MATHEMATICS OF COMPUTATION, VOLUME 33, NUMBER 145
JANUARY 1979, PAGES 171–183

Computer Solution and Perturbation Analysis of Generalized Linear Least Squares Problems

By C. C. Paige*

Abstract. A new formulation of the generalized linear least squares problem is given. This is based on some ideas in estimation and allows complete generality in that there are no restrictions on the matrices involved. The formulation leads directly to a numerical algorithm involving orthogonal decompositions for solving the problem. A perturbation analysis of the problem is obtained by using the new formulation and some of the decompositions used in the solution. A rounding error analysis is given to show that the algorithm is numerically stable.

1. Introduction. An important problem that has been treated at length in the numerical literature is the linear least squares problem: find the n-dimensional vector x that minimizes

$$\|Cx - y\|^2 = (Cx - y)^T(Cx - y), \tag{1}$$

where C is a given real m by n matrix, y is a given real m-dimensional vector and superscript T denotes transpose. See for example [2] to [14].

A closely related problem is the generalized linear least squares problem: find x that minimizes

$$(Cx - y)^T W^{-1}(Cx - y), \tag{2}$$

where in addition W is a given real symmetric positive definite m by m matrix. Perhaps the main use of this latter numerical problem is in the estimation of linear systems, see for example [15] and [16]. In such problems the vector y of measurements is given where y is known to be related to x by

$$y = Cx + w, \tag{3}$$

w being an unknown noise vector with zero mean and variance-covariance matrix (covariance matrix) $\sigma^2 W$. Here W is a known m by m nonnegative definite matrix, and σ^2 is an unknown nonnegative scalar. That is, if $E(\cdot)$ is used to denote the expected value,

$$E(w) = 0, \quad E(ww^T) = \sigma^2 W. \tag{4}$$

For this linear model, the vector x that minimizes (2) is called the least squares estimate of x, and its properties are discussed in [15]. The W^{-1} in (2) can be thought of as a way of taking into account the relative importance of noise elements, and the minimi-

Received July 11, 1977; revised April 6, 1978.
AMS (MOS) subject classifications (1970). Primary 62J05, 65F30; Secondary 65F35, 65G05.
Key words and phrases. Covariance matrices, error analysis, estimation of linear systems, linear least squares, matrix computations, perturbation analysis, regression analysis.
*Supported by National Research Council of Canada Grant A8652.

0025-5718/79/0000-0011/$04.50

zation then finds x corresponding to "smallest" noise in (3). A good introduction to the statistical ideas here is given in [16].

The numerical solution of this problem has not been fully treated in the literature, although methods are available. However, a method such as that described in [10, p. 185] and used in [17] and [8] and elsewhere in the numerical, engineering, econometric, and statistical literatures, can be numerically unstable when W in (2) is close to singular, and fail completely when W is singular. Björck [1] has designed a method to handle less than full rank W, and his method will work well when the nonzero eigenvalues of W are all of the same order. However, it is unstable in that it can lose accuracy unnecessarily when W is ill-conditioned for solution of equations.

It is the purpose of this paper to examine the problem in the setting of (2) and (3) in order to produce a natural formulation and solution. A proof of numerical stability of the algorithm will be given, along with a perturbation analysis for the whole problem. One advantage of the present method is that it follows directly from the new formulation of the problem, and this formulation appears to be the most natural and general one for the problem. Another advantage is that both formulation and method combine to give a reasonable perturbation analysis of the problem.

2. Problem Formulation. The formulation in (2) breaks down when W is singular, and yet a positive semidefinite W in (4) is perfectly meaningful. Here a formulation which allows any matrix C and any symmetric nonnegative definite W will be given. Any such m by m W of rank k has a factorization

$$W = BB^T, \tag{5}$$

where B is m by k of rank k. For example, the Cholesky factorization $W = LL^T$ could be carried out as in [10, p. 124], ensuring that a column of the lower triangular matrix L is zero whenever its diagonal element is also zero. B would then be obtained by deleting the zero columns of L. The decomposition can also be arranged to have L upper triangular. If there is uncertainty about the rank of W, it might be preferable to use the eigendecomposition of W.

In (4) and (5) B is a more basic matrix than W, and will often be directly available. Since computing W from a given B can lose information (see for example [6]), we will assume from now on that B is given.

It can be shown that a random vector w satisfying (4) with (5) can be expressed

$$w = Bv, \quad E(v) = 0, \quad E(vv^T) = \sigma^2 I, \tag{6}$$

where v is a k-dimensional random vector. As a result, (3), (4), and (5) give the linear model

$$y = Cx + Bv, \quad E(v) = 0, \quad E(vv^T) = \sigma^2 I. \tag{7}$$

Since all the elements of v can be treated equally, it makes sense to formulate the problem as

$$\underset{v,x}{\text{minimize}}\ v^T v \text{ subject to } y = Cx + Bv. \tag{8}$$

This is a very general formulation in that it allows all C and B, and any y that could have come from the linear model (7). It is straightforward to show that when W is non-

singular, a vector x solving (8) will also minimize (2), and so this formulation is consistent with present ideas. It can be shown that the formulation in (8) leads to the same solution that Rao obtains in his unified theory of least squares [15]. The advantage of the formulation (8) is that it appears to be easier to derive and work with than Rao's approach which is based on generalized inverses. Most importantly, formulation (8) will greatly facilitate perturbation analysis and will easily lead to numerically stable computations. It is also very amenable to treating special cases where C and B have special form—as in many engineering problems (see for example [17]). Rao's approach does not lead directly to good computations and does not appear to offer any advantages for special problems.

