
On the Reliable and Flexible Solution of Practical Subset Regression Problems

M. Verhaegen

(NASA-TM-89442) ON THE RELIABLE AND
FLEXIBLE SOLUTION OF PRACTICAL SUBSET
REGRESSION PROBLEMS (NASA) 7 p Avail:
NHS HC A02/MF A01 CSCI 12B

N87-24156

H1/66 Unclass
0080146

March 1987

On the Reliable and Flexible Solution of Practical Subset Regression Problems

M. Verhaegen, Ames Research Center, Moffett Field, California

March 1987



National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

ON THE RELIABLE AND FLEXIBLE SOLUTION OF PRACTICAL SUBSET REGRESSION PROBLEMS

M. Verhaegen*

Aerospace Engineer

Ames Research Center, Moffett Field, California

ABSTRACT

A new algorithm for solving subset regression problems is described. The algorithm performs a QR decomposition with a new column-pivoting strategy, which permits subset selection directly from the originally defined regression parameters. This, in combination with a number of extensions of the new technique, makes the method a very flexible tool for analyzing subset regression problems in which the parameters have a physical meaning.

INTRODUCTION

The subset regression problem analyzed in this paper is formulated for the following least-squares problem:

$$\min_{\underline{x}} \|\underline{Ax} - \underline{b}\|_2 \quad (1)$$

The system matrix $A \in R^{m \times n}$ ($m \geq n$) is assumed to contain a number of columns that are "nearly" linearly dependent on an independent set of columns of A . The task now is to find the minimal number of columns comprising that independent set. Furthermore, the adjective "nearly" is used to indicate that small perturbations of A , that is, of the order of the inaccuracies ($O(\epsilon)$) on the entries of A , establish that linear dependency.

A numerically reliable technique used so far to detect near dependencies is that of singular-value decomposition (SVD) [1]. This technique determines a minimal set of columns. Say that the dimension of this set is k . But, each of these selected columns now is a linear combination of the original columns of A . Hence, that minimal set will generally contain more than k individual columns of A . This is a major drawback for practical subset regression problems, such as aerodynamic model identification [2],[3].

This paper describes a new technique that allows the direct selection of k individual columns of A . The technique performs a QR decomposition with a new column-pivoting strategy. Furthermore, it is not necessary to compute the full SVD of A in order to determine k . This is because accurate estimates of the singular values, which allow such a determination, are also obtained.

This new technique is described first, and then a number of extensions and modifications of it are presented that allow its application to practical problems in subset regression analysis. These problems are in the order to be discussed:

1. The determination of the interrelationship between the defined independent and dependent columns of A by the new technique.
2. The inclusion of a priori information about the structure of the regression model, as well as about estimates of the individual components of \underline{x} in (1) and their corresponding uncertainties.
3. The capability to jointly process columns of A that are contaminated by noise and others that are noise-free without using scaling techniques.
4. The efficient solution of the closely related total least-squares problem, described in [4].

NEW COLUMN-PIVOTING STRATEGY IN QR DECOMPOSITION

The QR decomposition of the matrix A in (1) is denoted as

$$Q_o^T A = \begin{bmatrix} R \\ 0 \end{bmatrix} \quad (2)$$

where $R \in R^{n \times n}$ is upper triangular and $Q_o^T Q_o = I_n$. This transformation differs from the SVD in that no right orthogonal transformation on A is applied. It is precisely this right-transformation that obscures the selection of k individual columns of A . To avoid this problem, the only right-transformation on A allowed is a column permutation.

*National Research Council Research Associate

The idea of column pivoting is to exploit the freedom introduced by this column permutation to solve the rank-deficient least-squares problem (1). Based on the new pivoting strategy, it will be shown that this can be done as reliably as with the SVD method without relying on complete SVD.

Let us clarify this for the rank-one deficient case, that is, for $k = n - 1$, implying that $\sigma_n = O(\epsilon)$ and $\sigma_{n-1} \gg \sigma_n$. (For the sake of brevity, we restrict without loss of generality to this case throughout the whole paper.) Here the singular values σ_i of A are ordered in decreasing magnitude.

