A Generalized SVD Analysis of Some Weighting Methods for Equality Constrained Least Squares

Charles Van Loan
Department of Computer Science
Cornell University
Ithaca, New York, 14853, USA

Abstract

The method of weighting is a useful way to solve least squares problems that have linear equality constraints. New error bounds for the method are derived using the generalized singular value decomposition. The analysis clarifies when the weighting approach is succesful and suggests modifications when it is not.

1. Introduction

The problem we consider is how to find a vector $x \in R^n$ that solves the equality constrained problem

$$\text{(LSE)} \qquad \min_{Bx=d} ||Ax - b||_2$$

where $A \in R^{mxn}(m \geq n)$, $b \in R^m$, $B \in R^{pxn}(p \leq n)$ and $d \in R^p$. We will assume that rank(B) = p and that the nullspaces of the two matrices satisfy $N(A) \cap N(B) = \{0\}$. These conditions ensure that (LSE) has a unigue solution which we designate by x_{LSE}.

Important settings where this problem arises include constrained surface fitting, penalty function methods in nonlinear optimization, and geodetic least squares adjustment.

Several methods for solving the LSE problem are discussed in Lawson

and Hanson [7, Chapters 20-22].In one approach Q-R factorizations are used to compute the projections of $\mathbf{x}_{LSE}$ onto $N(B)^{\perp}$ and $N(B)$:

$$
\begin{aligned}
&\text{(a)} \quad B^T = [Q_1, Q_2] \begin{bmatrix} R_B \\ 0 \end{bmatrix} \begin{matrix} p \\ n-p \end{matrix} \qquad \text{(Q-R)} \\
&\text{(b)} \quad R_B^T y_1 = d \;;\; x_1 := Q_1 y_1 \\
&\text{(c)} \quad AQ_2 = [U_1, U_2] \begin{bmatrix} R_A \\ 0 \end{bmatrix} \begin{matrix} n-p \\ m-n+p \end{matrix} \qquad \text{(Q-R)} \\
&\text{(d)} \quad R_A y_2 = U_1^T (b - Ax_1) \;\;;\;\; x_{LSE} := x_1 + Q_2 y_2
\end{aligned}
\tag{1.1}
$$

This algorithm is easy to implement using the LINPACK routines. (It is a MATLAB "5-liner".)

Unfortunately, (1.1) is not a viable method for solving the large sparse LSE problem because the matrix AQ_2 will generally be dense. In this context the method of weighting is of interest. The idea behind this approach is simply to compute the solution $x(\mu)$ to the unconstrained problem

$$
\min_{x \in R^n} \left\| \begin{pmatrix} \mu B \\ A \end{pmatrix} x - \begin{pmatrix} \mu d \\ b \end{pmatrix} \right\|_2 \tag{1.2}
$$

for a large value of $\mu \in R$. It is widely known that $x(\mu) \to x_{LSE}$ as $\mu \to \infty$. Thus, existing software for sparse LS problems can "in principal" be used to generate an approximation to x_{LSE} of arbitrary quality.

However, several issues associated with the method of weighting demand our attention. At what rate do the quantities $x(\mu)$ and $d - Bx(\mu)$ converge? Are there practical ways to estimate the accuracy of $x(\mu)$? How can we cope with the numerical problems that can be expected to arise when μ is extremely large? We are prompted to discuss these issues because of the increasing importance of the sparse LSE problem. But our analytic and algorithmic developments will also be of interest to solvers of small LSE problems since the simplicity of the weighting method makes it extremely attractive and popular.

Our discussion is structured as follows. First, we analyze the properties of x_{LSE} and $x(\mu)$ using the generalized singular value decomposition. The limitations of the theory are then made obvious by reviewing the numerical difficulties associated with large μ. Next, we propose two techniques that can be used both to improve $x(\mu)$ and to estimate its error. One technique involves extrapolation and the other iterative improvement. We conclude with some remarks about the practical implementation of our ideas.