3. Problem Solution. C. L. Lawson and the referee both pointed out that once the realization is made that the problem can be formulated as in (8), the solution is fairly straightforward. For example (8) could be expressed as

$$\text{minimize} \left\| [0, I] \begin{pmatrix} x \\ v \end{pmatrix} \right\| \text{subject to } [C \quad B] \begin{pmatrix} x \\ v \end{pmatrix} = y,$$

a simple equality-constrained least squares problem. One of the general methods in [10] could then be applied. The method in [10, Chapter 20] appears to be the most numerically reliable of these, although no rounding error analysis is given. The method in [10, Chapter 21] has the same type of numerical instability we are trying to avoid here. If we extend the approach in [10, Chapter 20] to the rank deficient case, we will find orthogonal matrices Q and P to give

$$Q^T [C \quad B] P = \begin{pmatrix} T & 0 \\ 0 & 0 \end{pmatrix}, \quad T \text{ nonsingular},$$

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = P^T \begin{pmatrix} x \\ v \end{pmatrix}, \quad P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}, \quad \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = Q^T y.$$

The constraints then become

$$Tw_1 = s_1, \quad s_2 = 0.$$

The first of these can always be solved for w_1, but for consistency y must be such that the second automatically holds. We then need only solve

$$\text{minimize } \| P_{22} w_2 + P_{21} w_1 \|$$

for w_2 and then reconstruct x and v.

Using such a well-known technique could save time in finding solutions to problems of the form (8) for general C and B. However, such an approach does not treat x, v, C, B separately. In the statistical context where such problems usually arise, x is a vector of parameters while v is a noise vector, and it is important in the analysis to treat them separately. Here we will give a numerically stable algorithm that takes advantage of the special form of (8), and maintains x, v, C, B as separate throughout. This will allow us to carry out a perturbation analysis of the problem based on the resulting decompositions. For problems with special structure, as for example in [17], it is also important to maintain x, v, C, B as separate during the computation.

First, decompose C as is usual in the ordinary least squares problem [4], [6]

$$Q^TC = \begin{pmatrix} Q_1^TC \\ Q_2^TC \end{pmatrix} = \begin{pmatrix} R \\ 0 \end{pmatrix}, \quad Q = (Q_1, Q_2) \text{ orthogonal}, \tag{9}$$

so that R has full row rank. Column pivoting can be used; or if there is some uncertainty as to the rank of C, the singular value decomposition of C could be obtained [7]. The constraints in (8) then split in two

$$Q_1^Ty = Rx + Q_1^TBv, \tag{10}$$

$$Q_2^Ty = Q_2^TBv. \tag{11}$$

Since R has full row rank, (10) can always be solved for x once v is given, and so (11) gives the constraints on v, and (8) becomes

$$\min_v v^Tv \text{ subject to } Q_2^Ty = Q_2^TBv. \tag{12}$$

Next, decompose

$$Q_2^TBP = (0, S), \quad P = (P_1, P_2) \text{ orthogonal}, \tag{13}$$

so that S has full column rank. Row pivoting can be used, as can the singular value decomposition. The solution to (12) is then

$$\hat{v} = P_2\hat{u}, \quad \text{where } S\hat{u} = Q_2^Ty. \tag{14}$$

Since S has full column rank, $\hat{u}$ is unique if the set of equations is consistent, and so $\hat{v}$ is unique. These equations will be consistent if the constraints can be satisfied in (8), that is, if the original linear model is correct. Thus, a check can be provided on the correctness of the model. If Q_2^TB has full row rank, then S in (13) can be made triangular, in this case solve (14) for $\hat{u}$. Otherwise, decompose

$$\tilde{Q}^TS = \begin{pmatrix} \tilde{Q}_1^TS \\ \tilde{Q}_2^TS \end{pmatrix} = \begin{pmatrix} \tilde{S} \\ 0 \end{pmatrix}, \quad \tilde{Q} = (\tilde{Q}_1, \tilde{Q}_2) \text{ orthogonal}, \tag{15}$$

so that $\tilde{S}$ is triangular. Again column pivoting can be used, while if the singular value decomposition was used in (13), $\tilde{Q}_1^T$ and $\tilde{S}$ (diagonal) would already be available. Then solve

$$\tilde{S}\hat{u} = \tilde{Q}_1^TQ_2^Ty. \tag{16}$$

Now x rather than v is wanted, so (10) becomes

$$\begin{aligned} Rx &= Q_1^Ty - Q_1^TBP_2\hat{u} = b \quad \text{say}, \\ &= Q_1^T(I - BP_2\tilde{S}^{-1}\tilde{Q}_1^TQ_2^T)y = Gy \quad \text{say}. \end{aligned} \tag{17}$$

If R is square, solve for the unique solution $\hat{x}$. Otherwise, there will be many x satisfying (17); and we will usually want the solution $\hat{x}$ with minimum 2-norm, so decompose

$$R\tilde{P} = (R\tilde{P}_1, R\tilde{P}_2) = (0, \tilde{R}), \quad \tilde{P} \text{ orthogonal}, \tag{18}$$

so that $\tilde{R}$ is triangular. Row pivoting can be used, while if the singular value decompo-

sition was used in (9), $\tilde{P}_2$ and $\tilde{R}$ (diagonal) would already be available. Then

$$\hat{x} = \tilde{P}_2 z, \quad \text{where } \tilde{R}z = b, \tag{19}$$

and z, and so $\hat{x}$, can easily be computed.

