NBER WORKING PAPER SERIES

ROSEPACK Document No. 4

Rank Degeneracy and Least Squares Problems

Gene Golub*
Stanford University

Virginia Klema**
National Bureau of Economic Research

G. W. Stewart+
University of Maryland

Working Paper No. 165

Computer Research Center for Economics and Management Science
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

February 1977

*Supported in part by Contract No. Army DAHC04-75-G-0185 NSF DCR75-13497.

**Supported in part by the National Science Foundation under Contract No. DCR-75-08802.

+Supported in part by the Office of Naval Research under Contract No. N00014-76-C-0391.

1. Introduction

In this paper we shall be concerned with the following problem. Let A be an $m \times n$ matrix with $m \geq n$, and suppose that A is near (in a sense to be made precise later) a matrix B whose rank is less than n. Can one find a set of linearly independent columns of A that span a good approximation to the column space of B?

The solution of this problem is important in a number of applications. In this paper we shall be chiefly interested in the case where the columns of A represent factors or carriers in a linear model which is to be fit to a vector of observations b. In some such applications, where the elements of A can be specified exactly (e.g. the analysis of variance), the presence of rank degeneracy in A can be dealt with by explicit mathematical formulas and causes no essential difficulties. In other applications, however, the presence of degeneracy is not at all obvious, and the failure to detect it can result in meaningless results or even the catastrophic failure of the numerical algorithms being used to solve the problem.

The organization of this paper is the following. In the next section we shall give a precise definition of approximate degeneracy in terms of the singular value decomposition of A. In Section 3 we shall show that under certain conditions there is associated with A a subspace that is insensitive to how it is approximated by various choices of the columns of A, and in Section 4 we shall apply this result to the solution of the least squares problem. Sections 5, 6, and 7 will be concerned with algorithms for selecting a basis for the stable subspace from among the

columns of A.

The ideas underlying our approach are by no means new. We use the singular values of the matrix A to detect degeneracy and the singular vectors of A to rectify it. The squares of the singular values are the eigenvalues of the correlation matrix A^TA, and the right singular vectors are the eigenvectors of A^TA, that is the principal components of the problem. The use of principal components to eliminate colinearities has been proposed in the literature (e.g. see [4,9,16,17]). This paper extends these proposals in two ways. First we prove theorems that express quantitatively the results of deciding that certain columns of A can be ignored. Second we describe in detail how existing computational techniques can be used to realize our methods.

A word on notation is appropriate here. We have assumed a linear model of the form $b = Ax + e$, where b is an m-vector of observations and x is an n-vector of parameters. This is in contrast to the usual statistical notation in which the model is written in the form $y = X\beta + e$, where y is an n-vector of observations and β is a p-vector of parameters. The reason for this is that we wish to draw on a body of theorems and algorithms from numerical linear algebra that have traditionally been couched in the first notation. We feel that this dichotomy in notation between statisticians and numerical analysts has hindered communication between the two groups. Perhaps a partial solution to this problem is the occasional appearance of notation from numerical analysis in statistical journals and vice versa, so that each group may have a chance to learn the other's notation.

Throughout this paper we shall use two norms. The first is the Euclidean vector norm $\|\cdot\|_2$ defined for an n-vector x by

$$\|x\|_2^2 = \sum_{i=1}^{n} x_i^2$$

and its subordinate matrix norm defined by

$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2.$$

The second is the Frobenius matrix norm defined for the $m \times n$ matrix A by

$$\|A\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2.$$

Both these norms are consistent in the sense that

$$\|AB\|_p \leq \|A\|_p \|B\|_p \quad (p = 2,F)$$

whenever the product AB is defined. They are also unitarily invariant; that is if U and V are orthogonal matrices then

$$\|A\|_p = \|U^T A\|_p = \|AV\|_p \quad (p = 2,F).$$

For more on these matrix norms see [14].

2. Rank Degeneracy

The usual mathematical notion of rank is not very useful when the matrices in question are not known exactly. For example, suppose that A is an $m \times n$ matrix that was originally of rank $r < n$ but whose elements have been perturbed by some small errors (e.g. rounding or measurement errors). It is extremely unlikely that these errors will conspire to keep the rank of A exactly equal to r; indeed what is most likely is

that the perturbed matrix will have full rank n. Nonetheless, the nearness of A to a matrix of defective rank will often cause it to behave erratically when it is subjected to statistical and numerical algorithms.

One way of circumventing the difficulties of the mathematical definition of rank is to specify a tolerance and say that A is numerically defective in rank if to within that tolerance it is near a defective matrix. Specifically we might say that A has ε-rank r with respect to the norm $\|\cdot\|$ if

$$r = \inf \{\operatorname{rank}(B): \|A-B\| \leq \varepsilon\}. \tag{2.1}$$

However, this definition has the defect that a slight increase in ε can decrease the numerical rank. What is needed is an upper bound on the values of ε for which the numerical rank remains at least equal to r. Such a number is provided by any number δ satisfying

$$\varepsilon < \delta \leq \sup \{\eta: \|A-B\| \leq \eta \Rightarrow \operatorname{rank}(B) \geq r\}. \tag{2.2}$$

Accordingly we make the following definition.

Definition 2.1. A matrix A has numerical rank (δ,ε,r) with respect to the norm $\|\cdot\|$ if δ,ε, and r satisfy (2.1) and (2.2).

When the norm in definition 2.1 is either the 2-norm or the Frobenius norm, the problem of determining the numerical rank of a matrix can be solved in terms of the singular value decomposition of the matrix. This decomposition, which has many applications (e.g. see [7]), is described in the following theorem.

Theorem 2.2. Let A be an $m \times n$ matrix with $m \geq n$. Then there is an orthogonal matrix U or order m and an orthogonal matrix V of order n such that

$$U^T A V = \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} \tag{2.3}$$

where

$$\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$$

and

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

For proofs of this theorem and the results cited below see [14]. The numbers $\sigma_1, \sigma_2, \ldots, \sigma_n$, which are unique, are called the singular values of A. The columns $u_1, u_2, \ldots, u_m$ of U are called the left singular vectors of A, and the columns $v_1, v_2, \ldots, v_n$ are called the right singular vectors of A. The matrix A has rank r if and only if

$$\sigma_r > 0 = \sigma_{r+1}, \tag{2.4}$$

in which case the vectors $u_1, u_2, \ldots, u_r$ form an orthonormal basis for the column space of A (hereafter denoted by $R(A)$).