From [5] (for example) it is known that the magnitude of the last diagonal element of R in (2) is an upper bound for σ_n . Therefore, the rank-one deficiency can be revealed with column pivoting if we can find a column permutation matrix π_n such that

$$Q_n^T R \pi_n = \begin{bmatrix} R_{11}^n & r_{12}^n \\ 0 & r_{22}^n \end{bmatrix} \quad (3)$$

where $|r_{22}^n|$ is "small" of $O(\epsilon)$. The existence of such a permutation is indicated by exercise P-6-4-4 of [6].

Furthermore this exercise gives a constructive way to find this permutation. The key information is the computation of the right singular vector corresponding to the smallest singular value σ_n . This can be done, for example, by the inverse iteration method [6].

The decomposition (3) combined with (2) produces:

$$Q_n^T \cdot Q_0^T A \pi_n = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \\ 0 & 0 \end{bmatrix} \quad (4)$$

and

$$Q_n^T Q_0^T b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \begin{matrix} k \\ n - k \\ m - n \end{matrix} \quad (5)$$

Therefore, if we explicitly accumulate Q_n^T (which can be done efficiently because we are working only in the n -dimensional parameter space as opposed to the sample space) and perform Q_0^T with Householder transformations such as described in [7], the basic solution \hat{x}_b , the residual y , and the estimate \hat{b} of (1) can be computed in a manner that is completely similar to that in [7]. These quantities can now be given as follows:

$$\hat{x}_b = R_{11}^{-1} b_1 \in R^k \quad (6)$$

$$y = (A \pi_{\text{pivot}})_k \hat{x}_b - b = Q_0 \begin{bmatrix} Q_n \begin{bmatrix} 0 \\ b_2 \end{bmatrix} \\ b_3 \end{bmatrix} \quad (7)$$

$$\hat{b} = (A \pi_n)_k \hat{x}_{k-\hat{b}} Q_0 \begin{bmatrix} Q_n \begin{bmatrix} b_1 \\ 0 \end{bmatrix} \\ 0 \end{bmatrix} \quad (8)$$

where $(A \pi_n)_k$ denotes the k selected columns of A .

DETERMINING INTERRELATIONSHIPS IN THE REGRESSION MODEL

The above algorithm rearranges the columns of A as $[a_{c_1} \dots a_{c_k} a_{c_{k+1}} \dots a_{c_n}]$. Here the vectors $[a_{c_1} \dots a_{c_k}]$ correspond to the so-called identifiable or independent components of x , denoted previously as x_b , and $[a_{c_{k+1}} \dots a_{c_n}]$ corresponds to the dependent ones.

For a number of applications it is necessary to specify these dependencies more precisely. In this case, it is necessary to know on which minimal set of vectors from $[a_{c_1} \dots a_{c_k}]$ each vector from $[a_{c_{k+1}} \dots a_{c_n}]$ depends.

The decomposition obtained in (4) reveals the information necessary to answer this question.

First, focus on the following least-squares problem:

$$\min_{\underline{\beta}} \left\| \begin{bmatrix} R_{11} \\ 0 \end{bmatrix} \underline{\beta} - \begin{bmatrix} r_{12} \\ r_{22} \end{bmatrix} \right\|_2 \quad (9)$$

Elements of the solution $\hat{\underline{\beta}} = R_{11}^{-1} r_{12}$ are the coordinates of the projection of a_{c_n} in the space spanned by $[a_{c_1} \dots a_{c_{n-1}}]$. Of course, "large" components in $\hat{\underline{\beta}}$ suggest a strong correlation between the corresponding columns of $[a_{c_1} \dots a_{c_{n-1}}]$ and a_{c_n} . However, it is more convenient to normalize these coordinates. Let us therefore define the following cosines:

$$\cos \theta_i = \frac{\hat{\underline{\beta}}(i) \|r_{12}\|_2}{\|r_{11}\|_2} \quad \text{for } i = 1:k \quad (10)$$

with r_{1i} denoting the i th column of R_{11} in (9). This set of numbers ranges from -1 to $+1$ and, therefore, can be interpreted completely analogously to the commonly used cross-correlation coefficients [5]. Normally, the latter are

retrieved from the covariance matrix of the least-squares estimates of \underline{x} in (1). Since A here is assumed to be rank-deficient, this covariance matrix cannot be computed unless use is made of pseudoinverses.