The matrix $\tilde{P}_2$ can effectively be kept and used in (19). The matrix Q in (9) could be kept for forming b in (17), or it need not be kept if $Q^T B$ is formed at the same time as $Q^T C$ and $Q^T y$ are formed. Again, P could be kept for forming b, if only $Q_2^T B$ is transformed in (13). P need not be kept if

$$Q^T BP = \begin{pmatrix} Q_1^T BP_1 & Q_1^T BP_2 \\ 0 & S \end{pmatrix} = \begin{pmatrix} L^T & F \\ 0 & S \end{pmatrix} \quad \text{say} \tag{20}$$

is formed. In this case if B has special form such as upper triangular, it could be advantageous to carry out the P and Q transformations together, making careful use of rotations to maintain the triangular form throughout. This would result in S and L^T being upper triangular. If B has lower triangular form, the computations in (9) and (13) can be rearranged to maintain this form too. This latter form has been programmed in ALGOLW by Stavros Kourouklis [18] for the McGill University IBM 370 system.

Some properties of the problem will be clarified by examining the matrices appearing in this method of solution. From (9) we see that the columns of Q_1 span $\mathcal{R}(C)$ the range of C, so the columns of Q_2 span the orthogonal complement of $\mathcal{R}(C)$; that is, the columns of Q_2 span the space of all vectors q such that $q^T C = 0$. Next from (13) and (15)

$$\tilde{Q}^T Q_2^T B = \begin{pmatrix} \tilde{S} \\ 0 \end{pmatrix} P_2^T, \quad \tilde{S} \text{ nonsingular},$$

so that the columns of $Q_2\tilde{Q}_2$ span the space of all vectors q such that $q^T(C, B) = 0$. This means that the columns of $(Q_1, Q_2\tilde{Q}_1)$ span $\mathcal{R}(C) + \mathcal{R}(B)$. So for the model (7) to be meaningful we need $y \in \mathcal{R}(C) + \mathcal{R}(B)$, that is $\tilde{Q}_2^T Q_2^T y = 0$. If this is so, then from (10) and (11) we see that v will be chosen to account for $\tilde{Q}_1^T Q_2^T y$, and then x will be chosen to account for the rest. In other words, $\|\tilde{Q}_2^T Q_2^T y\|$ will give a measure of how wrong y is for this model; and when y is allowable, $\|\tilde{Q}_1^T Q_2^T y\|$ shows how much of y can only be accounted for by noise.

4. Properties of the Estimator. For a given measurement vector y, $\hat{x}$ in (19) is our least squares estimate of x in (7). Now since v is a random vector in (7), y is a random vector before it is measured; and then $\hat{x}$ which is a linear function of y, see (17), will also be a random vector, called the estimator of x. Here we examine some of the important properties of $\hat{x}$ as an estimator of x.

We have seen that $\hat{x}$ is the minimum 2-norm solution of (17). From (13) and (15)

$$Q_2^T B = SP_2^T = \tilde{Q}_1 \tilde{S} P_2^T, \tag{21}$$

so in (17) $R\hat{x} = Gy$ becomes with (7) and (9)

$$R\hat{x} = G(Cx + Bv) = Rx + Q_1^T B(I - P_2 P_2^T)v;$$

thus with the orthogonality of P and the notation in (20)

$$R(\hat{x} - x) = Q_1^T B P_1 P_1^T v = L^T P_1^T v. \tag{22}$$

From (7) it follows that

$$RE(\hat{x} - x) = 0, \tag{23}$$

$$RHR^T = \sigma^2 L^T L, \quad H \equiv E[(\hat{x} - x)(\hat{x} - x)^T]. \tag{24}$$

If R is square, it is nonsingular, and

$$E(\hat{x} - x) = 0, \quad H = \sigma^2 R^{-1} L^T L R^{-T}, \tag{25}$$

so $\hat{x}$ is an unbiased estimator of x whose covariance matrix H has the right-hand factor $\sigma L R^{-T}$. If R is upper triangular and L is lower triangular, then this factor is lower triangular.