It is the intimate relation of the singular values of a matrix to its spectral and Frobenius norms that enables us to characterize numerical rank in terms of singular values. Specifically the spectral norm of A is given by the expression.

$$\|A\|_2 = \sigma_1.$$

Moreover, if $\tau_1 \geq \tau_2 \geq \ldots \geq \tau_n$ are the singular values of $B = A + E$, then

$$|\sigma_i - \tau_i| \leq \|E\|_2 \quad (i = 1,2,\ldots,n).$$

In view of (2.4) this implies that

$$\inf_{\text{rank}(B)\leq r} \|A-B\|_2 = \sigma_{r+1}, \tag{2.5}$$

and this infimum is actually attained for the matrix B defined by

$$B = U \begin{pmatrix} \Sigma' \\ 0 \end{pmatrix} V^T, \tag{2.6}$$

where $\Sigma' = \text{diag}(\sigma_1,\sigma_2,\ldots,\sigma_r,0,\ldots,0)$.

Likewise

$$\|A\|_F^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2,$$

and

$$\inf_{\text{rank}(B)\leq r} \|A-B\|_F^2 = \sigma_{r+1}^2 + \ldots + \sigma_n^2.$$

The infimum is attained for the matrix B defined by (2.6).

Using these facts we can characterize the notion of numerical rank. In the following theorem we use the notation $\text{rank}\ (\delta,\varepsilon,r)_p$ to mean numerical rank with respect to the norm $\|\cdot\|_p$.

Theorem 2.3. Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$ be the singular values of A. Then A has numerical rank $(\delta,\varepsilon,r)_2$ if and only if

$$\sigma_r \geq \delta > \varepsilon \geq \sigma_{r+1}. \tag{2.7}$$

Also A has numerical rank $(\delta,\varepsilon,r)_F$ if and only if

$$\sigma_r^2 + \sigma_{r+1}^2 + \ldots + \sigma_n^2 \geq \delta^2 > \varepsilon^2 \geq \sigma_{r+1}^2 + \ldots + \sigma_n^2.$$

Proof. We prove the result for the spectral norm, the proof for the Frobenius norm being similar. First suppose that (2.7) holds. Then by (2.5) if $\|B-A\|_2 < \delta$ we must have rank $(B) \geq r$. Consequently δ satisfies (2.2). This also shows that

$$\min \{\text{rank}(B): \|B-A\| \leq \varepsilon\} \geq r.$$

But the matrix B of (2.6) is of rank r and satisfies $\|A-B\|_2 \leq \varepsilon$. Hence ε satisfies (2.1).

Conversely, suppose $\delta, \varepsilon,$ and r satisfy (2.1) and (2.2). Then by (2.5), $\delta \leq \sigma_r$. Also $\varepsilon \geq \sigma_{r+1}$; for if not by (2.1) there is a matrix B of rank r satisfying $\|A-B\| < \sigma_{r+1}$, which contradicts (2.5). □

Because of the simplicity of the characterization (2.7) we shall restrict ourselves to rank defectiveness measured in terms of the spectral norm.

We shall need two other facts about singular values in the sequel. First define

$$\text{inf}(A) = \inf_{\|x\|_2=1} \|Ax\|_2. \tag{2.8}$$

Then

$$\text{inf}(A) = \sigma_n$$

where σ_n is the smallest singular value of A. Second, let X and Y be any matrices with orthonormal columns and let $\tau_1 \geq \tau_2 \geq \ldots \geq \tau_k$ be

the singular values of $C = X^TAY$. Then

$$\sigma_i \geq \tau_i \qquad (i = 1,2,\ldots,k) \tag{2.9}$$

and

$$\tau_{k-i+1} \geq \sigma_{n-i+1} \qquad (i = 1,2,\ldots,k). \tag{2.10}$$

3. The ε-Section of $R(A)$

Having confirmed that a matrix A has numerical rank $(\delta,\varepsilon,r)_2$ with $r < n$, one must decide what to do about it. If the singular value decomposition has been computed as a preliminary to determining the numerical rank, one solution naturally presents itself. This is to work with the matrix B defined by (2.6). Because B has an explicit representation in terms of Σ', the usual difficulties associated with zero singular values can be avoided. Moreover, the solution so obtained is the exact solution of a small perturbation of A.

However, this solution has the important defect that it does not reduce the size of the problem. For example, if the problem at hand is to approximate a vector of observations b, the procedure sketched above will express the approximation as a linear combination of all the columns of A, even though some of them are clearly redundant. What is needed is a device for selecting a set of r linearly independent columns of A. In Sections 5 and 6 we shall discuss numerical techniques for actually making such a selection. In this section and the next we shall concern ourselves with the question of when making such a selection is sensible.

The main difficulty is that there are many different sets of r

linearly independent columns of the matrix A, and not all these sets may be suitable for the problem at hand. For example, if the problem is again that of approximating a vector of observations b, then for each set of columns we shall attempt to find a vector in the subspace spanned by the columns that is in some sense a best approximation to b. Now if the subspace determined by a set varies widely from set to set, then our approximation to b will not be stable. Therefore, we turn to the problem of determining when these subspaces are stable.

We shall attack the problem by comparing the subspaces with a particular subspace that is determined by the singular value decomposition. Let A have numerical rank (δ,ε,r). Let the matrix U in (2.3) be partitioned in the form

$$U = (U_\varepsilon, \hat{U}_\varepsilon),$$

where U_ε has the r columns $u_1,u_2,\ldots,u_r$. Then we shall call $R(U_\varepsilon)$ the <u>ε-section</u> of $R(A)$. Note that the ε-section of $R(A)$ is precisely the column space of the matrix B defined in (2.6).

