27.
9-29-80 948

SAND80-1260

Unlimited Release

UC-32

MASTER

The Weighted Linear Least Squares Problem: An Interval Analysis Approach to Rank Determination

Thomas A. Manteuffel

Prepared by Sandia National Laboratories, Albuquerque, New Mexico 87185
and Livermore, California 94550 for the United States Department of Energy
under Contract DE-AC04-76DP00789

Printed August 1980

Sandia National Laboratories

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Issued by Sandia Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

PAGES 1 to 2
WERE INTENTIONALLY
LEFT BLANK

Master

SAND80-1260

The Weighted Linear Least Squares Problem:
An Interval Analysis Approach to Rank Determination

Published August 1980

Thomas A. Manteuffel
Numerical Mathematics Division 5642
Sandia National Laboratories
Albuquerque, NM 87185

ABSTRACT

This is an extension of the work in Manteuffel [11] to the weighted linear least squares problem. Given the weighted linear least squares problem $WAX \cong Wb$, where W is a diagonal weighting matrix, and bounds on the uncertainty in the elements of A , we define an interval matrix A^I that contains all perturbations of A due to these uncertainties and say that the problem is rank deficient if any member of A^I is rank deficient. It is shown that if $WA = QR$ is the QR decomposition of WA , then Q and R^{-1} can be used to bound the rank of A^I . A modification of the Modified Gram-Schmidt QR decomposition yields an algorithm that implements these results. The extra arithmetic is $O(MN)$. Numerical results show the algorithm to be effective on problems in which the weights vary greatly in magnitude.

DISCLAIMER

This work was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Fig

1. Introduction

Given an $M \times N$ ($M > N$) matrix A , an $M \times 1$ vector $\underline{b}$, and an $M \times M$ positive diagonal matrix W whose diagonal elements are weights, the weighted least squares problem entails finding an $\underline{x}$ that minimizes

$$\|W(A\underline{x} - \underline{b})\|_2.$$

Let us write this as

$$(1.1) \quad W A \underline{x} \cong W \underline{b}.$$

Linear least squares problems with equality constraints are often solved by heavily weighting the equality equations, adjoining them to the least squares equations, and solving the resulting weighted linear least squares problem.

A standard technique for solving the weighted least squares problem is to factor WA such that

$$WA = QR$$

where Q is orthogonal and R is upper triangular, in which case the solution is given by

$$(1.2) \quad \underline{x} = R^{-1}Q^*W\underline{b}.$$

Another technique is to perform a singular value decomposition on WA to get

$$WA = U \Sigma V^*$$

where U and V are orthogonal and Σ is diagonal. The solution is given by

$$(1.3) \quad \underline{x} = V \Sigma^{-1} U^* W \underline{b}.$$

Problem (1.1) has a unique solution only if the matrix A is of full column rank. Numerical determination of rank is not always easy, especially in light of uncertainties in the elements of A . It is often the case that

$A = (a_{ij})$ is known to be correct only within certain tolerances and the desired but unknown $\hat{A} = (\hat{a}_{ij})$ satisfies

$$(1.4) \quad \hat{a}_{ij} = a_{ij} + \psi_{ij}, \quad |\psi_{ij}| < \max(\gamma_{ij} |a_{ij}|, \delta_{ij})$$

where γ_{ij} , δ_{ij} are relative and absolute error bounds. The uncertainty in $\hat{A}$ may cause rather large uncertainties in the solution $\underline{x}$. If any $\hat{A}$ in (1.4) is rank deficient, then a numerical solution may be completely incorrect.

Singular value decomposition of WA is of little value in detecting possible rank deficiencies when the weights vary greatly in magnitude. This is because a singular value analysis ignores the position of uncertainty in the matrix. Indeed, the ratio of the largest to the smallest singular values will be on the order of the ratio of the largest to smallest weights.

In Manteuffel [11] it was shown that a modification of the Businger-Golub QR decomposition (Businger and Golub [2]) was effective in determining the numerical rank of the unweighted least squares problem. In this paper we will show that a modification of the Modified Gram-Schmidt QR decomposition (cf. Lawson and Hanson [9], page 130) will yield bounds that are nearly as effective for the weighted least squares problem. Certain important quantities readily available in the Modified Gram-Schmidt algorithm are difficult to obtain when the Businger-Golub algorithm is used.

1.1 Notation

We will use the following notation. Capital letters will denote matrices, underlined small letters will denote vectors, and small letters will denote scalars. The $M \times N$ ($M > N$) real valued matrix

$$A = (a_{ij}) = \begin{pmatrix} | & | & \cdots & | \\ \underline{a}_1 & \underline{a}_2 & \cdots & \underline{a}_N \\ | & | & & | \end{pmatrix}$$

has a singular value decomposition

$$A = U \Sigma V^*, \Sigma = \text{diag}(\sigma_1, \dots, \sigma_N), 0 < \sigma_N < \dots < \sigma_1,$$

where $U_{M \times N}$ and $V_{N \times N}$ are orthogonal matrices. We will use the vector norms

$$\|\underline{v}\|_{\infty} = \max_i |v_i|$$

$$\|\underline{v}\|_1 = \sum_i |v_i|$$

$$\|\underline{v}\|_2 = (\sum_i |v_i|^2)^{1/2}$$

and matrix norms

$$\|A\|_{\infty} = \max_i \sum_j |a_{ij}|$$

$$\|A\|_1 = \max_j \sum_i |a_{ij}| = \max_j \|\underline{a}_j\|_1$$

$$\|A\|_2 = \max_{\|\underline{x}\|_2=1} \|\underline{Ax}\|_2 = \sigma_1$$

$$\|A\|_F = (\sum_{i,j} (a_{ij})^2)^{1/2} = (\sum_i \sigma_i^2)^{1/2}.$$

The 2-norm may appear without subscript. The vector $\underline{e}_j$ will denote the j^{th} basis vector and $\underline{e} = (1, 1, \dots, 1)^T$. We will let

$$C_i(A) = \|A\|_i \|A^+\|_i,$$

where A^+ is the pseudo-inverse of A , be the condition of A in the appropriate norm. The spectral radius of A will be denoted $S(A)$.