If R which has full row rank is not square, then C in (7) and (9) has less than full column rank. In this case write

$$x = x_c + \bar{x}_c, \quad x_c \perp N(C), \quad \bar{x}_c \in N(C), \tag{26}$$

where $N(C)$ is the null space of C. Then $\hat{x}$ is an unbiased estimator of x_c, with covariance matrix, from (18) and (24)

$$\sigma^2 \tilde{P}_2 \tilde{R}^{-1} L^T L \tilde{R}^{-T} \tilde{P}_2^T. \tag{27}$$

5. Perturbation Analysis. It is important to understand what effect changes in the data will have on our estimate $\hat{x}$. Let our perturbed data be

$$\bar{y} = y + \delta y, \quad \bar{C} = C + \delta C, \quad \bar{B} = B + \delta B \tag{28}$$

leading to the solution $\hat{v} + \delta v$, $\hat{x} + \delta x$ of the perturbed problem (8). Considering (8) for both the original and perturbed problems, we see that δv and δx give

$$\underset{\delta x, \delta v}{\text{minimum}} (2\hat{v}^T \delta v + \delta v^T \delta v), \tag{29}$$

$$\text{subject to } \bar{C}\delta x + \bar{B}\delta v = \delta \bar{y}, \tag{30}$$

$$\delta \bar{y} \equiv \delta y - \delta C \hat{x} - \delta B \hat{v}. \tag{31}$$

The constraints (30) have the same form as in (8), so we can proceed as in (9) and (20)

$$\bar{Q}^T \bar{C} = \begin{pmatrix} \bar{R} \\ 0 \end{pmatrix}, \quad \bar{Q}^T \bar{B} \bar{P} = \begin{pmatrix} \bar{L}^T & \bar{F} \\ 0 & \bar{S} \end{pmatrix}, \tag{32}$$

where $\bar{Q} = (\bar{Q}_1, \bar{Q}_2)$, $\bar{P} = (\bar{P}_1, \bar{P}_2)$ are orthogonal, and $\bar{R}$ and $\bar{S}^T$ have full row rank.

In the following, superscript + will denote the pseudo-inverse, and $\sigma(\cdot)$ will denote the smallest nonzero singular value, so for example in (15)

$$S^+ = \tilde{S}^{-1} \tilde{Q}_1^T, \quad \|S^+\| = 1/\sigma(S), \tag{33}$$

where $\|\cdot\|$ will always denote the 2-norm.

Combining (30) and (32) shows that the constraints on δv are

$$\bar{Q}_2^T \delta \bar{y} = \bar{Q}_2^T \bar{B} \delta v = \bar{S} \bar{P}_2^T \delta v, \tag{34}$$

and this must be a consistent system for the perturbation to be meaningful for this problem. We can then express

$$\delta v = \bar{P}_1 z_1 + \bar{P}_2 z_2, \quad z_2 = \bar{S}^+ \bar{Q}_2^T \delta \bar{y},$$

for all z_1. Substituting this in (29) and taking the derivative with respect to z_1 gives

$$z_1 = -\overline{P}_1^T \hat{v}, \quad \delta v = \overline{P}_2 \overline{S}^+ \overline{Q}_2^T \delta \overline{y} - \overline{P}_1 \overline{P}_1^T \hat{v}. \tag{35}$$

The first term in (35) can easily be bounded, but the second is more difficult. From (14)

$$\overline{P}_1^T \hat{v} = \overline{P}_1^T P_2 S^+ Q_2^T y = \overline{P}_1^T P_2 P_2^T \hat{v}, \tag{36}$$

and we will seek an expression for $\overline{P}_1^T P_2$. To do this we first consider $Q_2^T \overline{Q}_1$. Combining (9) and (32), and realizing that $\overline{R}$ has full row rank,

$$Q_2^T \overline{C} = Q_2^T (C + \delta C) = Q_2^T \delta C = Q_2^T \overline{Q}_1 \overline{R}, \quad Q_2^T \overline{Q}_1 = Q_2^T \delta C \overline{R}^+, \tag{37}$$

$$\|Q_2^T \overline{Q}_1\| \leqslant \epsilon_c / \sigma(\overline{C}), \quad \epsilon_c \equiv \|\delta C\|, \tag{38}$$

where $\sigma(\overline{C}) = \sigma(\overline{R})$ is the smallest nonzero singular value of $\overline{C} = C + \delta C$. Note that if C and $\overline{C}$ have the same rank, then (see for example [19, p. 321])

$$|\sigma(C) - \sigma(\overline{C})| \leqslant \epsilon_c. \tag{39}$$

Continuing our search for $\overline{P}_1^T P_2$, we see from (32)

$$(B + \delta B)\overline{P} = \overline{Q} \begin{pmatrix} \overline{L}^T & \overline{F} \\ 0 & \overline{S} \end{pmatrix}, \tag{40}$$

so from the first set of colunns of this

$$Q_2^T \overline{Q}_1 \overline{L}^T - Q_2^T \delta B \overline{P}_1 = Q_2^T B \overline{P}_1 = S P_2^T \overline{P}_1$$

from (21). This combines with (37) to give

$$P_2^T \overline{P}_1 = S^+ (Q_2^T \delta C \overline{R}^+ \overline{L}^T - Q_2^T \delta B \overline{P}_1), \tag{41}$$

$$\|P_2^T \overline{P}_1\| \leqslant [\epsilon_B + \|\overline{L}\| \epsilon_c / \sigma(\overline{C})] / \sigma(S), \quad \epsilon_B \equiv \|\delta B\|. \tag{42}$$

This can now be used with (35) and (36) to give an expression for δv which can be bounded

$$\delta v = \overline{P}_2 \overline{S}^+ \overline{Q}_2^T \delta \overline{y} + \overline{P}_1 [\overline{P}_1^T \delta B^T Q_2 - \overline{L} (\overline{R}^+)^T \delta C^T Q_2] (S^+)^T P_2^T \hat{v}. \tag{43}$$