We shall compare subspaces in terms of the difference of the orthogonal projections upon them. Specifically for any matrix X let P_X denote the orthogonal projection onto $R(X)$. Then for two subspaces $R(X)$ and $R(Y)$ we shall measure the distance between them by $\|P_X - P_Y\|_2$ (for the various geometric interpretations of this number, which is related to canonical correlations and the angle between subspaces, see [1,2,13]). It is known that if Y has orthonormal columns and $\hat{X}$ has orthonormal

columns spanning the orthogonal complement of $R(X)$, then

$$\|P_X - P_Y\|_2 = \|\hat{X}^T Y\|_2. \tag{3.1}$$

The selection of r columns $a_{i_1}, a_{i_2}, \ldots, a_{i_r}$ from the matrix $A = (a_1, a_2, \ldots, a_n)$ has the following matrix interpretation. Let W be the $n \times r$ matrix formed by taking columns $i_1, i_2, \ldots, i_r$ from the $n \times n$ identity matrix. Then it is easily verified that $(a_{i_1}, a_{i_2}, \ldots, a_{i_r}) = AW$. Of course $W^T W = I$, so that W has orthonormal columns, and this is all that is needed for the following comparison theorem.

Theorem 3.1. Let A have numerical rank $(\delta, \varepsilon, r)_2$ and let U_ε be defined as above. Let W be an $n \times r$ matrix with orthonormal columns and suppose that

$$\gamma = \inf(AW) > 0, \tag{3.2}$$

where $\inf(X)$ is defined by (2.8). Then

$$\|P_{U_\varepsilon} - P_{AW}\|_2 \leq \varepsilon/\gamma. \tag{3.3}$$

Proof. The matrix $W^T A^T A W$ is positive definite and hence has a nonsingular positive definite square root. Set $Y = AW(W^T A^T AW)^{-1/2}$. It is easily verified that Y has orthonormal columns spanning $R(AW)$. Moreover, from (3.2)

$$\|(W^T A^T AW)^{-1/2}\|_2 = \gamma^{-1}. \tag{3.4}$$

The matrix $\hat{U}_\varepsilon$ also has orthonormal columns, and they span the orthogonal

complement of $R(U_\varepsilon)$. It follows from (2.3) that

$$\|U_\varepsilon^T A\|_2 \leq \varepsilon. \tag{3.5}$$

Hence from (3.1), (3.4), and (3.5)

$$\begin{aligned} \|P_{U_\varepsilon} - P_{AW}\|_2 &= \|\hat{U}_\varepsilon^T AW(W^T A^T AW^T)^{-1/2}\|_2 \\ &\leq \|U_\varepsilon^T A\|_2 \|W\|_2 \|(W^T A^T AW)^{-1/2}\|_2 \\ &\leq \varepsilon/\gamma. \square \end{aligned}$$

Theorem 3.1 has the following interpretation. The number γ measures the linear independence of the columns of AW. If it is small compared to $\|AW\|$ then the columns of AW themselves must be nearly dependent. Thus Theorem 3.1 says that <u>if we can isolate a set of</u> r <u>columns of</u> A <u>that are strongly independent, then the space spanned by them must be a good approximation to the</u> ε-<u>section</u> $R(U_\varepsilon)$.

However, there are limits to how far we can go with this process. By (2.8) the number γ satisfies $\sigma_r \geq \gamma$, and by the definition of numerical rank $\varepsilon \geq \sigma_{r+1}$. Consequently, the best ratio we can obtain in (3.3) is σ_{r+1}/σ_r. Thus the theorem is not very meaningful unless there is a well defined gap between σ_{r+1} and σ_r. One cure for this problem is to increase ε in an attempt to find a gap; however, such a gap need not exist (e.g. suppose $\sigma_{i+1} = \sigma_i/2$ $(i = 1,2,\ldots,n-1)$). What to do when the matrix A exhibits a gradual rather than a precipitous decline into degeneracy is a difficult problem, whose solution must almost certainly depend on additional information.

A second difficulty is that it may be impossible to obtain the ideal ratio because in practice we must restrict our choice of W to columns of the identity matrix; i.e. we must choose from among columns of A. That this is a real possibility is shown by the following example.

Example 3.2. Let $e^{(n)}$ denote the vector $(1,1,\ldots,1)^T$ with n components. The matrix

$$A_n = I_n - \frac{e^{(n)}e^{(n)T}}{n}$$

has singular values $1,1,\ldots,1,0$, so that it has numerical rank $(1,0,n-1)_2$. Thus we should like to remove a single column of A_n to obtain an approximation to the 0-section of A. Owing to symmetry, it does not matter which column we remove. If we remove the last one, the resulting matrix A'_n has the form

$$A'_n = E_n - \frac{e^{(n)}e^{(n-1)T}}{n}$$

where E_n consists of the first $n-1$ columns of the identity matrix. Thus

$$A'_n \frac{e^{(n-1)}}{\sqrt{n-1}} = \frac{1}{\sqrt{n-1}}\left[\begin{pmatrix} e^{(n-1)} \\ 0 \end{pmatrix} - \frac{n-1}{n}\begin{pmatrix} e^{(n-1)} \\ 1 \end{pmatrix}\right]$$

$$= \frac{1}{n\sqrt{n-1}}\begin{pmatrix} e^{(n-1)} \\ n-1 \end{pmatrix},$$

from which it follows that

$$\left\| A_n' \frac{e^{(n-1)}}{\sqrt{n-1}} \right\|_2 = \frac{1}{\sqrt{n}}$$

and

$$\gamma = \inf(A_n') \leq \frac{1}{\sqrt{n}} .$$

It should be observed that the factor $n^{-1/2}$ exhibited in the example is not extremely small. For $n = 25$ it is only 1/5. Unfortunately no lower bound on γ is known, although with the computational algorithms to be described in Sections 5 and 6 it is easy enough to check the computed value.

A final problem associated with Theorem 3.1 is that it is not invariant under scaling. By scaling we mean the multiplicative scaling of rows and columns of A and not additive scaling such as the subtraction of means or a time factor from the columns of A (this latter scaling can be handled by including the factors explicitly in the model). Since by multiplying a column by a sufficiently small constant one can produce as small a singular value as one desires without essentially altering the model, Theorem 3.1 can be coaxed into detecting degeneracies that are not really there. This means that one must look outside the hypotheses of Theorem 3.1 for a natural scaling. While we are suspicious of pat scaling strategies, we think that the following criterion is reasonable for many applications. Specifically, the rows and columns of A should be scaled so that the errors in the individual elements of A are

as nearly as possible equal. This scaling has also been proposed in [4], and an efficient algorithm for accomplishing it is described in [5].