1.2 Problem Definition

Consider the set of matrices that includes all perturbations in (1.4).

Let

$$(1.5) \quad \mathcal{I}_0^I = \{A + \Psi : |\psi_{ij}| < \max(\gamma_{ij}|a_{ij}|, \delta_{ij})\}.$$

We will say that $\mathcal{I}_0^I$ is of full rank if every $A \in \mathcal{I}_0^I$ is of full rank. A subset of columns of $\mathcal{I}_0^I$ will be said to be linearly independent if there

exists no $A \in A_0^I$ whose corresponding columns are linearly dependent.

An independent set of interval columns will be said to be maximal if no larger independent set includes it. The Rank (A_0^I) will be the index of the largest maximal independent set of columns of A_0^I . This yields

$$(1.6) \quad \text{Rank } (A_0^I) < \min_{A \in A_0^I} \text{Rank } (A).$$

Notice that $\{a_1^I, a_j^I\}$ and $\{a_1^I, a_k^I\}$ may both be dependent sets, while $\{a_j^I, a_k^I\}$ may be an independent set. In fact, one can construct matrices for which the inequality in (1.6) is strict. However, if the right hand side equals N , then equality holds.

A desirable result would be to find a largest maximal independent set of columns of A_0^I and produce an interval solution $\underline{x}^I$ that contains all possible solutions based upon this maximal set. Such a result is possible but requires an inordinate amount of computation and storage (cf. Moore [12], Hanson and Smith [7], [8]). The goal of this paper is to obtain as much of this information as possible at a moderate increase in computation and storage over the Modified Gram-Schmidt algorithm.

As in Manteuffel [11] we will consider the larger interval matrix

$$(1.7) \quad A_2^I = \{A + \Psi: \|\Psi_j\|_2 \leq \epsilon_j, j = 1, \dots, N\}$$

where

$$\epsilon_j = \left(\sum_{i=1}^M \max(\gamma_j |a_{ij}|, \delta_j)^2 \right)^{1/2},$$

$$\gamma_j = \max_i \gamma_{ij},$$

$$\delta_j = \max_i \delta_{ij}.$$

The interval matrix A_2^I has the property that is invariant to multiplication by a unitary matrix.

1.3 Overview

In Section 2 it will be shown that the rank of A_2^I can be bounded using Q and R^{-1} of the QR decomposition of WA . In Section 3 an algorithm will be presented, based upon the Modified Gram-Schmidt QR decomposition, that implements these bounds. The rank of A_2^I may not be determined exactly. However, the rank tests presented here are more effective than those of methods that do not consider the position of the weights. The algorithm here requires the construction of R^{-1} , but remains more economical than performing an analysis on A to determine a maximal set of independent columns and then performing a QR decomposition on WA to solve the problem. Section 4 will present numerical results. The algorithm is shown to be effective on a range of problems.

Rank determination is difficult because the algorithm decomposes WA while the interval matrix is centered about A . While Ψ_2^I is invariant to multiplication by a unitary matrix, $W\Psi_2^I$ is not. This work may be viewed as rank determination of an unweighted problem with the more general interval matrix:

$$(1.12) \quad A_3^I = \{A + \Psi : |\psi_{ij}| < w_i \epsilon_j\}.$$

For A_3^I the hypothesis of Lemma 2.1 gives both necessary and sufficient conditions.

Powell and Reid [13] developed bounds on the growth of round-off errors for the Businger-Golub algorithm applied to a weighted least squares problem with weights that vary greatly in magnitude. They showed that a row pivot was necessary to prevent loss of information. The Modified Gram-Schmidt algorithm has the advantage that no row pivot is necessary. Björck [1] has studied the effect of round-off errors on the Modified Gram Schmidt algorithm. His results show that if $\bar{QR}$ is the computed decomposition of WA , then

$$(1.13) \quad \|WA - \bar{QR}\|_F < 1.5 (N-1) \tau \|WA\|_F,$$

where τ is machine precision. These results are independent of row or column pivots. Here we find that column order may have an effect on how precise the determination of rank is (see Section 3). However, numerical results show that the additional computation required to produce an optimal column order may be unwarranted.

In this paper we are concerned with uncertainties that are many orders of magnitude greater than the bounds given in (1.13). Björck's result allows us to ignore the effects of round-off in our analysis.

2. Rank of A_2^I

Let A_2^I be as given in (1.7) and let

$$(2.1) \quad \Psi_2^I = \{\Psi: \|\Psi_j\|_2 \leq \varepsilon_j, j = 1, \dots, N\}.$$

If $\hat{A} \in A_2^I$, then $\hat{A} = A + \Psi$ for some $\Psi \in \Psi_2^I$ and conversely.