We now obtain an expression for $\delta \hat{x}$, the smallest δx in (29) and (30). From (32)

$$\overline{R} \delta x = \overline{Q}_1^T (\delta \overline{y} - \overline{B} \delta v),$$

and combining this with (43) and (32) gives

$$\delta \hat{x} = \overline{R}^+ [\overline{Q}_1^T - \overline{F} \overline{S}^+ \overline{Q}_2^T] \delta \overline{y} - \overline{R}^+ \overline{L}^T [\overline{P}_1^T \delta B^T Q_2 - \overline{L} (\overline{R}^+)^T \delta C^T Q_2] (S^+)^T P_2^T \hat{v}. \tag{44}$$

Note that for large $\hat{v}$ this second term can be quite large, this is the equivalent of the possibly large residual term in the analyses in [3], [9], [12], [13], [19], and [20], for the ordinary least squares problem. If here $B = \overline{B} = I$, then $\delta B = 0$, $P = Q$, $\overline{P} = \overline{Q}$, $L = \overline{L} = I$, $S = \overline{S} = I$, $F = \overline{F} = 0$ and

$$\delta \hat{x} = \overline{R}^+ \overline{Q}_1^T \delta \overline{y} + \overline{R}^+ (\overline{R}^+)^T \delta C^T \hat{v}, \tag{45}$$

which corresponds to the results in those references. We also note that if the columns of δB and δC lie in the range of C, then the second term in (44) is zero. A simple bound for (44) is

$$\|\delta x\| \leqslant \{[1 + \|\bar{F}\|/\sigma(\bar{S})]\|\delta\bar{y}\| + [\epsilon_B + \|\bar{L}\|\epsilon_c/\sigma(\bar{C})]\|\bar{L}\| \cdot \|\hat{v}\|/\sigma(S)\}/\sigma(\bar{C}), \tag{46}$$

where $\delta\bar{y}$ is as in (31) and $\sigma(\bar{C})$ satisfies (39) if C and $\bar{C}$ have the same rank. We also have

$$\|\bar{L}\|, \|\bar{F}\| \leqslant \|B + \delta B\| \leqslant \|B\| + \epsilon_B. \tag{47}$$

For (46) to give a rigorous a priori bound, we see we will have to obtain a lower bound for $\sigma(\bar{S})$. Now

$$\sigma(S) = \sigma(Q_2^T B), \quad \sigma(\bar{S}) = \sigma(\bar{Q}_2^T \bar{B}) \tag{48}$$

and to relate these we note that

$$\bar{Q}_2^T \bar{B} = \bar{Q}_2^T Q Q^T B + \bar{Q}_2^T \delta B = \bar{Q}_2^T Q_1 Q_1^T B + (0, I)\bar{Q}^T Q_2 Q_2^T B + \bar{Q}_2^T \delta B, \tag{49}$$

$$\begin{pmatrix} 0 \\ \bar{Q}_2^T \bar{B} \end{pmatrix} - \bar{Q}^T Q_2 Q_2^T B = \begin{pmatrix} -\bar{Q}_1^T Q_2 Q_2^T B \\ \bar{Q}_2^T Q_1 Q_1^T B \end{pmatrix} + \begin{pmatrix} 0 \\ \bar{Q}_2^T \delta B \end{pmatrix}$$

$$= \begin{pmatrix} 0 & -\bar{Q}_1^T Q_2 \\ \bar{Q}_2^T Q_1 & 0 \end{pmatrix} Q^T B + \begin{pmatrix} 0 \\ \bar{Q}_2^T \delta B \end{pmatrix}. \tag{50}$$

But from (9) and (32)

$$\bar{Q}_2^T C = \bar{Q}_2^T Q_1 R = \bar{Q}_2^T(\bar{C} - \delta C) = -\bar{Q}_2^T \delta C, \tag{51}$$

$$\bar{Q}_2^T Q_1 = -\bar{Q}_2^T \delta C R^+, \tag{52}$$

and we will assume $\epsilon_c = \|\delta C\|$ is sufficiently small so that

$$\|\bar{Q}_2^T Q_1\| \leqslant \epsilon_c/\sigma(C) < 1. \tag{53}$$

If also C and $\bar{C}$ have the same rank, then (38) and (39) give

$$\|\bar{Q}_1^T Q_2\| \leqslant \epsilon_c/(\sigma(C) - \epsilon_c), \tag{54}$$

so if S and $\bar{S}$ have the same rank, (48) and (50) give

$$|\sigma(S) - \sigma(\bar{S})| \leqslant \epsilon_B + \epsilon_c\|B\|/(\sigma(C) - \epsilon_c). \tag{55}$$

Happily we note that if the perturbations are sufficiently small, and C and $\bar{C}$ have the same rank, while S and $\bar{S}$ do too, then $\sigma(\bar{C})$ and $\sigma(\bar{S})$ in (46) can effectively be replaced by $\sigma(C)$ and $\sigma(S)$, respectively, where $\sigma(S) = \sigma(Q_2^T B)$, and the columns of Q_2 form an orthogonal basis for the null space of C^T. In this case it is the smallest nonzero singular value of C and that of the projection of B onto the null space of C^T, which determine the condition of the problem.