The rationale for this scaling is the following. From the definition of the singular value decomposition it follows that

$$Av_i = \sigma_i u_i \quad (i = 1,2,\ldots,n).$$

Now if we imagine that our matrix is in error and that our true matrix is A + E, then

$$(A+E)v_i = \sigma_i u_i + Ev_i. \tag{3.6}$$

If we have balanced our matrix as suggested above, then all of the elements of E are roughly the same size, and $\|Ev_i\|_2 \cong \|E\|_2$. Thus if $|\sigma_i| \leq \|E\|_2$, equation (3.6) says that up to error v_i is a null vector of A + E, and the matrix is degenerate.

We recognize that this scaling criterion raises as many questions as it answers. An important one is what to do when such scaling cannot be achieved. Another question is raised by the observation that in regression row scaling is equivalent to weighting observations, which amounts to changing the model.* Is this justified simply to make Theorem 3.1 meaningful? Although this question has no easy answer, we should like to point out that it may be appropriate to use one scaling to eliminate colinearities in A and another for subsequent regressions.

*We are indebted to John Chambers and Roy Welsh for pointing this out.

In the next section we are going to examine the implications of Theorem 3.1 for the linear least squares problem in which a vector of observations b is optimally approximated in the 2-norm by linear combinations of the columns of A:

$$b \cong Ax.$$

In some applications the 2-norm is not the best possible choice, and one may wish to minimize $\varphi(b-Ax)$, where φ is a function that may not even be a norm. For example, in robust regression one approach is to minimize a function that may reduce the influence of wild points. We shall not pursue this subject here; but we believe that Theorem 3.1 has important implications for these problems. Namely, if we are searching for an approximation to b in $R(A)$, we cannot expect the solution to be well determined unless $R(A)$ itself is. Theorem 3.1 provides a theoretical basis for finding stable subspaces of $R(A)$; however, specific theorems must wait the development of a good perturbation theory for approximation in norms other than the 2-norm.

4. The Linear Least Squares Problem

In this section we shall consider the linear least squares problem

$$\text{minimize } \|b-Ax\|_2^2. \tag{4.1}$$

It is well known that this problem always has a solution, which is unique if and only if A is of full column rank. At the solution, the residual vector

$$r = b - Ax$$

is the projection of b onto the orthogonal complement of $R(A)$.

When A has numerical rank $(\delta,\varepsilon,r)_2$, the solution to (4.1) may be large, and some of the individual components of the solution will certainly have large variances. If the ratio ε/δ is sufficiently small a stable solution can be computed by restricting oneself to the ε-section of A. Computationally this can be done as follows. Define U_ε and $\hat{U}_\varepsilon$ as in Section 3, and further define

$$V_\varepsilon = (v_1,v_2,\ldots,v_r), \quad \hat{V}_\varepsilon = (v_{r+1},\ldots,v_n)$$

and

$$\Sigma_\varepsilon = \operatorname{diag}(\sigma_1,\sigma_2,\ldots,\sigma_r), \ \hat{\Sigma}_\varepsilon = \operatorname{diag}(\sigma_{r+1},\ldots,\sigma_n).$$

Then the matrix B of (2.6) is given by

$$B = U_\varepsilon \Sigma_\varepsilon V_\varepsilon^T.$$

Moreover the vector

$$x_\varepsilon = V_\varepsilon \Sigma_\varepsilon^{-1} U_\varepsilon^T b$$

is the unique solution of the problem of minimizing

$$\|b-Bx\|_2$$

that is of minimum 2-norm. It is easily seen that

$$r_\varepsilon = b - Ax_\varepsilon = b - Bx_\varepsilon.$$

As we indicated in the last section, this solution is not entirely satisfactory, since it involves all the columns of A, whereas we might hope to obtain a satisfactory representation of b in terms of r suitably chosen columns; that is with a model having only r carriers. It is a consequence of Theorem 3.1 that any set of r reasonably independent columns will do, although in practice additional considerations may make some choices preferable to others.

Theorem 4.1. Assuming the notation and hypothesis of Theorem 3.1, let x_ε and r_ε be defined as above. Let y_W be the solution of the linear least squares problem

$$\text{minimize } \|b-AWy\|_2^2$$

and let r_W be the residual

$$r_W = b - AWy_W.$$

Then

$$\frac{\|r_\varepsilon - r_W\|_2}{\|b\|_2} \leq \varepsilon/\gamma.$$

Proof. By the properties of the least squares residual $r_\varepsilon = (I-P_{U_\varepsilon})b$ and $r_W = (I-P_{AW})b$. Hence

$$\|r_\varepsilon - r_W\|_2 = \|(P_{U_\varepsilon} - P_{AW})b\|_2 \leq \frac{\varepsilon}{\gamma}\|b\|_2. \square$$

Theorem 4.1 partially answers a question raised by Hotelling [10]; namely if carriers are chosen to eliminate dependencies, what guarantees

that one such set will not fit b better than another? The answer is that if there is a well defined gap between δ and ε, then any set of r strongly independent columns will give approximately the same residual. However, there remains the possibility that by including more columns of A a considerably smaller residual could be obtained. We stress that such a solution cannot be very stable. By (2.8) any matrix consisting of more than r columns of A must have a singular value less than or equal to ε, and it follows from the perturbation theory for the least squares problem [15] that the solution must be sensitive to perturbations in A and b. (Another way of seeing this is to note that ε^{-2} is a lower bound for $\|(A^TA)^{-1}\|_2$, so that the solution must have a large covariance matrix.)

However, one might be willing to put up with the instabilities in the solution provided it gives a good approximation to b. We shall now show that any solution that substantially reduces the residual over r_ε is not only unstable, it is also large.