Suppose $WA = QR$ is the QR decomposition of WA . We would like to use Q and R to determine the numerical rank of A_2^I . If $\hat{A} = A + \Psi \in A_2^I$ we have

$$\text{Rank}(\hat{A}) = \text{Rank}(W\hat{A}) = \text{Rank}(WA + W\Psi).$$

Assuming R is of full rank we have

$$\text{Rank}(\hat{A}) = \text{Rank}(QR + W\Psi) = \text{Rank}(I + Q^*W\Psi R^{-1}).$$

Let us define

$$(2.2) \quad \Phi = \Psi R^{-1}, \quad \Theta = WQ$$

and

$$(2.3) \quad T = \Theta^* \Phi.$$

We have then

$$(2.4) \quad \text{Rank}(\hat{A}) = \text{Rank}(I + T).$$

Since T is a square matrix we may consider the spectral radius of T , $S(T)$, as Ψ varies over Ψ_2^I . First we make the following definitions. Let

$$(2.5) \quad R^{-1} = \Delta = (\delta_{ij}),$$

$$(2.6) \quad \rho_j = \sum_{i=1}^j |\delta_{ij}| \epsilon_i,$$

$$(2.7) \quad \alpha_i = \|\underline{\theta}_i\|_2 = \|\underline{w}_{q_i}\|_2.$$

Lemma 2.1

If $\sum_{i=1}^N \alpha_i \rho_i < 1$, then $\text{Rank}(A_2^I) = N$.

Further, if there exists a subset $C \subseteq \{1, \dots, N\}$ of index k such that

$$\sum_{i \in C} \alpha_i \rho_i < 1, \text{ then } \text{Rank}(A_2^I) > k.$$

Proof First we establish the inequality:

$$(2.8) \quad \begin{aligned} \|\underline{\phi}_j\|_2 &= \left\| \sum_{i=1}^j \delta_{ij} \underline{\psi}_i \right\|_2 \leq \sum_{i=1}^j |\delta_{ij}| \|\underline{\psi}_i\|_2 \\ &\leq \sum_{i=1}^j |\delta_{ij}| \epsilon_i = \rho_j. \end{aligned}$$

Now consider the elements of $T = (\tau_{ij})$. We have

$$(2.9) \quad |\tau_{ij}| = |\langle \underline{\theta}_i, \underline{\phi}_j \rangle| \leq \|\underline{\theta}_i\|_2 \|\underline{\phi}_j\|_2 \leq \alpha_i \rho_j.$$

Thus, if we let $\underline{\alpha} = (\alpha_1, \dots, \alpha_N)^T$, $\underline{\rho} = (\rho_1, \dots, \rho_N)^T$, we have

$$|T| \leq \underline{\alpha} \underline{\rho}^T.$$

By the Perron-Frobenius Theorem (Varga [14], page 30) we have

$$S(T) \leq S(\underline{\alpha} \underline{\rho}^T) = \sum_{i=1}^N \alpha_i \rho_i.$$

If $S(T) < 1$ for every $\Psi \in \Psi_2^I$, then A_2^I is of full rank.

The second result follows from the fact that the principal submatrix of $I + T$ consisting of the k rows and columns corresponding to the subset C is of full rank for every $\Psi \in \Psi^I$. Thus, there are at least k independent interval columns in A_2^I . ■

The inequality (2.9) is sharp. We can always find $\Psi \in \Psi_2^I$ such that

$$|\tau_{kl}| = \alpha_k \rho_l. \text{ Let}$$

$$(2.10) \quad \underline{v}_\ell = (\text{Sign}(\delta_{1\ell})\varepsilon_1, \dots, \text{Sign}(\delta_{\ell\ell})\varepsilon_\ell, 0, \dots, 0)^T$$

and

$$(2.11) \quad \psi(k, \ell) = \frac{1}{\|\theta_k\|_2} \theta_k^* \underline{v}_\ell^T.$$

Then $\psi(k, \ell) \in \Psi_2^I$ and

$$\begin{aligned} T(k, \ell) &= \Theta \star \psi(k, \ell) R^{-1} \\ &= \left(\frac{1}{\|\theta_k\|_2} \Theta \star \theta_k \right) (\underline{v}_\ell^T R^{-1}). \end{aligned}$$

Let

$$(2.12) \quad \underline{\rho}^{(\ell)} = R^{-T} \underline{v}_\ell, \quad \rho_i^{(\ell)} = \sum_{m=1}^{\ell} \text{Sign}(\delta_{m\ell}) \delta_{mi} \varepsilon_m,$$

$$(2.13) \quad \underline{\alpha}^{(k)} = \frac{1}{\|\theta_k\|_2} \Theta \star \theta_k, \quad \alpha_i^{(k)} = \frac{\langle \theta_i, \theta_k \rangle}{\|\theta_k\|_2}.$$

Clearly,

$$(2.14) \quad \rho_\ell^{(\ell)} = \rho_\ell, \quad |\rho_i^{(\ell)}| \leq \rho_i,$$

$$(2.15) \quad \alpha_k^{(k)} = \alpha_k, \quad |\alpha_i^{(k)}| \leq \alpha_i.$$

We have

$$(2.16) \quad T(k, \ell) = \underline{\alpha}^{(k)} \underline{\rho}^{(\ell)T}$$

and

$$(2.17) \quad \tau_{k\ell}^{(k, \ell)} = \alpha_k \rho_\ell.$$

One can think of $T^{(k, \ell)}$ $k = 1, \dots, N; \ell = 1, \dots, N$ as corner elements in the set

$$(2.18) \quad T_2^I = \{T : T = \Theta \star \Psi R^{-1}, \Psi \in \Psi_2^I\},$$

because there is no $T \in T_2^I$ for which $|\tau_{k\ell}| > \tau_{k\ell}^{(k, \ell)}$. In fact, since T_2^I is a convex set, the smallest convex set containing the corner elements is a subset of T_2^I . The advantage to the corner elements is that they each have only one nonzero eigenvalue which is real and can be easily calculated.