Based on the following theorem we can however gain an understanding of this covariance matrix because of the inclusion of a dependent column.

Theorem 1:

If an upper triangular matrix $R \in R^{n \times n}$ is given as

$$R = \left[\begin{array}{c|c} R_{11} & r_{12} \\ \hline 0 & r_{22} \end{array} \right]$$

with R_{11} of rank $n-1$ and r_{22} arbitrary, then

$$(R^T R)^{-1} = \left[\begin{array}{c|c} (R_{11}^T R_{11})^{-1} + \alpha \hat{\underline{g}} \hat{\underline{g}}^T & -\alpha \hat{\underline{g}} \\ \hline -\alpha \hat{\underline{g}}^T & \alpha \end{array} \right] \quad (11)$$

with $\hat{\underline{g}} = R_{11}^{-1} r_{12}$ and $\alpha = 1/r_{22}^2$.

Proof: The proof of Theorem 1 could be obtained directly from the definition of the inverse of a matrix.

For the decomposition obtained in (6), Theorem 1 allows us to exactly calculate the influence of any r_{22} on the estimated variances of $\hat{\underline{x}}_b$ (given by the diagonal elements of $(R_{11}^T R_{11})^{-1}$).

When we make the practical observation that "a component of \underline{x} is called not identifiable if its corresponding estimated variance becomes unacceptably large," then another procedure for determining the dependencies more precisely corresponds to determining whether an unacceptable increase in the variances of the components of \underline{x}_b occurs.

Both of the procedures described above supply parameters and insights that have commonly been used to analyze identifiability problems in practical regression problems.

INCLUSION OF A PRIORI INFORMATION

For the physical applications addressed in this paper, information is often available about (1) the structure of the regression model, that is, which of the components of \underline{x} in (1) have to be in \underline{x}_b , and (2) a priori estimates of individual components of \underline{x} with corresponding variance. For the sake of brevity, only the capabilities of the new procedure in dealing with the first a priori information source are presented here.

We can include that information after the decomposition given in (4) has been obtained. Here we would then use the cosines defined in (10) or the revealed influences of the dependent terms on the estimated variances of the independent terms of \underline{x} . This information would allow us to interchange columns of $[a_{c_1} \dots a_{c_k}]$ with columns of $[a_{c_{k+1}} \dots a_{c_n}]$. Algorithmically, the necessary permutation can again be done efficiently, since we can remain working on the compressed matrix structure defined in (4). Furthermore, information could be revealed about the effect of this interchange by computing the change in residual \underline{v} , efficiently computed as given in (7).

JOINTLY PROCESSING NOISE-FREE AND NOISY COLUMNS IN A

In many regression models there is a so-called offset term. This results in a column of A of all ones. The other columns might result from observations and are, therefore, contaminated by errors, often of different magnitudes for each column. A rough partitioning of the A -matrix in (1) is therefore

$$A = \left[\underbrace{A_1}_{\text{Noise-free}} \mid \underbrace{A_2}_{\text{Contaminated by errors}} \right] \quad (12)$$

The described algorithm requires, as the SVD, a threshold $\bar{\sigma}$ to determine the rank of A in (1) [5]. The use of such a single $\bar{\sigma}$ requires column scaling [5], so that the magnitude of the errors on the columns of this scaled matrix is of the same order. The mixing as given in (12) results in a very bad conditioning of this scaled matrix and, therefore, should be avoided. The new algorithm might handle the situation in (12) without scaling when modified in the following stages.

Stage 1:

Process A_1 by the algorithm given in the second section to produce

$$[Q_1^T A_1 \mid Q_1^T A_2] = \begin{array}{cc|c} R_{11} & R_{12} & R_{13} \\ 0 & R_{22} & \\ \hline 0 & 0 & R_{23} \end{array} \quad (13)$$

This investigates the dependencies among the noise-free columns and, therefore, a threshold $\bar{\sigma}_1 = O(\text{machine precision})$ should be used. Before continuing to the second stage, check whether columns of A_2 are dependent on the columns represented in R_{11} in (13). This corresponds to checking whether

$$\|r_{23}(i)\|_2 \leq \sigma_i \sqrt{m} \quad (14)$$

as outlined in [8]. Here $r_{23}(i)$ denoted the i th column of R_{23} in (13) and σ_i the corresponding standard deviation of the errors on that column. This procedure reduces R_{13}, R_{23} in (13) to $\bar{R}_{13}, \bar{R}_{23}$, respectively.