Theorem 4.2. Let r_ε be defined as above. Given the vector x, let $r = b - Ax$. If $\|r_\varepsilon\|_2 > \|r\|_2$, then

$$\|x\|_2 \geq \frac{\|r_\varepsilon\|_2 - \|r\|_2}{\varepsilon} .$$

Proof. Let $z = V^Tx$ and let

$$U^Tb = \begin{pmatrix} c \\ d \end{pmatrix}$$

where c is an n-vector. Then if we partition $z = (z_\varepsilon^T, \hat{z}_\varepsilon^T)^T$ and $c = (c_\varepsilon^T, \hat{c}_\varepsilon^T)^T$ conformally with the previous partitions of U, V, and Σ, we have

$$\begin{aligned}
\|r\|_2^2 &= \|U^T(b-AVV^Tx)\|_2^2 \\
&= \left\| \begin{pmatrix} c \\ d \end{pmatrix} - \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} z \right\|_2^2 \\
&= \|c-\Sigma z\|_2^2 + \|d\|_2^2 \\
&= \|c_\varepsilon - \Sigma_\varepsilon z_\varepsilon\|_2^2 + \|\hat{c}_\varepsilon - \Sigma_\varepsilon \hat{z}_\varepsilon\|_2^2 + \|d\|_2^2.
\end{aligned}$$

Consequently

$$\|r\|_2^2 \geq \|\hat{c}_\varepsilon - \hat{\Sigma}_\varepsilon \hat{z}_\varepsilon\|_2^2 + \|d\|_2^2. \tag{4.2}$$

Now the vector $y_\varepsilon = V^T x_\varepsilon$ is given by

$$y_\varepsilon = \begin{pmatrix} \Sigma_\varepsilon^{-1} c_\varepsilon \\ 0 \end{pmatrix}$$

so that

$$\|r_\varepsilon\|_2^2 = \|\hat{c}_\varepsilon\|_2^2 + \|d\|_2^2. \tag{4.3}$$

From (4.2)

$$\begin{aligned}
\sqrt{\|r\|_2^2 - \|d\|_2^2} &\geq \|\hat{c}_\varepsilon - \hat{\Sigma}_\varepsilon \hat{z}_\varepsilon\|_2 \geq \|\hat{c}_\varepsilon\|_2 - \|\hat{\Sigma}_\varepsilon\|_2 \|\hat{z}_\varepsilon\|_2 \\
&\geq \|\hat{c}_\varepsilon\|_2 - \varepsilon \|\hat{z}_\varepsilon\|_2.
\end{aligned}$$

Hence

$$\|x\|_2 \geq \|\hat{z}_\varepsilon\|_2 \geq \frac{\|c_\varepsilon\|_2 - \sqrt{\|r\|_2^2 - \|d\|_2^2}}{\varepsilon} ,$$

and from (4.3)

$$\|x\|_2 \geq \frac{\sqrt{\|r_\varepsilon\|_2^2 - \|d\|_2^2} - \sqrt{\|r\|_2^2 - \|d\|_2^2}}{\varepsilon}$$

$$\geq \frac{\|r_\varepsilon\|_2 - \|r\|_2}{\varepsilon} . \square$$

The theorem shows that even a slight decrease in the residual must result in a great increase in the size of the solution. It is hardly necessary to add that a large solution is seldom acceptable in practice: it must have high variance, and it may be physically meaningless.

The results of this section have implications for a common practice in data analysis, namely that of fitting a large number of subsets of the columns of A in an attempt to obtain a good fit with fewer than the full complement of columns (for example, see [6]). We have, in effect, shown that if the ratio ε/δ is reasonable, this procedure is not likely to be very productive. Any set of r independent columns will give about the same residual, and any larger set that significantly reduces the residual must produce an unacceptably large solution. There are, however, two cases where this procedure might be of some help. First when it is hoped that fewer than r columns can produce a good fit, and second when the ε-δ ratio is not very small. An approach to the second problem that uses the singular value decomposition of the augmented matrix (A,b) is described in [9] and [16,17].

5. Extraction of Independent Columns: the QR Factorization

We now turn to the problem of extracting a set of numerically independent columns. The first method we shall consider is based on the QR factorization of the matrix A. Specifically, if A is an $m \times n$ matrix with $m \geq n$, then A can be written in the form

$$A = QR,$$

where Q has orthonormal columns ($Q^TQ=I$) and R is upper triangular. If A has full column rank, then the factorization is unique up to the signs of the columns of Q and the corresponding rows of R. It should be noted that the columns of Q form an orthonormal basis for $R(A)$.

A knowledge of the QR factorization of A enables one to solve the least squares problem (4.1). Specifically, any solution x of (4.1) must satisfy the equation

$$Rx = Q^Tb,$$

which can be easily solved since R is upper triangular. Moreover, since $A^TA = R^TR$, we have

$$(A^TA)^{-1} = R^{-1}R^{-T}$$

so that one can use the matrix R in the factorization to estimate the covariance matrix of the solution.

An especially desirable feature of the QR factorization is that it can be used to solve a truncated least squares problem in which only an

initial set of columns are fit. If $A^{|r}$ denotes the matrix consisting of the first r columns of A and $R^{\overline{|r}}$ denotes the leading principal submatrix of order r of R then

$$A^{|r} = Q^{|r}R^{\overline{|r}}. \tag{5.1}$$

Since $R^{\overline{|r}}$ is upper triangular and $Q^{|r}$ has orthonormal columns, equation (5.1) gives the QR factorization of $A^{|r}$ and can be used as described above to solve least squares problems involving $A^{|r}$.

The basis for using the QR factorization to extract a linearly independent set of columns from the matrix A is contained in the following theorem.

Theorem 5.1. Let the QR factorization of A be partitioned in the form

$$(A_1,A_2) = (Q_1,Q_2)\begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}$$

where $A_1,Q_1 \in \mathbb{R}^{m\times r}$ and $R_{11} \in \mathbb{R}^{r\times r}$. If

$$\|R_{22}\|_2 = \varepsilon < \delta = \inf(R_{11}),$$

then A has rank $(\delta,\varepsilon,r)_2$. Moreover,

$$\inf(A_1\ ) = \delta.$$

Proof. Because the columns of Q are orthonormal, the singular values of A and of R are the same. Now δ is the r-th singular value of R_{11}, and hence by (2.9) δ is less than or equal to the r-th

singular value of A; i.e. $\sigma_r \geq \delta$. Likewise from (2.10), $\varepsilon \geq \sigma_{r+1}$. Thus A has rank (δ, ε, r). Moreover, since Q_1 has orthonormal columns,

$$\inf(A_1) = \inf(Q_1 R_{11}) = \inf(R_{11}) = \delta. \square$$

The application of this theorem is obvious. If, after having computed the QR factorization of A, we encounter a small matrix R_{22} and a matrix R_{11} with a suitably large infinum, then the columns of A_1 span a good approximation to the ε-section of A. Because of (5.1), we have at hand the QR factorization of A_1 and can proceed immediately to the solution of least squares problems involving A_1. There remain two problems. First how can one insure that the first r columns of A are linearly independent, and second how can one estimate $\inf(R_{11})$?