We have

$$(2.19) \quad S(T^{(k,\ell)}) = |\langle \underline{\alpha}^{(k)}, \underline{\rho}^{(\ell)} \rangle| = \left| \sum_{i=1}^N \alpha_i^{(k)} \rho_i^{(\ell)} \right|.$$

This yields the following result.

Lemma 2.2

If $S(T^{(k,\ell)}) > 1$ for any (k,ℓ) , then A_2^I is rank deficient.

Proof

If $S(T^{(k,\ell)}) > 1$, then there exists $T \in T_2^I$ with an eigenvalue equal to -1 , namely

$$T = \frac{-1}{\langle \underline{\alpha}^{(k)}, \underline{\rho}^{(\ell)} \rangle} T^{(k,\ell)}.$$

Thus $I + T$ is rank deficient. ■

A more useful result for our purposes is the following.

Lemma 2.3

If $\alpha_j \rho_j > 1 + \sum_{i \neq j} \alpha_i \rho_i$ for some j , then A_2^I is rank deficient.

Proof. Consider $T^{(j,j)} = \underline{\alpha}^{(j)} \underline{\rho}^{(j)T}$. We have

$$\begin{aligned} S(T^{(j,j)}) &= \left| \sum_{i=1}^N \alpha_i^{(j)} \rho_i^{(j)} \right| \\ &= \alpha_j \rho_j + \left| \sum_{i \neq j} \alpha_i^{(j)} \rho_i^{(j)} \right| \\ &> \alpha_j \rho_j - \sum_{i \neq j} \alpha_i \rho_i. \end{aligned}$$

Thus $S(T^{(j,j)}) > 1$ and, as in the proof of Lemma 2.2, A_2^I is rank deficient. ■

There is another measure of possible rank deficiency. Let us define

$$(2.20) \quad \beta_i = \|W^{-1} \underline{q}_i\|_2 = \|W^{-2} \underline{\theta}_i\|_2$$

We have the following result:

Lemma 2.4

If $\rho_j > \beta_j$ for some j , then A_2^I is rank deficient.

Proof Let $\underline{v}_j$ be defined as in (2.10). Define

$$\Psi = \frac{-1}{\rho_j} W^{-1} Q \underline{e}_j \underline{v}_j^T;$$

then,

$$\|\underline{\psi}_\ell\|_2 = \frac{1}{\rho_j} \|W^{-1} Q \underline{e}_j\|_2 \epsilon_\ell = \frac{\beta_j}{\rho_j} \epsilon_\ell < \epsilon_\ell,$$

which yields $\Psi \in \Psi_2^I$. Now

$$T = Q^* W \Psi R^{-1} = -\frac{1}{\rho_j} \underline{e}_j \underline{\rho}^{(j)T},$$

and

$$(I + T) \underline{e}_j \equiv 0.$$

The result follows. ■

The results of this section can be combined to yield the following theorems:

Theorem 2.5

If $\rho_j > \beta_j$ for some j , then A_2^I is rank deficient.

If $\rho_j < \frac{1}{\alpha_j \sqrt{N}}$ for every j , then A_2^I is of full rank.

Further, if $k = \text{index} \{ \rho_j : \rho_j < \frac{1}{\alpha_j \sqrt{N}} \}$, then $\text{Rank} (A_2^I) > k$.

Proof If $\alpha_j \rho_j < \frac{1}{\sqrt{N}}$ for every j , then $\sum_{i=1}^N \alpha_i \rho_i < 1$.

The results follow from the above lemmas. ■

Theorem 2.6

If $\alpha_j \rho_j > 1 + \sum_{i \neq j} \alpha_i \rho_i$ for some j , then A_2^I is rank deficient.

If $\sum_{i=1}^N \alpha_i \rho_i < 1$, then A_2^I is of full rank.

Further, if there exists a subset $C \subseteq \{1, \dots, N\}$ of index k such that

$\sum_{i \in C} \alpha_i \rho_i < 1$, then $\text{Rank} (A_2^I) > k$. ■

Notice that since Q is orthogonal we have

$$1 = \langle W\underline{q}_i, W^{-1}\underline{q}_i \rangle \leq \|W\underline{q}_i\|_2 \|W^{-1}\underline{q}_i\|_2 = \alpha_i \beta_i.$$

Thus, $\frac{1}{\alpha_i} \leq \beta_i$ and the rank test in Theorem 2.5 may have a gap larger

than $\sqrt{N}$. It is clear that

$$\alpha_i \beta_i \leq \|W\|_2 \|W^{-1}\|_2 = w_{\max}/w_{\min},$$

the ratio of the largest weight to the smallest weight. If the elements of A are all of the same order of magnitude, a singular value decomposition of WA will yield $C_2(WA) = \sigma_1/\sigma_N$ at least as large as $w_{\max}/w_{\min}$. At its worst Theorem 2.5 is no worse than rank determination based on an interval matrix centered about WA .