2. A GSVD Analysis of $x(\mu)$

The generalized singular value decomposition (GSVD) of A and B is useful as a tool for analyzing the method of weighting. This decomposition is as follows:

Theorem 2.1

If $A \in R^{m \times n}$ $(m \geq n)$ and $B \in R^{p \times n}$ $(p \leq n)$ satisfy $N(A) \cap N(B) = \{0\}$ then there exist

$$U = [u_1, \ldots, u_m] \in R^{m \times m} \quad \text{(orthogonal)}$$

$$V = [v_1, \ldots, v_p] \in R^{p \times p} \quad \text{(orthogonal)}$$

$$X = [x_1, \ldots, x_n] \in R^{n \times n} \quad \text{(nonsingular)}$$

such that

$$U^T A X = D_A = \operatorname{diag}(\alpha_1, \ldots, \alpha_n)$$

and

$$V^T B X = D_B = \operatorname{diag}(\beta_1, \ldots, \beta_p).$$

Without loss of generality we may assume

$$\|X\|_2 = 1 \quad \text{and} \quad \|X^{-1}\|_2 = \sigma_1/\sigma_n$$

where σ_1 and σ_n are the largest and smallest singular values of $\begin{bmatrix} A \\ B \end{bmatrix}$.

Proof

Let

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \text{ diag } (\sigma_i) Z^T$$

be the SVD of $\begin{bmatrix} A \\ B \end{bmatrix}$ with $Q_1^T Q_1 + Q_2^T Q_2 = I_n$, $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$, and $Z^T Z = I_n$.

Let

$$\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} C \\ S \end{bmatrix} W^T$$

be the C-S decomposition of $\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}$ where $U \in R^{mxm}$, $V \in R^{pxp}$, and $W \in R^{nxn}$ are orthogonal and

$$C = \text{diag}(c_1, \ldots, c_n) \in R^{mxn} \qquad c_i \geq 0$$

and

$$S = \text{diag}(s_1, \ldots, s_p) \in R^{pxn} \qquad 0 \leq s_1 \leq \ldots \leq s_p$$

satisfy $C^T C + S^T S = I_n$. This decomposition is discussed in Stewart [9] who also presents an effective algorithm for computing it in [10].

The theorem follows by setting $D_A = \sigma_n C$, $D_B = \sigma_n S$, and $X^{-1} = \frac{1}{\sigma_n} W^T \text{diag}(\sigma_i) Z^T$. Note that $\sigma_n > 0$ because $N(A) \cap N(B) = N(\begin{bmatrix} A \\ B \end{bmatrix}) = \{0\}$ ∎.

A number of elementary consequences of the GSVD are repeatedly used in the sequel. These are summarized in the following result:

Corollary 2.2

Suppose the GSVD is computed as indicated by the proof of Theorem 2.1. If rank(B) = p and

$$\mu_i = \alpha_i / \beta_i \qquad i = 1, \ldots, p$$

then

(a) $\alpha_i^2 + \beta_i^2 = \sigma_n^2 \qquad i = 1,\ldots,p$

(b) $\alpha_1 \geq \ldots \geq \alpha_q > \alpha_{q+1} = \ldots = \alpha_p = 0$ where $q = \dim[N(A)^{\perp} \cap N(B)^{\perp}]$

(c) $\alpha_{p+1} = \ldots = \alpha_n = \sigma_n$

(d) $0 < \beta_1 \leq \ldots \leq \beta_p$

(e) $\mu_1 \geq \ldots \geq \mu_p \geq 0$

(f) $Ax_i = \alpha_i u_i \qquad i = 1,..,n$

(g) $Bx_i = \beta_i v_i \qquad i = 1,\ldots,p.$

Proof

Contentions (a) and (c) follow from $D_A^T D_A + D_B^T D_B = \sigma_n^2 I_n$ while (d) and (e) are true because $s_1 \leq \ldots \leq s_p$ and $p = \text{rank}(B)$. The equations $AX = UD_A$ and $BX = VD_B$ establish (f) and (g). Finally, if $\alpha_q > \alpha_{q+1} = \ldots = \alpha_p = 0$ then it follows from $A^T U = X^{-T} D_A^T$ and $B^T V = X^{-T} D_B^T$ that the first q columns of X^{-T} span the subspace $\text{Ran}(A^T) \cap \text{Ran}(B^T) = N(A)^{\perp} \cap N(B)^{\perp}$. This proves (b). □

The μ_i are called the generalized singular values of (A,B). We mention that a more general version of the GSVD is given in [11].