The solution to the first problem depends on the method by which the QR factorization is computed. Probably the best numerical algorithm is one based on Householder transformations in which the QR factorizations $A^{|k} = Q^{|k} R^{\overline{|k}}$ are computed successively for $k = 1,2,\ldots,n$ (e.g. see [14]). At the k-th step, just before $Q^{|k}$ and $R^{\overline{|k}}$ are computed, there is the possibility of replacing the k-th column of A by one of the columns $a_{k+1}, a_{k+2}, \ldots, a_n$. If the column that maximizes the (k,k) - element of R is chosen to replace a_k, then there will be a tendency for independent columns to be processed first, leaving the dependent columns at the end of the matrix. An ALGOL program incorporating this "column pivoting" is given in [3] and a FORTRAN program is given in [11].

Once a satisfactory QR decomposition has been calculated, we can estimate $\|R_{22}\|_2$ by the bound

$$\|R_{22}\|_2 \leq \sqrt{\|R_{22}\|_1 \|R_{22}\|_\infty},$$

where

$$\|X\|_1 = \max_j \sum_i |x_{ij}|$$

and

$$\|X\|_\infty = \max_i \sum_j |x_{ij}|.$$

Likewise one can estimate $\inf(R_{11})$ by computing R_{11}^{-1} (an easy task since R_{11} is upper triangular) and using the relations

$$\inf(R_{11}) = \|R_{11}^{-1}\|_2^{-1} \geq \frac{1}{\sqrt{\|R_{11}^{-1}\|_1 \|R_{11}^{-1}\|_\infty}}.$$

The procedure sketched above is completely reliable in the sense that it cannot fool one into thinking a set of dependent columns are independent. However, it can fail to obtain a set of linearly independent columns, as the following example shows.

Example 5.2. Let A_n be the matrix of order n illustrated below for $n = 5$:

$$A_5 = \begin{pmatrix} 1 & -1/\sqrt{2} & -1/\sqrt{3} & -1/\sqrt{4} & -1/\sqrt{5} \\ 0 & 1/\sqrt{2} & -1/\sqrt{3} & -1/\sqrt{4} & -1/\sqrt{5} \\ 0 & 0 & 1/\sqrt{3} & -1/\sqrt{4} & -1/\sqrt{5} \\ 0 & 0 & 0 & 1/\sqrt{4} & -1/\sqrt{5} \\ 0 & 0 & 0 & 0 & 1/\sqrt{5} \end{pmatrix}$$

Letting $x_n^T = (1, \sqrt{2}/2, \sqrt{3}/4, \sqrt{4}/8, \ldots, \sqrt{n}/2^{n-1})$, it is easily verified that

$$A_n x_n = 2^{-n} e$$

where $e^T = (1,1,\ldots,1)$. Thus A_n has the approximate null vector x_n and must have nearly dependent columns. However, computing the QR factorization of A_n, even with column pivoting, leaves A_n undisturbed. Since no element of A_n is very small, we shall have R_{22} void; i.e. no dependent column will be found.

It should be observed that in the above example there is no danger of the degeneracy in A_n going undetected. Since R_{22} is void, $R_{11} = A_n$ and any attempt to estimate $\inf(R_{11})$ will reveal the degeneracy.

It may be objected that the matrix A_n in Example 5.2 shows an obvious sign of degeneracy; viz. its determinant $(n!)^{-1/2}$ goes rapidly to zero with increasing n. However, the matrix $|A_n|$, obtained from A_n by taking the absolute value of its elements, has the same determinant yet its columns are strongly independent. Thus the example confirms a fact well known to practical computers: the value of a determinant is worthless as an indication of singularity.

6. Extraction of Independent Columns: the Singular Value Decomposition

When the singular value decomposition of A has been computed (an ALGOL program is given in [8] and a FORTRAN program in [11]), a different way of selecting independent columns is available. The method is based on the following theorem.

Theorem 6.1. Let A have the singular value decomposition

$$U^T A V = \begin{pmatrix} \Sigma \\ 0 \end{pmatrix}.$$

Let V be partitioned in the form

$$V = \begin{pmatrix} V_{\varepsilon 1} & \hat{V}_{\varepsilon 1} \\ V_{\varepsilon 2} & \hat{V}_{\varepsilon 2} \end{pmatrix},$$

where $V_{\varepsilon 1}$ is $r \times r$, and let A be partitioned in the form

$$A = (A_1, A_2),$$

where A_1 has r columns. Let $\delta = \sigma_r$, $\varepsilon = \sigma_{r+1}$ and

$$\gamma = \delta \inf(V_{\varepsilon 1}).$$

Then A has numerical rank $(\delta, \varepsilon, r)_2$ and

$$\inf(A_1) \geq \gamma. \tag{6.1}$$

Proof. The fact that A has numerical rank $(\delta, \varepsilon, r)_2$ follows immediately from Theorem 2.3. To establish (6.1), observe that if we write

$$AV = S = (S_1, S_2)$$

where S_1 has r columns, then $S_1^T S_2 = 0$. Now since $A = SV^T$, we have

$$A_1 = S_1 V_{\varepsilon 1}^T + S_2 \hat{V}_{\varepsilon 1}^T.$$

Since $S_1^T S_2 = 0$,

$$\inf(A_1) \geq \inf(S_1 V_{\varepsilon 1}^T) \geq \inf(S_1)\inf(V_{\varepsilon 1}^T)$$

$$= \sigma_r \inf(V_{\varepsilon 1}) = \gamma. \square$$

As with the QR factorization, Theorem 6.1 provides us with a way of determining when an initial set of r columns of A are independent. Since an initial set may be degenerate, we must adopt some kind of interchange strategy to bring an independent set of columns into the initial positions. If P is any permutation matrix, then

$$U^T(AP)(P^T V) = \begin{pmatrix} \Sigma \\ 0 \end{pmatrix},$$

so that in the singular value decomposition an interchange of columns of A corresponds to an interchange of the corresponding rows of V. This suggests that we exchange rows of V until $\inf(V_{\varepsilon 1})$ becomes acceptably large. One way of accomplishing this is to start with the $r \times n$ matrix

$$V_1^T = (V_{\varepsilon 1}^T, V_{\varepsilon 2}^T)$$

and compute its QR factorization with column pivoting to force a set of independent columns into the first r positions. Alternatively one

could apply an algorithm such as Gaussian elimination with complete pivoting to V_1^T (e.g. see [14]).

