ψ_Y are diagonal matrices, $\Psi_X = \text{diag}(\alpha_i)$, $\Psi_Y = \text{diag}(\beta_i)$ with

$$\alpha_i^2 + \beta_i^2 = 1, \quad \alpha_i \geq 0, \quad \beta_i > 0. \tag{8.2}$$

The pairs (α_i,β_i) are called the generalized singular values and can be chosen with the α_i in descending order and the β_i in ascending order (Van Loan, 1976). This and the unrestricted case are discussed by Paige and Saunders (1981) and we strongly recommend reference to their paper.

From (8.1) we see that

$$X^TX = Z^{-T} \begin{pmatrix} \Psi_X^2 & 0 \\ 0 & 0 \end{pmatrix} Z^{-1}, \quad Y^TY = Z^{-T} \begin{pmatrix} \Psi_Y^2 & 0 \\ 0 & 0 \end{pmatrix} Z^{-1} \tag{8.3}$$

so that Z is the congruence matrix that simultaneously diagonalizes the normal matrices (X^TX, Y^TY) and hence the columns of Z are the eigenvectors of the generalized symmetric eigenvalue problem

$$(X^TX)z = \lambda(Y^TY)z \tag{8.4}$$

and the (α_i^2/β_i^2) are the corresponding eigenvalues. Thus, just as the SVD allows us to avoid the numerically damaging step forming X^TX, the GSVD allows us to avoid the numerically damaging step of forming the pair (X^TX, Y^TY).

Unlike the SVD there is not yet quality software available for computing the GSVD, but numerically stable algorithms are beginning to emerge (Stewart, 1983; Paige, 1984b) and such software will surely be available in the near future. This will mean that we can use the natural tool for tackling multivariate problems involving

matrix pairs (X,Y), rather than using the SVD, which is really only the natural tool when a single data matrix is involved.

Two such examples are the generalized least squares problem and the canonical correlation problem, discussed in the previous two sections. Paige (1984a) has given an elegant analysis of the generalized least squares problem in terms of the GSVD, and for the canonical correlation problem it can readily be shown that we simply have to replace the Q_x and Q_y of (7.5) by those of (8.1) and then we again solve the principal component problem for the matrix $\tilde{Q}_x^T\tilde{Q}_y$.

9. Acknowledgement

This is a revised version of an article that first appeared as "How to live without covariance matrices: Numerical stability in multivariate statistical analysis", in NAG Newsletter 1/83.

10. References

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

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

BMPD (1977). Biomedical Computer Programs. University of California Press, London.

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

CHAN, T.F. (1982). Algorithm 581: An improved algorithm for computing the singular value decomposition. ACM Trans. Math. Softw. 8, 84-88.

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". Eds. Caussinus, H., Ettinger, P. and Tomassone, R. Physica-Verlag, Vienna.

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

COX, M.G. (1981). The least squares solution of overdetermined linear equation shaving band or augmented band structure. IMA J. Num. Anal., 1, 3-22.

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

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

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

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

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.

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

GOLUB, G.H. (1969). Matrix decompositions and statistical calculations. In "Statistical Computation". Eds. Milton, R.C. and Nelder, J.A., Academic Press, London.

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

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, CA 94305, USA.

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

GOLUB, G.H. and VAN LOAN, C.F. (1980). An analysis of the total least squares problem. SIAM J. Num. Anal., 17, 883-893.

GOLUB, G.H. and VAN LOAN, C.F. (1983). Matrix Computations. North Oxford Academic, Oxford.

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

HAMMARLING, S.J., LONG, E.M.R. and MARTIN, D.W. (1983). A generalized linear least squares algorithm for correlated observations, with special reference to degenerate data. NPL Report DITC 33/83. National Physical Laboratory, Teddington, Middlesex, TW11 OLW, UK.

IMSL, International Mathematical and Statistical Libraries, 7500 Bellaire Blvd., Houston, TX 77036-5085, USA.

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.

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.

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

NAG, Numerical Algorithms Group, NAG Central Office, 256 Banbury Road, Oxford, OX2 7DE, UK.

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

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

PAIGE, C.C. (1984a). The general linear model and the generalized singular value decomposition. School of Computer Science, McGill University, Montreal, Quebec, Canada, H3A 2K6. (Submitted to Linear Algebra Applic., Special Issue on Statistics.)

PAIGE, C.C. (1984b). Computing the generalized singular value decomposition. School of Computer Science, McGill University, Montreal, Quebec, Canada, H3A 2K6. (Submitted to SIAM J. Sci. Stat. Comput.)

PAIGE, C.C. and SAUNDERS, M.A. (1981). Towards a generalized singular value decomposition. SIAM J. Num. Anal., 18, 398-405.

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

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

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

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

STEWART, G.W. (1983). A method for computing the generalized singular value decomposition. In "Matrix Pencils". Eds. Kagstrom, B. and Ruhe, A., Springer-Verlag, Berlin.

STEWART, G.W. (1984). Rank degeneracy. SIAM J. Sci. Stat. Comput., 5, 403-413.

VAN HUFFEL, S., VANDEWALLE, J. and STAAR, J. (1984). The total linear least squares problem: Formulation, algorithm and applications. Katholieke Universiteit Leuven, ESAT Laboratory, 3030 Heverlee, Belgium.

VAN LOAN, C.F. (1976). Generalizing the singular value decomposition. SIAM J. Num. Anal., 13, 76-83.

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

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

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

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

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.

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