If a tighter bound than (46) is needed, we can write

$$Q^T \delta C = \begin{pmatrix} \delta C_1 \\ \delta C_2 \end{pmatrix}, \quad Q^T \delta B P = \begin{pmatrix} \delta B_{11} & \delta B_{12} \\ \delta B_{21} & \delta B_{22} \end{pmatrix}, \quad Q^T \delta y = \begin{pmatrix} \delta y_1 \\ \delta y_2 \end{pmatrix}, \tag{56}$$

and use (41) to give

$$Q_2^T \delta B \bar{P}_1 = \delta B_{21} P_1^T \bar{P}_1 + \delta B_{22} S^+(\delta C_2 \bar{R}^+ \bar{L}^T - Q_2^T \delta B \bar{P}_1), \tag{57}$$

$$\|Q_2^T \delta B \overline{P}_1\|_2 \leqslant \frac{\|\delta B_{21}\| + \|\overline{L}\| \cdot \|\delta C_2\| \cdot \|\delta B_{22}\| (\sigma(S)\sigma(\overline{C}))^{-1}}{1 - \|\delta B_{22}\| (\sigma(S))^{-1}}, \tag{58}$$

where we assume the denominator is positive. For small errors this bound is nearly $\|\delta B_{21}\|$. These results can be used to bound the second term on the right-hand side of (44). For the first term we use

$$\overline{Q}_j^T = (\overline{Q}_j^T Q_1, \overline{Q}_j^T Q_2) Q^T, \qquad j = 1, 2, \tag{59}$$

with (31), (56), (14), (37) and (52) to give

$$\begin{aligned} \overline{Q}_1^T \delta \overline{y} &= \overline{Q}_1^T Q_1 [\delta y_1 - \delta C_1 \hat{x} - \delta B_{12} P_2^T \hat{v}] \\ &\quad + (\overline{R}^+)^T \delta C^T Q_2 [\delta y_2 - \delta C_2 \hat{x} - \delta B_{22} P_2^T \hat{v}], \end{aligned} \tag{60}$$

$$\begin{aligned} \overline{Q}_2^T \delta \overline{y} &= -\overline{Q}_2^T \delta C R^+ [\delta y_1 - \delta C_1 \hat{x} - \delta B_{12} P_2^T \hat{v}] \\ &\quad + \overline{Q}_2^T Q_2 [\delta y_2 - \delta C_2 \hat{x} - \delta B_{22} P_2^T \hat{v}]. \end{aligned} \tag{61}$$

From the expressions we have derived we could produce a correct, fairly tight, but extremely messy a priori bound on $\|\delta \hat{x}\|$. Instead we will assume the perturbations are small enough to ignore some of the second order error terms. Thus, if C and $\overline{C}$ have the same rank, and S and $\overline{S}$ have the same rank, we effectively have from (44), (60), (61), and (58)

$$\begin{aligned} \|\delta \hat{x}\| \lesssim \frac{1}{\sigma(C)} \Bigg[& \left(1 + \frac{\|F\| \cdot \|\delta C_2\|}{\sigma(S)\sigma(C)}\right) (\|\delta y_1\| + \|\delta C_1\| \cdot \|\hat{x}\| + \|\delta B_{12}\| \cdot \|\hat{v}\|) \\ & + \left(\frac{\|F\|}{\sigma(S)} + \frac{\|\delta C_2\|}{\sigma(C)}\right) (\|\delta y_2\| + \|\delta C_2\| \cdot \|\hat{x}\| + \|\delta B_{22}\| \cdot \|\hat{v}\|) \\ & + \frac{\|L\|}{\sigma(S)} \left(\|\delta B_{21}\| + \frac{\|\delta C_2\|}{\sigma(C)} \|L\|\right) \|\hat{v}\| \Bigg]. \end{aligned} \tag{62}$$

This separates the perturbation effect into its main components. For example, we see that the term $Q_1^T \delta B P_1$ has no effect on $\delta \hat{x}$. Next we note that if all the perturbations are orthogonal to Q_2, that is if they all lie in the range of C, the condition of the problem is proportional to $1/\sigma(C)$. If $Q_2^T \delta C = 0$, then the condition is proportional to $1/(\sigma(C)\sigma(S))$. In the worst case the condition depends on $1/(\sigma(C)^2 \cdot \sigma(S))$.

This perturbation analysis can be used to see what effect errors or changes in the original data will have on the solution. It can also be used along with the backward error analysis in the next section to understand what effect rounding errors can have on $\hat{x}$.