If either of the above suggestions is followed, the final matrix $V_{\varepsilon 1}^T$ will be upper triangular, and its infimum can be bounded by the method suggested in the last section.

If r is small, significant savings can be obtained by observing that the singular values in $[0,1)$ of $V_{\varepsilon 1}$ and $\hat{V}_{\varepsilon 2}$ are the same (see the appendix of [15] for a proof). Thus one can start with the smaller matrix

$$V_2^T = (\hat{V}_{\varepsilon 1}^T, \hat{V}_{\varepsilon 2}^T) \tag{6.2}$$

and use the QR factorization with column pivoting to determine the dependent columns of A. Note that when $r = n-1$ the column to be stricken corresponds to the largest element of the row vector V_2^T.

The question of whether to use the QR factorization or the singular value decomposition is primarily one of computational efficiency. Although Example 5.2 shows that the QR factorization can fail to isolate a set of independent columns in a case where the singular value decomposition does, this is an unusual phenomenon (see Example 7.2) and in most cases the QR factorization with column pivoting is effective in locating independent columns. When m is not too much greater than n, the calculation of the singular value decomposition is considerably more expensive than the calculation of the QR factorization, and it is more efficient to stick with the latter, if possible.

When $m \gg n$, we can begin by computing the QR factorization of A. The matrix R has the same singular values as A, and indeed if

$$\tilde{U}^T R V = \Sigma \tag{6.3}$$

is the singular value decomposition of R, then V is the matrix of right singular vectors of A. Since R is an $n \times n$ matrix, the reduction (6.3) is computationally far less expensive than the initial computation of R, and there seems to be no reason not to use the singular value decomposition.

7. Examples

In this section we shall give some examples illustrating the preceding material. The numerical computations were done in double precision on an IBM 360; i.e. to about sixteen decimal digits.

Example 7.1. This example has been deliberately chosen to be uncomplicated. For fixed n, let

$$H_n = I - \frac{2}{n} ee^T,$$

where $e^T = (1,1,\ldots,1)$. It is easily verified that H_n is orthogonal. Let

$$\Sigma = \mathrm{diag}(1,1,1,1,1,0,0,0,0,0)$$

and

$$A = H_{50} \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} H_{10}.$$

Then A has five nonzero singular values equal to unity and five zero singular values, and thus it should have five linearly independent columns.

The singular values of A were computed to be $1,1,1,1,1,.35\times10^{-16}$, $0,0,0,0$, so that A can be regarded as having rank $(1,\varepsilon,5)$ where $\varepsilon = 10^{-16}$. The pivoting strategy described in Section 6 was used to isolate a set of five linearly independent columns. These turned out to be columns 1,2,4,5, and 9. The associated matrix $V_{\varepsilon 1}$ had an infimum of .45 which is very close to the optimal value of unity. As a final check, we compute $\|P_{U_\varepsilon} - P_{AW}\|$, where $W = (e_1,e_2,e_4,e_5,e_9)$ is the matrix that selects the independent columns from A (cf. Theorem 3.1). The result is

$$\|P_{U_\varepsilon} - P_{AW}\|_2 = .37 \times 10^{-14},$$

which shows that columns 1,2,4,5, and 9 of the matrix A almost exactly span the ε-section of A.

The QR factorization with column pivoting that is described in Section 5 was also applied to A. The pivot columns and their norms were

5	.89
4	.86
2	.81
3	.71
6	.44
1	$.45 \times 10^{-16}$
7	$.13 \times 10^{-30}$
8	0
9	0
10	0

If the gap is taken to lie after the fifth vector we have

$$\inf(R_{11}) = 1, \quad \|R_{22}\|_2 = .20 \times 10^{-16}.$$

Thus the QR factorization exhibits the same sharp gap as the singular value decomposition. However, the five columns 2,3,4,5, and 6 designated as independent are different from those chosen by means of the singular value decomposition. Nonetheless, for $W = (e_2,e_3,e_4,e_5,e_6)$ we have

$$\|P_{U_\varepsilon} - P_{AW}\| = .37 \times 10^{-14},$$

so that this choice of columns is as good as the one predicted by the singular value decomposition.

Incidentally the estimate of $\|R_{22}\|_2$ using the 1- and ∞-norms is

$$\sqrt{\|R_{22}\|_1 \|R_{22}\|_\infty} = .94 \times 10^{-16},$$

which is not a gross overestimate.

Example 7.2. This is the matrix A_{25} of Example 5.2. The singular values of this matrix are

$$\sigma_1=3.7,\sigma_2=1.6,\ldots,\sigma_{24}=.31,\sigma_{25}=.77 \times 10^{-7}.$$

Again there is a well defined gap, and we may take A to have rank $(.31,\varepsilon,24)$ where $\varepsilon = 10^{-7}$. This time there is only a single dependent vector which can be found by looking for the largest component of the right singular vector v_{25} corresponding to σ_{25} (cf. the comments at equation (6.2)). This component, .75, is the first, which indicates

that column one should be discarded. For this selection we have

$$\|P_{U_\varepsilon} - P_{AW}\|_2 = .49 \times 10^{-7}.$$

In principle, the QR factorization should fail to isolate a dependent column of A_{25}. However, because the elements of A_{25} were entered with rounding error, the pivot order with column norms turned out to be

1	1.0
25	.98
6	.88
.	.
.	.
.	.
24	.37
2	$.15 \times 10^{-16}$

This again gives a well defined gap and indicates that column 2 should be thrown out (the second component of v_{25} is .53 so that also from the point of view of the singular value decomposition the second column is a candidate for rejection). For this subspace we have

$$\|P_{U_\varepsilon} - P_{AW}\| = .11 \times 10^{-6}.$$

Thus the QR factorization gives only slightly worse results than the singular value decomposition, in spite of the fact that the example was concocted to make the QR decomposition fail.