In practice one would expect $\alpha_i \beta_i$ to be of order unity. Consider a matrix A whose elements are all of order unity and W with

$$\begin{aligned} w_i &= 10^{+6} & i=1, \dots, k, \\ w_i &= 1 & i=k+1, \dots, M. \end{aligned}$$

The first k rows are heavily weighted. Thus, each column of WA is nearly in the subspace spanned by $\{\underline{e}_1, \dots, \underline{e}_k\}$. Consider $\underline{q}_i$, $i=1, \dots, k$. Since $\underline{q}_i$ is of unit length we would have

$$\underline{q}_i \sim \sum_{j=1}^k \underline{e}_j + \sum_{j=k+1}^M 10^{-6} \underline{e}_j$$

which yields

$$\begin{aligned} \alpha_i &= \|W\underline{q}_i\|_2 \sim 10^{+6}, \\ \beta_i &= \|W^{-1}\underline{q}_i\|_2 \sim 10^{-6}. \end{aligned}$$

Now consider $\underline{q}_i$, $i = k+1, \dots, N$. Since Q is orthogonal we would have

$$\underline{q}_i \sim \sum_{j=1}^k 10^{-6} \underline{e}_j + \sum_{j=k+1}^M \underline{e}_j$$

which yields

$$\begin{aligned} \alpha_i &= \|W\underline{q}_i\|_2 \sim 1, \\ \beta_i &= \|W^{-1}\underline{q}_i\|_2 \sim 1. \end{aligned}$$

This pattern is borne out in numerical tests. In fact, even when the weights are gradually decreased from $w_1 = 10^{+6}$ to $w_M = 1$ we have $\alpha_1 \beta_1 \sim 1$, that is, until round-off error causes a loss of orthogonality. Consider the example where the first k rows are heavily weighted. If for some $j > k+1$ round-off error has caused $\underline{q}_j$ to no longer be nearly orthogonal to $\{\underline{e}_1, \dots, \underline{e}_k\}$ then we could expect $\alpha_j \sim 10^{+6}$, $\beta_j \sim 1$. This provides an indication that reorthogonalization is necessary. Implementation will be discussed in Section 3.

Finally, notice that if all the weights are set to unity, that is if $W = I$, then $\alpha_i = \beta_i = 1$ and Theorem 2.5 reduces to Theorem 2.3 in Manteuffel [11] and Theorem 2.6 is similar to Theorem 2.2.

3. The Algorithm

The modified Gram-Schmidt QR decomposition is carried out in N steps. If we let $WA = A^{(0)}$, $D^{(0)} = I$, $R^{(0)} = I$, then at each step

$$(3.1) \quad WA = A^{(k)} D^{(k)-1} R^{(k)},$$

and after N steps

$$(3.2) \quad A^{(N)} D^{(N)-1} = Q,$$

$$R^{(N)} = R.$$

Here, $D^{(k)} = \text{diag}(d_1^{(k)}, \dots, d_N^{(k)})$ is a normalization matrix and $R^{(k)}$ is upper triangular. At each step the matrices are updated by the formulas.

$$(3.3) \quad \underline{a}_k^{(k)} = \underline{a}_k^{(k-1)},$$

$$d_k^{(k)} = \|\underline{a}_k^{(k)}\|_2,$$

$$r_{kk}^{(k)} = d_k^{(k)},$$

for $j = k+1, \dots, N$

$$r_{kj}^{(k)} = \frac{\langle \underline{a}_k^{(k)}, \underline{a}_j^{(k)} \rangle}{d_k^{(k)}},$$

$$\underline{a}_j^{(k)} = \underline{a}_j^{(k-1)} - \frac{r_{kj}^{(k)}}{d_k^{(k)}} \underline{a}_k^{(k)}.$$

The algorithm can be rearranged to avoid taking square roots for $d_k^{(k)}$ (cf. Lawson and Hanson [9], page 130). For our purposes, however, it will be more convenient to use the above form. Since $d_j^{(k)} = r_{jj}^{(k)}$ for all j, k it will be unnecessary to store D .

Let us write

$$(3.4) \quad \Lambda^{(k)} = \begin{pmatrix} \Lambda_1^{(k)} & \Lambda_2^{(k)} \end{pmatrix}$$

where $\Lambda_1^{(k)}$ is the first k columns of $\Lambda^{(k)}$ and

$$(3.5) \quad R^{(k)} = \begin{pmatrix} R_{11}^{(k)} & R_{12}^{(k)} \\ 0 & R_{22}^{(k)} \end{pmatrix}$$

where $R^{(k)}$ is $k \times k$ and upper triangular and $R_{22}^{(k)} = I$. The elements of $R_{11}^{(k)}$ and $R_{12}^{(k)}$ are fixed (up to column interchanges in $R_{12}^{(k)}$) for step $k+1, \dots, N$. We can construct the inverse of $R^{(k)}$ as we proceed. Consider the matrix

$$(3.6) \quad \Lambda^{(k)} = (\delta_{ij}^{(k)})$$

where the elements are updated at each step according to the formulas

$$(3.7) \quad \delta_{kk}^{(k)} = \frac{1}{r_{kk}};$$

for $i = 1, \dots, k-1$

$$(3.8) \quad \delta_{ik}^{(k)} = -\delta_{kk}^{(k)} \cdot \delta_{ik}^{(k-1)};$$

for $j = k+1, \dots, N$

for $i = 1, \dots, k-1$

$$(3.9) \quad \delta_{ij}^{(k)} = \delta_{ij}^{(k)} + \delta_{ik}^{(k)} \cdot r_{kj}^{(k)},$$

$$\delta_{kj}^{(k)} = \delta_{kk}^{(k)} r_{kj}^{(k)}.$$

We may write

$$(3.10) \quad \Delta^{(k)} = \begin{pmatrix} \Delta_{11}^{(k)} & \Delta_{12}^{(k)} \\ 0 & \Delta_{22}^{(k)} \end{pmatrix},$$

where

$$(3.11) \quad \Delta_{11}^{(k)} = (R_{11}^{(k)})^{-1},$$

$$\Delta_{12}^{(k)} = (R_{11}^{(k)})^{-1} R_{12}^{(k)},$$

$$\Delta_{22}^{(k)} = I.$$

At step k we may compute the quantities ρ_k of (2.6) and α_k of (2.7).