6. Rounding Error Analysis. The analysis of the most likely computation will be given. This will occur when orthogonal-triangular decompositions are used in (9) and

(13) rather than singular value decompositions, and when R in (9) and S in (13) are both square. This last will always occur when C has full column rank and B is square in (7) and (8). When no rounding errors are present, the resulting computations in Section 3 can be summarized by

$$Q^T(C, B, y)\begin{pmatrix} I & P & \\ & & 1 \end{pmatrix} = \begin{pmatrix} R, L^T, F, y_1 \\ 0, 0, S, y_2 \end{pmatrix}, \quad y_i \equiv Q_i^T y, \tag{63}$$

$$\begin{pmatrix} R & F \\ 0 & S \end{pmatrix}\begin{pmatrix} \hat{x} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \tag{64}$$

For this case well-known backward rounding error analysis results will be used to show that the computation is numerically stable. By this is meant that the computed solution will be exact for a problem of the same form as (8) but with slightly different initial data. It will simplify the presentation if multiplicative terms involving the dimensions of the problem are omitted from the error bounds, since the exact terms are not needed for proving stability, and can be found for any particular computation in [21] and elsewhere in the literature. Results of basic rounding error analyses will be quoted from the work of Wilkinson (see for example [21]) without further reference, and the symbols ϵ_i will indicate nonnegative quantities which are bounded above by the product of ϵ, the floating point arithmetic computer precision, and constants dependent only on the dimensions of the problem. Terms of the form $\epsilon_i\epsilon_j$ will be ignored.

When a matrix A is transformed by stable orthogonal transformations or fast stable square root free rotations (see, for example, [22]) so that ideally $\tilde{Q}^T A = \tilde{G}$, $\tilde{Q}$ orthogonal, it can be shown that the computed matrix G satisfies

$$\bar{Q}^T(A + E_1) = G, \quad \bar{Q} \text{ orthogonal}, \quad \|E_1\| = \epsilon_1\|A\|, \tag{65}$$

and this holds even when $\tilde{Q}$ is directly chosen to make part of $\tilde{G}$ zero, and the corresponding part of G is set to zero. If G is then transformed from the right, the computed result N satisfies

$$(G + E_2)\bar{P} = N, \quad \bar{P} \text{ orthogonal}, \quad \|E_2\| = \epsilon_2\|A\|.$$

These combine to give

$$\bar{Q}^T(A + E_3)\bar{P} = N, \quad \|E_3\| = \|E_1 + \bar{Q}E_2\| = \epsilon_3\|A\|, \tag{66}$$

and further left and right transformations can be applied in any order to give the same form of result.

Thus, when rounding errors are present, (63) becomes

$$\bar{Q}^T(C + E_1, B + E_2, y + \delta y)\begin{pmatrix} I & \bar{P} & \\ & & 1 \end{pmatrix} = \begin{pmatrix} R, L^T, F, y_1 \\ 0, 0, S, y_2 \end{pmatrix},$$

$$\|E_1\| = \epsilon_1\|C\|, \quad \|E_2\| = \epsilon_2\|B\|, \quad \|\delta y\| = \epsilon_0\|y\|, \tag{67}$$

where $\bar{Q}$ and $\bar{P}$ are orthogonal, and R etc. are the computed results. We see that this holds whether $\bar{Q}$ is applied first and then followed by $\bar{P}$, or the two transformations

are "interleaved" as in [18]. In the presence of rounding errors it can be shown that the computed solution $\hat{x}$, $\hat{u}$ of (64) satisfies

$$\begin{pmatrix} R + E_3, & F + E_4 \\ 0, & S + E_5 \end{pmatrix} \begin{pmatrix} \hat{x} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \tag{68}$$

$$\|E_3\| = \epsilon_3 \|C\|, \quad \left\| \begin{matrix} E_4 \\ E_5 \end{matrix} \right\| = \epsilon_4 \|B\overline{P}_2\|.$$

Combining (67) and (68) shows that $\hat{x}$ and $\hat{u}$ give the exact solution of (8) for the initial data

$$\begin{cases} \overline{C} = C + E_1 + \overline{Q}_1 E_3, & \|\overline{C} - C\| \leqslant (\epsilon_1 + \epsilon_3)\|C\|, \\ \overline{B} = B + E_2 + \overline{Q}_1 E_4 \overline{P}_2^T + \overline{Q}_2 E_5 \overline{P}_2^T, & \|\overline{B} - B\| \leqslant (\epsilon_2 + \epsilon_4)\|B\|, \\ \overline{y} = y + \delta y, & \|\overline{y} - y\| \leqslant \epsilon_0 \|y\|, \end{cases} \tag{69}$$

that is, the computed solution is exact for a nearby problem. Note that the computed R and L are exact for the data $C + E_1$, $\overline{B}$; and that R differs by E_3 from the true matrix corresponding to (69), while L is correct. As a result, the algorithm is numerically stable for computing the estimate $\hat{x}$ and is also numerically stable for computing full available information R and L on the covariance matrix in (24). Since the triple $\hat{x}$, R, L is not exact for (69), but is very close to the exact solution of (69), the statement that the algorithm is numerically stable for this triple can be interpreted in the wide sense described by Stewart [19, p. 76].

7. **Comments.** The rounding error analysis does not say that $R^{-1}L^T$ in (25) can be computed in a numerically stable way, and so it has not been shown that the algorithm is stable for computing the covariance matrix factor. However, it is the author's opinion that (24) gives a more general and useful representation of the covariance matrix than the more standard (25) or (27), and the algorithm is numerically stable for computing R and L in (24). In this wider sense the algorithm can be said to be stable for finding the covariance matrix representation.

The cases where either R in (9) or S in (13) is square and nearly singular, or not square, can lead to difficulties in deciding just what ranks are meaningful for any given problem. This has been examined in [14] for the ordinary linear least squares problem, and needs further work here.