The GSVD can be used to diagonalize the LSE problem. In particular, by setting

$$\begin{aligned} \tilde{b} &= U^T b = (u_1^T b,\ldots,u_m^T b)^T \\ \tilde{d} &= V^T d = (v_1^T d,\ldots,v_p^T d)^T \\ y &= X^{-1} x = (y_1,\ldots, y_n)^T \end{aligned} \tag{2.1}$$

we obtain

$$\text{(LSE')} \qquad \min_{D_B y = \tilde{d}} \| D_A y - \tilde{b} \|_2 \quad .$$

It is not hard to show that

$$y_{LSE} = \left(\frac{v_1^T d}{\beta_1}, \ldots, \frac{v_p^T d}{\beta_p}, \frac{u_{p+1}^T b}{\alpha_{p+1}}, \ldots, \frac{u_n^T b}{\alpha_n} \right)^T$$

solves this problem. Since $\alpha_{p+1} = \ldots = \alpha_n = \sigma_n$ we have

$$(2.2) \qquad x_{LSE} = Xy_{LSE} = \sum_{i=1}^{p} \frac{v_i^T d}{\beta_i} x_i + \frac{1}{\sigma_n} \sum_{i=p+1}^{n} (u_i^T b) x_i .$$

If we define

$$(2.4) \qquad \rho_i = u_i^T b - \mu_i v_i^T d \qquad i=1,\ldots,q$$

and

$$(2.5) \qquad r_{LS} = \sum_{i=q+1}^{p} (u_i^T b) u_i + \sum_{i=n+1}^{m} (u_i^T b) u_i$$

then it is easy to show that the constrained minimum residual is given by

$$(2.6) \qquad r_{LSE} = b - Ax_{LSE} = r_{LS} + \sum_{i=1}^{q} \rho_i u_i .$$

Note that $r_{LS} = b - Ax_{LS}$ where

$$x_{LS} = \sum_{i=1}^{q} \frac{u_i^T b}{\alpha_i} x_i + \frac{1}{\sigma_n} \sum_{i=p+1}^{n} (u_i^T b) x_i$$

is the solution of minimum 2-norm to the unconstrained LS problem $\min ||Ax-b||_2$. Observe that

$$(2.7) \qquad \Delta^2 \equiv ||r_{LSE}||_2^2 - ||r_{LS}||_2^2 = \sum_{i=1}^{q} \rho_i^2$$

measures how much the residual increases as a result of the constraints.

It is clear from (2.2) that x_{LSE} is sensitive to perturbation whenever σ_n or $\beta_1 = \sigma_n/\sqrt{1+\mu_1^2}$ is small. Because it "displays" these critical quantities, it is advisable to solve ill-conditioned LSE problems via the GSVD.The sensitivity of the LSE problem is investigated in[5] and [12].

The GSVD is also convenient for analyzing $x(\mu)$. Since this vector solves (1.2) it must satisfy the normal equations

$$(2.8) \qquad (A^TA + \mu^2 B^TB)x(\mu) = A^Tb + \mu^2 B^Td.$$

Under (2.1) this transforms to $(D_A^T D_A + \mu^2 D_B^T D_B)y(\mu) = D_A^T\tilde{b} + \mu^2 D_B^T\tilde{d}$ where $x(\mu) = Xy(\mu)$. It is easy to deduce the solution to this diagonal system and that

$$(2.9) \qquad x(\mu) = \sum_{i=1}^{p} \frac{\alpha_i u_i^T b + \mu^2 \beta_i v_i^T d}{\alpha_i^2 + \mu^2 \beta_i^2} x_i + \frac{1}{\sigma_n} \sum_{i=p+1}^{n} (u_i^T b) x_i .$$

By subtracting (2.2) from this equation we find

$$(2.10) \qquad e(\mu) \equiv x(\mu) - x_{LSE} = \sum_{i=1}^{q} \left(\frac{\mu_i}{\mu_i^2+\mu^2}\right) * \left(\frac{\rho_i}{\beta_i}\right) x_i \quad .$$

Note that the error is confined to span $\{x_1,\ldots,x_q\}$ and that it tends to zero as $\mu \to \infty$. An alternative proof of the latter fact may be found in [7, Chapter 22].

In Section 4 we establish the following inequalities:

$$||x(\mu) - x_{LSE}||_2 \leq \frac{\Delta}{2\mu\sigma_n} \sqrt{1+\mu_1^2}$$

$$||d - Bx(\mu)||_2 \leq \frac{\Delta}{2\mu}$$

$$0 \leq ||r_{LSE}||_2 - ||b - Ax(\mu)||_2 \leq \frac{\Delta\mu_1}{2\mu}$$

These results suggest that a large weight might be required if either μ_1 is large or σ_n is small.

3. Difficulties Associated With a Large Weight

It is widely appreciated that care must be exercised when Householder matrices are used to compute the Q-R factorization of a matrix whose rows vary greatly in norm, e.g., the matrix in (1.2) . Powell and Reid [8] examined this problem in conjunction with the Businger-Golub algorithm [4] and advise incorporation of row interchanges, much as in Gaussian elimination. Specifically, they recommend that the k-th

column be searched and its largest entry pivoted to the (k,k) position before the k-th Householder matrix is applied.