We have

$$(3.12) \quad \rho_k = \sum_{i=1}^k |\delta_{ik}^{(k)}| \epsilon_i,$$

$$\alpha_k = \frac{1}{d_k^{(k)}} \|W_{\underline{k}}^{(k)}\|_2 = \delta_{kk}^{(k)} \|W_{\underline{k}}^{(k)}\|_2.$$

This provides the following test for rank deficiency.

a) If $\alpha_k \rho_k + \sum_{i=1}^{k-1} \alpha_i \rho_i < 1$, then $\{\underline{a}_1^I, \dots, \underline{a}_k^I\}$

is a linearly independent set and $\text{Rank}(\underline{A}_2^I) > k$.

b) If $\alpha_k \rho_k > 1 + \sum_{i=1}^{k-1} \alpha_i \rho_i$, then $\{\underline{a}_1^I, \dots, \underline{a}_k^I\}$

is a linearly dependent set and $\underline{a}_k$ may not be used in computing a solution.

This column should be ordered last and only the remaining columns need be considered.

c) If $1 - \sum_{i=1}^{k-1} \alpha_i \rho_i < \alpha_k \rho_k < 1 + \sum_{i=1}^{k-1} \alpha_i \rho_i$, then

the above tests are inconclusive. We may compute

$$(3.13) \quad \beta_k = \delta_{kk}^{(k)} \|W^{-1} \underline{a}_k^{(k)}\|_2,$$

as in (2.20). If $\rho_k > \beta_k$, then $\{\underline{a}_1^I, \dots, \underline{a}_k^I\}$ is a dependent set.

d) If all tests are inconclusive then $\underline{a}_k$ will be accepted.

It was found in practice that when test a) failed it might be due to loss of orthogonality. An effective strategy is to reorthogonalize $\underline{a}_k^{(k)}$ in reverse order, that is, for $j = k-1, \dots, 1$

$$\underline{a}_k^{(k)} + \underline{a}_k^{(k)} - \frac{\langle \underline{a}_j^{(k)}, \underline{a}_k^{(k)} \rangle}{(d_j^{(k)})^2} \underline{a}_j^{(k)},$$

and repeat test a).

In this way a maximal set of linearly independent columns is constructed. If we let K_1 be the maximum number of columns that pass test a) and K_2 be the number of columns that are accepted, then we have

$$(3.14) \quad K_1 < \text{Rank}(A_2^I) < K_2.$$

The difference $K_2 - K_1$ will yield a measure of how precise the rank determination is. One can construct pathological examples in which $K_1 \ll K_2$, but such a case will not go undetected.

It is possible to order the columns so that $\alpha_{k+1} \rho_{k+1}$ is as small as possible at step $k+1$. Suppose after k steps we rearrange the order of the columns of $A_2^{(k)}$ and correspondingly the columns of $R_{12}^{(k)}$. Let $P^{(k+1)}$ be the column permutation matrix that performs this. If

$$W A P^{(1)} \dots P^{(k)} = A^{(k)} D^{(k)} R^{(k)-1},$$

then

$$W A P^{(1)} \dots P^{(k+1)} = A^{(k)} D^{(k)-1} P^{(k+1)} P^{(k+1)} R^{(k)} P^{(k+1)}.$$

Since $P^{(k+1)}$ only involves the positions $k+1, \dots, N$ and since $R_{22}^{(k)} = I$, $P^{(k+1)} R^{(k)} P^{(k+1)}$ merely rearranges the column order of $R_{12}^{(k)}$.

We would like to choose column $k+1$ to yield the smallest $\alpha_{k+1} \rho_{k+1}$. This can be accomplished by examining the columns $k+1, \dots, N$ in turn. For $j = k+1, \dots, N$ let

$$(3.15) \quad d_j^{(k+1)} = \|a_j^{(k)}\|_2,$$

$$(3.16) \quad \alpha_j^{(k+1)} = \frac{1}{d_j^{(k+1)}} \|W a_j^{(k)}\|_2,$$

$$(3.17) \quad \rho_j^{(k+1)} = \frac{1}{d_j^{(k+1)}} \left(\epsilon_j + \sum_{i=1}^k |\delta_{ij}^{(k)}| \epsilon_i \right).$$

Then choose column $k+1$ so that

$$(3.18) \quad \alpha_{j+k} \rho_{k+1} = \alpha_{k+1}^{(k+1)} \rho_{k+1}^{(k+1)} < \alpha_j^{(k+1)} \rho_j^{(k+1)}, \quad j = k+1, \dots, N.$$

The basic algorithm described in equations (3.3) requires $MN^2 + MN$ operations.⁽¹⁾ The computation of Δ requires approximately $N^3/6 - N^2$ operations. The computation of α_k and ρ_k at each step requires $3MN + N^2/2$ operations. This yields a total of about $MN^2 + N^3/6 + 4MN - N^2/2$ operations. If, however, we use the equations (3.18) - (3.20) to select a column order the algorithm requires an additional $3/2 MN^2 + 3/2 MN + N^3/6 - N^2$ operations. This seemingly inordinate amount of extra arithmetic may not be as costly as it appears on a vector machine as it is the computation of the norm of a vector in (3.15) that produces the $3/2 MN^2$ term. Numerical experiments indicate that pivoting does make rank determination more precise. However, the algorithm is quite effective without pivoting.