Stage 2:

Apply the algorithm given in the second section to \bar{R}_{23} :

$$Q_2^T \bar{R}_{23} \pi_2 = \begin{bmatrix} \bar{R}_{11} & \bar{R}_{12} \\ 0 & \bar{R}_{22} \end{bmatrix} \quad (15)$$

using a threshold $\bar{\sigma}_2 = 0$ (magnitude of errors on A_2).

Remark: This two-stage procedure might be extended to the following three-stage procedure, where A is partitioned as

$$A = \left[\underbrace{A_1}_{\text{Noise-free}} \mid \underbrace{A_2}_{\text{Mild errors}} \mid \underbrace{A_3}_{\text{Large errors}} \right] \quad (16)$$

EFFICIENT SOLUTION OF TOTAL LEAST-SQUARES PROBLEM

This problem is not addressed here; the reader is referred to [8] for that discussion.

CONCLUDING REMARKS

A new scheme was presented for efficiently solving subset regression problems. The scheme is as reliable as the singular value decomposition technique, but does the subset selection directly from the defined set of regression parameters. Furthermore, a number of interesting extensions and modifications of this new procedure have been pre-

sented that allow the flexible solution of subset regression problems where the regression parameters are of physical interest.

REFERENCES

- [1] G.H. Golub, V. Klema, and G.W. Stewart, "Rank Degeneracy and Least Squares Problems," TR-456, Department of Computer Science, U. of Maryland, College Park, Md., 1976.
- [2] M.H. Verhaegen, "A New Class of Algorithms in Linear System Theory: With Applications to Real-Time Aircraft Model Identification," Ph.D. dissertation, Catholic University, Leuven, Belgium, Nov. 1985.
- [3] V. Klein, J.G. Batterson, and P.C. Murphy, "Determination of Airplane Structure from Flight Data by Using Modified Stepwise Regression," NASA TP-1916, 1981.
- [4] G.H. Golub and C.F. Van Loan, "An Analysis of the Total Least Squares Problem," SIAM Journal on Numerical Analysis, Vol. 11, 1980, pp. 472-479.
- [5] C.L. Lawson and R.J. Hanson, Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, N.J., 1974.
- [6] G.H. Golub and C.F. Van Loan, Matrix Computations, The Johns Hopkins University Press, Baltimore, 1983.
- [7] J.J. Dongarra, C. Moler, J. Buch, and G. Stewart, LINPACK User's Guide, SIAM, Philadelphia, 1979.
- [8] M.H. Verhaegen, "The Minimal Residual QR Decomposition for Reliably Solving Rank Deficient Least Squares Problems," submitted to International Symposium on Mathematical Theory of Networks and Systems, Phoenix, June 1987.



Report Documentation Page

1. Report No. NASA TM 89442	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle On the Reliable and Flexible Solution of Practical Subset Regression Problems		5. Report Date March 1987	6. Performing Organization Code
		7. Author(s) M. Verhaegen	8. Performing Organization Report No. A87158
9. Performing Organization Name and Address Ames Research Center Moffett Field, CA 94035		10. Work Unit No. 505-66-11	11. Contract or Grant No.
		12. Sponsoring Agency Name and Address National Aeronautics and Space Administration Washington, DC, 20546	
13. Type of Report and Period Covered Technical Memorandum		14. Sponsoring Agency Code	
15. Supplementary Notes Point of contact: M. Verhaegen, Ames Research Center, MS 210-9, Moffett Field, California 94035 (415) 694-5983 or FTS 446-5983			
16. Abstract <p>A new algorithm for solving subset regression problems is described. The algorithm performs a QR decomposition with a new column-pivoting strategy, which permits subset selection directly from the originally defined regression parameters. This, in combination with a number of extensions of the new technique, makes the method a very flexible tool for analyzing subset regression problems in which the parameters have a physical meaning.</p>			
17. Key Words (Suggested by Author(s)) Subset regression QR factorization Column pivoting		18. Distribution Statement Unlimited - Unclassified Subject Category: 66	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of pages 6	22. Price A02