Finally, it is the author's opinion that the formulation (8) of the problem contributes greatly to its solution and analysis, as well as generalizing the problem. For example, if the formulation (8) is used for the ordinary least squares case of $B = I$, then it seems to lead to a slightly easier perturbation analysis, and in the general case (8) makes the perturbation analysis definitely more tractable . Also, since $y = Cx + Bv$ is now just a set of constraints, it is clear that the transformation in (9) can be carried out by any well-conditioned nonsingular matrix Q. Then it is clear that the algorithm can often be speeded up by using stabilized nonunitary transformations [11], [21].

Again it is not necessary to use orthogonal transformations to solve (12), see [11], although orthogonal transformations are elegant for this problem. Cline [5] has used stabilized nonunitary transformations for the ordinary linear least squares problem, thereby saving computations in some situations, with no loss in accuracy. This type of approach is also especially useful for problems of the form (8) when the matrices have special form such as being large and sparse with structure. Paige [23] has suggested some fast algorithms for the present problem, especially in structured cases. Numerically stable updating techniques are given in [24].

Acknowledgments. The author is grateful to Dr. Michael Saunders for thoughtful and stimulating discussions on this material at the Operations Research Department, Stanford University. Stavros Kourouklis programmed and tested the algorithm with exemplary speed and clarity, and contributed a lot to the understanding of the problem and algorithms. Gene Golub made some helpful comments on the manuscript, and he and Åke Björck pointed out reference [1] to the author. The referee gave a very helpful and thorough report, which led to several improvements in the manuscript.

School of Computer Science
McGill University
Montreal, Quebec, Canada H3A 2K6

1. Å. BJÖRCK, *A Uniform Numerical Method for Linear Estimation from General Gauss-Markoff Models*, Proc. 1st. Sympos. on Computational Statistics (COMPSTAT), Vienna, 1974, pp. 131–140.

2. Å. BJÖRCK, "Solving linear least squares problems by Gram-Schmidt orthogonalization," *BIT*, v. 7, 1967, pp. 1–21.

3. Å. BJÖRCK, "Iterative refinement of linear least squares solution. I, II," *BIT*, v. 7, 1967, pp. 251–278; *BIT*, v. 8, 1968, pp. 8–30.

4. P. BUSINGER & G. GOLUB, "Linear least squares solutions by Householder transformations," *Numer. Math.*, v. 7, 1965, pp. 269–276.

5. A. K. CLINE, "An elimination method for the solution of linear least squares problems," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 283–289.

6. G. GOLUB, "Numerical methods for solving linear least squares problems," *Numer. Math.*, v. 7, 1965, pp. 206–216.

7. G. H. GOLUB & C. REINSCH, "Singular value decomposition and least squares solutions," *Numer. Math.*, v. 14, 1970, pp. 403–420.

8. GENE H. GOLUB & GEORGE P. H. STYAN, "Numerical computations for univariate linear models," *J. Statist. Comp. and Simulation*, v. 2, 1973, pp. 253–274.

9. G. H. GOLUB & J. H. WILKINSON, "Note on the iterative refinement of least squares solutions," *Numer. Math.*, v. 9, 1966, pp. 139–148.

10. C. L. LAWSON & R. J. HANSON, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, N. J., 1974.

11. G. PETERS & J. H. WILKINSON, "The least squares problem and pseudo-inverses," *Comput. J.*, v. 13, 1970, pp. 309–316.

12. G. W. STEWART, "On the continuity of the generalized inverse," *SIAM J. Appl. Math.*, v. 17, 1969, pp. 33–45.

13. G. W. STEWART, "On the perturbation of pseudo-inverses, projections, and linear least squares problems," *SIAM Rev.*, v. 19, 1977, pp. 634–662.

14. G. GOLUB, V. KLEMA, & G. W. STEWART, *Rank Degeneracy and Least Squares Problems*, Stanford University Computer Science Report STAN-CS-76-559, August, 1976.

15. C. R. RAO, *Linear Statistical Inference and its Applications*, Chapter 4, 2nd ed., Wiley, New York, 1973.

16. G. A. F. SEBER, *Linear Regression Analysis*, Wiley, New York, 1977.

17. C. C. PAIGE & M. A. SAUNDERS, "Least squares estimation of discrete linear dynamic systems using orthogonal transformations," *SIAM J. Numer. Anal.*, v. 14, 1977, pp. 180–193.

18. S. KOUROUKLIS, *Computing Weighted Linear Least Squares Solutions*, McGill University School of Computer Science, M.Sc. Project, May 1977.

19. G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

20. A. VAN DER SLUIS, "Stability of the solutions of linear least squares problems," *Numer. Math.*, v. 23, 1975, pp. 241–254.

21. J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford, 1965.

22. S. HAMMARLING, "A note on modifications to the Givens plane rotation," *J. Inst. Math. Appl.*, v. 13, 1974, pp. 215–218.

23. C. C. PAIGE, "Fast numerically stable computations for generalized linear least squares problems," *SIAM J. Numer. Anal.* (To appear.)

24. C. C. PAIGE, "Numerically stable computations for general univariate linear models," *Comm. Statist. Ser. B*, v. B7, No. 5, 1978.