(1) Here an operation is one addition and one multiplication.

4. Numerical Results

The questions that must be answered by numerical tests are as follows:

1. How accurately does the algorithm predict the rank of A_2^I when the weights vary greatly in magnitude? This can be answered by comparing the numerical rank computed by this algorithm (WLS) operating on WA with the numerical rank produced by the algorithm described in Manteuffel [11] (LS) operating on A.

2. Is this algorithm an improvement to applying an algorithm that ignores the weights? For this we compare (WLS) to (LS) both applied to WA.

3. Does pivoting yield a better prediction of the numerical rank? This can be measured by comparing the numerical ranks computed by (WLS) with and without pivoting operating on WA with the numerical rank produced by (LS) operating on A. It will be seen that pivoting increases the precision. That is, it reduces the gap between K_1 and K_2 .

In each of the trials an $M \times N$ matrix $A = (a_{ij})$ was constructed with $N \in [10, 50]$, $M \in [N, 2N]$ and $a_{ij} \in [-1, 1]$ (each uniformly distributed). Then an integer K was chosen so that $N-K \in [0, 5]$ and the last $N-K$ columns of A were made to be linear combinations of the first K columns. Finally a matrix $\Psi = (\psi_{ij})$ with $\psi_{ij} \in [-10^{-5}, 10^{-5}]$ was added to A . Thus, if the uncertainty were set so that $\epsilon_i > 10^{-5}$, $i = 1, \dots, N$ one would expect that A_2^I would have rank K or less. If the uncertainty were set so that $\epsilon_i < 10^{-5}$, $i = 1, \dots, N$ one would expect each matrix to be of full rank. Since the matrices are randomly chosen the rank cannot be predicted precisely a priori.

Three different weighting schemes were used. In the first, W_1 , a number $L \in [1, 5]$ was chosen and the first L weights were set to 10^{+6} .

The remaining weights were set to 1. In the second, w_2 , a number $NEX \in [0,6]$ was chosen for each i and $w_1 = 10^{NEX}$. In the third, w_3 , a number of $\gamma > 0$ was calculated so that $\gamma^{(M-1)} = 10^{+6}$. Then $w_i = \gamma^{(i-1)}$, $i = 1, \dots, M$.

Weighting scheme w_1 was used with relative uncertainty $\gamma_j = 10^{-3}$, 10^{-5} , 10^{-7} , 10^{-9} and absolute uncertainty $\delta_j = 0$, $j = 1, \dots, N$. One hundred trials were run at each value for a total of 400 trials. In each trial (WLS) operating on WA predicted the same rank as (LS) operating on A. Further, in each trial $K_1 = K_2$, that is, there was no question about the rank of A_2^I . On the other hand, (LS) operating on WA computed the proper rank 0 times for $\gamma_j = 10^{-3}$, 0 times for $\gamma_j = 10^{-5}$, 16 times for $\gamma_j = 10^{-7}$ and 47 times for $\gamma_j = 10^{-9}$. The rank predicted by (LS) on WA was as much as 35 less than the correct rank. This reflects the fact that the interval matrix used by (LS) is centered at WA and is much larger than A_2^I .

Without pivots, (WLS) still predicted the correct rank in each trial. Now, however, for $\gamma_j = 10^{-3}$ the gap between K_1 and K_2 was no longer zero in each trial. Of the 400 trials the gap was equal to 1 in 4 trials and 2 in 2 trials.

When weighting scheme w_1 was used with $\gamma_j = 10^{-1}$, (WLS) operating on WA predicted the same rank as (LS) operating on A in only 47 out of 100 trials. Both (WLS) on WA and (LS) on A had problems with the precision of the numerical rank. For (LS) on A, the gap between K_1 and K_2 was as large as 17 while for (WLS) on WA the gap was as large as 38. This is because the uncertainty is on the same order of magnitude as the length of the columns of A. The algorithm (LS) applied to WA did even worse. None of these algorithms was designed to sail in such rough waters.

The results for weighting schemes W_2 and W_3 were similar. One hundred trials were run for each weighting scheme at each of the uncertainties $\gamma_j = 10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}$, $j = 1, \dots, N$. In each of the 800 trials (WLS) on WA predicted the same rank as (LS) on A. For $\gamma_j = 10^{-5}, 10^{-7}, 10^{-9}$, there was no gap between K_1 and K_2 in any trial. For $\gamma_j = 10^{-3}$ the gap was larger. For W_2 , 87 of the 100 trials had zero gap while for W_3 , 82 of the 100 trials had zero gap. Values for $K_2 - K_1$ as large as 9 occurred for W_2 and as large as 11 for W_3 . It is expected that the algorithm will not perform as well when the weights take on all values between 1 and 10^6 . This type of weighting is more likely to make the quantity $\alpha_i \beta_i \gg 1$ (See Section 3). Again, (LS) on WA was not very effective. For W_3 , the predicted numerical rank was the same as that calculated by (LS) on A in only 8 trials for $\gamma_j = 10^{-3}$, 79 trials for $\gamma_j = 10^{-5}$, 24 trials for $\gamma_j = 10^{-7}$, and 83 for $\gamma_j = 10^{-9}$. Results for W_2 were similar.

We may conclude that the algorithm (WLS) applied to WA is nearly as effective as the algorithm (LS) applied to A in determining the rank of A_2^I for a wide range of problems. Thus rank determination and decomposition can be carried out simultaneously in the weighted least squares problem. We may also conclude that the extra computation required to choose a pivot sequence is unwarranted.