Example 7.3. To show that our theory may be of some use even where there is not a sharply defined gap in the singular values, we consider the Longley test data [12], which has frequently been cited in the literature.

Since it is a common practice to subtract means from raw data, we have included a column of ones in the model. Specifically the columns of A are as follows:

1 -- ones

2 -- GNP Implicit Price Deflator, 1954 - 100

3 -- GNP

4 -- Unemployment

5 -- Size of armed forces

6 -- Noninstitutional population ≥ 14 years old

7 -- Time (years)

The scaling of this data will critically affect our results. For the purposes of this experiment we assume that columns two through six are known to about three significant figures. Accordingly each of these columns was multiplied by a factor that made its mean equal to 500. The column of ones is known exactly and by the equal error scaling criterion ought to be scaled by a factor of infinity. As an approximation we took the scaling factor to be 10^{10}.

The column of years can be treated in two ways. First the errors in the time of measurement can be attributed to the column itself, which would result in the column being assigned a low accuracy. However, we observe that any constant bias in the time of measurement is accounted for by the column of ones, and any other errors can be attributed to the measured data. Consequently we have preferred to regard the years as known exactly and scale the seventh column by 10^{10}.

The singular values of the matrix thus scaled are

$$.78 \times 10^{14}$$
$$.94 \times 10^{8}$$
$$.58 \times 10^{3}$$
$$.26 \times 10^{3}$$
$$.26 \times 10^{2}$$
$$.22 \times 10^{2}$$
$$.51 \times 10^{1}$$

Since the error in A is of order unity, the last singular value must be regarded as pure noise, and we may take A to have rank $(22,5.1,6)_2$. The largest component of the seventh singular vector is the sixth and has a value of .90. When the sixth column is removed from the matrix, the resulting subspace compares with $U_{5.1}$ as follows:

$$\|P_{U_{5.1}} - P_{AW}\|_2 = .12.$$

The relatively poor determination of the 5.1-section of A suggests that not much useful information can be obtained from a least squares fit, even when the sixth column is ignored. The next gap that presents itself is between the fourth and fifth singular values. If we regard A as having rank $(260,26,4)_2$ and use the pivoting strategy of Section 6 to isolate a set of four independent columns, we choose columns 1,4,5, and 7 with

$$\inf(V_{\epsilon 1}) = .991.$$

For this choice of columns

$$\|P_{U_{260}} - P_{AW}\| = 0.011,$$

a far more satisfactory result.

If the QR factorization is applied to A, there results the following sequence of pivot columns and norms:

7	$.78 \times 10^{14}$
1	$.94 \times 10^{8}$
5	$.47 \times 10^{3}$
4	$.31 \times 10^{3}$
2	$.24 \times 10^{2}$
3	$.21 \times 10^{2}$
6	$.57 \times 10^{1}$

This agrees completely with the results from the singular value decomposition. Either one or three columns should be discarded, and columns 6, 2, and 3, in that order, are candidates.

Although these results indicate that columns 2, 3, and 6 should be discarded from the model, they are not conclusive, since there may be other sets containing some of these columns that give a satisfactory approximation to the 260-section of A. However, a singular value decomposition of the matrix consisting of columns 1,2,3,6, and 7 gives the singular values

$$.78 \times 10^{14}$$
$$.94 \times 10^{8}$$
$$.50 \times 10^{2}$$
$$.25 \times 10^{2}$$
$$.10 \times 10^{2}$$

which shows that none of these columns is a really good candidate for inclusion in the model.

To sum up: if the raw Longley data is taken to be accurate to three significant figures, if years are assumed to be exact, and if means are subtracted from the columns, then the column corresponding to noninstitutional population is redundant, and the columns corresponding to the GNP implicit price deflator and the GNP are so nearly redundant that their inclusion in the model will affect the stability of the residuals from any regressions.

References

1. S. N. Afriat, Orthogonal and oblique projectors and the characteristics of pairs of vector spaces, Proc. Cambridge Philos. Soc. 53 (1957) 800-816.

2. Å Bjork and G. H. Golub, Numerical methods for computing angles between linear subspaces, Math. Comp. 27 (1973) 579-594.

3. P. Businger and G. H. Golub, Linear least squares solutions by Householder transformations, Numer. Math. 1 (1965) 269-276.

4. J. M. Chambers, Stabilizing linear regression against observational error in independent variates, unpublished manuscript, Bell Laboratories, Murray Hill, New Jersey (1972).

5. A. R. Curtis and J. K. Reid, On the automatic scaling of matrices for Gaussian elimination, J. Inst. Math. Appl. 10 (1972) 118-124.

6. C. Daniel and F. S. Wood, Fitting Equations to Data, Wiley, New York (1971).

7. G. H. Golub, Least squares, singular values, and matrix approximations, Aplikace Mathematiky 13 (1968) 44-51.

8. _____ and C. Reinsch, Singular value decomposition and least squares solution, Numer. Math. 14 (1970) 403-420.

9. D. M. Hawkins, On the investigation of alternative regressions by principal component analysis, Appl. Statist. 22 (1973) 275-286.

10. H. Hotelling, The relations of the newer multivariate statistical methods to factor analysis, Brit. J. Statist. Psychol. 10 (1957) 69-79.

11. C. L. Lawson and R. J. Hanson, Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, New Jersey (1974).

12. J. W. Longley, An appraisal of least squares programs for the electronic computer from the point of view of the user, J. Amer. Statist. Assoc. 62 (1967) 819-841.

13. G. W. Stewart, Error and perturbation bounds for subspaces associated with certain eigenvalue problems, SIAM Rev. 15 (1973) 727-764.

14. _____, Introduction to Matrix Computations, Academic Press, New York (1973).

15. _____, On the perturbation of pseudo-inverses, projections and linear least squares problems, to appear SIAM Rev.

16. J. T. Webster, R. F. Gunst, and R. L. Mason, Latent root regression analysis, Technometrics 16 (1974) 513-522.

17. _____, A comparison of least squares and latent root regression estimators, Technometrics 18 (1976) 75-83.