As a final test, the length of the shortest least squares solution and the length of the residual vector produced by (WLS) were compared to the corresponding values produced by the psuedo-inverse of the same rank. A singular value decomposition was used to construct the psuedo-inverse. As might be expected both pairs of values agreed to within the relative uncertainty used to compute the numerical rank, sometimes many orders of

magnitude less. This indicates that (WLS) is producing a good approximation to the pseudo-inverse.

References

- [1] A. Bjorck, "Solving Linear Least Squares Problems By Gram-Schmidt Orthogonalization," BIT, 7, (1967), 1-21.
- [2] P. Businger and G. Golub, "Linear Least Squares by Householder Transformations" Numer. Math. 7, 269-276, (1965).
- [3] A. K. Cline, C. B. Moler, G. W. Stewart, and J. H. Wilkinson, "An Estimate for the Condition Number of a Matrix," SIAM J. Numer. Anal. Vol. 16, No. 2, 368-375, (1979).
- [4] G. Golub and W. Kahan, "Calculating the Singular Values and Pseudo-Inverse of a Matrix," SIAM J. of Numer. Anal., Ser. B, Vol. 2, No. 2, 205-224 (1965).
- [5] G. Golub, "Numerical Methods for Solving Linear Least Squares Problems," Numer. Math. 7, 206-216, (1965).
- [6] G. Golub, V. Klement, G. W. Stewart, "Rank Degeneracy and Least Squares Problems," Stanford Report, Stan-CS-76-559, Stanford University, (1976).
- [7] E. Hanson, "Interval Arithmetic in Matrix Computations, Part I," SIAM J. Numer. Anal., Vol. 2, No. 2, 308-320, (1965).
- [8] E. Hanson and R. Smith, "Interval Arithmetic in Matrix Computations, Part II," SIAM J. Numer. Anal., Vol. 4, No. 1, 1-9, (1967).
- [9] C. Lawson and R. Hanson, Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, NJ, (1974).
- [10] T. A. Manteuffel, "Numerical Rank Determination in Linear Least Squares Problems," Sandia Laboratories Report, SAND79-8243, (August 1979).
- [11] T. A. Manteuffel, "An Interval Analysis Approach to Rank Determination in Linear Least Squares Problems" SAND80-0655 (June 1980).

- [12] R. E. Moore, Interval Analysis, Prentice-Hall, Englewood Cliffs, NJ, (1966).
- [13] M. J. D. Powell and J. K. Reid, "On Applying Householder Transformations to Linear Least Squares Problems," Proc. IFIP 1968 Vol. 1, 122-126, (1969).
- [14] R. S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1962.

Distribution:

H. van der Vorst
Academic Computer Center Utrecht
Budapestlaan 6
De Uithof-Utrecht
The Netherlands 030 531436

L. W. Ehrlich
Applied Physics Laboratory
The Johns Hopkins University
Johns Hopkins Road
Laurel, MD 20810

Dr. John K. Reid
Computer Science System Division
Building 8.9 AERE
Harwell, Didcot
England OX11 0RA

Professor Richard S. Varga
Department of Mathematics
Kent State University
Kent, OH 44242

James C. T. Pool
ER-17, J-309, GTN
Basic Energy Sciences
Department of Energy
Washington, DC 20545

B. L. Buzbee, C-3
Los Alamos Scientific Laboratory
P. O. Box 1663
Los Alamos, NM 87545

Paul Concus
University of California
Lawrence Berkeley Laboratory
5DA-2129
Berkeley, CA 94720

I. Karasallo
University of California
Lawrence Berkeley Laboratory
5DA-2129
Berkeley, CA 94720

J. Dongarra
Applied Mathematics Department
Argonne National Laboratories
9700 S. Cass Avenue
Argonne, IL 60439

F. N. Fritsch, L-300
Lawrence Livermore Laboratory
University of California
P. O. Box 808
Livermore, CA 94550

A. Greenbaum, L-300
Lawrence Livermore Laboratory
University of California
P. O. Box 808
Livermore, CA 94550

G. Rodrique
Lawrence Livermore Laboratory
University of California
P. O. Box 808
Livermore, CA 94550

G. H. Golub
Dept. of Computer Science
Stanford University
Stanford, CA 94305

R. Sweet
National Center for Atmospheric Research
P. O. Box 3000
Boulder, CO 80303

Paul Saylor
Department of Computer Science
University of Illinois
Urbana, IL 61820

Professor Cleve B. Moler
Department of Mathematics
University of New Mexico
Albuquerque, NM 87106

2620 R. J. Detry
2646 M. R. Scott
2646 K. Haskell
2646 H. Watts
3141 T. L. Werner (5)
3151 W. L. Garner (3)
For DOE/TIC (Unlim. Release)
3154-3 R. P. Campbell (25) (For DOE/TIC)
5600 D. B. Shuster
Attn: 5610 A. A. Lieber
5620 M. M. Newsom
5630 R. C. Maydew
5640 G. J. Simmons

5641 R. J. Thompson
5642 L. F. Shampine
5642 R. J. Hanson
5642 K. L. Hiebert
5642 T. A. Manteuffel (25)
8000 T. B. Cook
Attn: 8100 W. J. Spencer
8200 A. N. Blackwell
8300 B. F. Murphey
8400 L. Gutierrez

8330 G. W. Anderson
8331 R. J. Kee
8332 R. E. Huddleston
8332 R. E. Chang
8332 D. L. Crawford
8332 T. H. Jefferson
8333 A. G. Schuknecht
8354 B. Sanders