Note that near-domination of the pivot elements will result if the Businger - Golub algorithm is applied to our heavily weighted matrix in (1.2) but <u>not</u> if we apply it to the mathematically equivalent problem

$$(3.1) \qquad \min \left\| \begin{pmatrix} A \\ \mu B \end{pmatrix} x - \begin{pmatrix} b \\ \mu d \end{pmatrix} \right\|_2 \qquad \mu >> 1.$$

To appreciate the difference between the "B-over-A" and the "A-over-B" formulations consider the LSE problem

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad B = (1,-1), \quad d = (2).$$

This problem is well-conditioned and has exact solution $x_{LSE} = \frac{1}{29}(39, -19)^T$. In the following table we record the error $\| x(\mu) - x_{LSE} \|_2$ for both approaches as a function of μ:

Method \ μ	10^1	10^3	10^5	10^7	10^9	10^{11}	10^{13}	10^{15}	10^{17}
B-over-A	10^{-3}	10^{-7}	10^{-11}	10^{-15}	10^{-17}	10^{-17}	10^{-17}	10^{-17}	10^{-17}
A-over-B	10^{-3}	10^{-7}	10^{-11}	10^{-11}	10^{-10}	10^{-7}	10^{-6}	10^{-4}	10^{-2}

These computations were performed using VAX double precision arithmetic (macheps $\approx 10^{-17}$) in the MATLAB environment.

The divergence of the two methods in the vicinity of $\mu = (\text{macheps})^{-1/2}$ is fairly typical, even for ill-conditioned examples. However, for ill-conditioned LSE problems, the "optimum" weight for the B-over-A approach is usually several orders of magnitude greater than $(\text{machep})^{-1/2}$. Thus it is preferable (although not always critical) to solve (1.2) instead of (3.1). In the remainder of this paper, we will always use the B-over-A formulation (1.2).

However, it is not always easy to arrange the rows of the unconstrained problem so that the constraint equations come first. The minimization of fill-in may dictate some other row ordering in the case when A and B are large and sparse.

Column ordering is also important in the method of weighting. At the suggestion of the referee, we considered the example

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \qquad b = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \end{bmatrix} \qquad d = \begin{bmatrix} 7 \\ 4 \end{bmatrix},$$

This example is well conditioned and has exact solution $x_{LSE} = \frac{1}{8}(46,-2,12)^T$. Let $R(\mu) = (r_{ij}(\mu))$ be the 3-by-3 upper triangular matrix obtained by computing the Q-R decomposition of $\begin{pmatrix} \mu B \\ A \end{pmatrix}$. It is clear that $r_{11}(\mu)/r_{22}(\mu)$ becomes very large as $\mu \to \infty$. This occurs because the first two columns of B are dependent. On the other hand, if we apply Q-R with column pivoting to $\begin{pmatrix} \mu B \\ A \end{pmatrix}$ then $r_{11}(\mu)$ and $r_{22}(\mu)$ have the same order of magnitude as $\mu \to \infty$. The decision to column pivot is important as the following table of relative errors in $x(\mu)$ indicates:

Column Pivoting \ μ	10^1	10^3	10^5	10^7	10^9	10^{11}	10^{13}	10^{15}
No	10^{-2}	10^{-7}	10^{-11}	10^{-11}	10^{-9}	10^{-7}	10^{-5}	10^{-3}
Yes	10^{-2}	10^{-7}	10^{-10}	10^{-14}	10^{-16}	10^{-16}	10^{-16}	10^{-16}

We infer that the weighting method can be unstable if (1.2) is solved using the Q-R decomposition without pivoting.

4. An Extrapolation Procedure

As we have seen, the importance of row and column ordering in (1.2) increases with increasing μ. Hence, it is potentially interesting to see how the method might be used with "safe" weights since the

ordering of rows and columns is often inconvenient.

Suppose $x(\mu)$ and $x(\gamma\mu)$ have been computed for some $\mu > 0$ and $\gamma > 1$. Using (2.10) it can be shown that if

$$x^{(1)}(\gamma\mu) = x(\gamma\mu) + \frac{1}{\gamma^2-1}\left[x(\gamma\mu) - x(\mu)\right] \tag{4.1}$$

then

$$x^{(1)}(\gamma\mu) = x_{LSE} + \sum_{i=1}^{q} \frac{\mu_i^2}{\mu_i^2+\mu^2} \cdot \frac{\mu_i}{\mu_i^2+(\gamma\mu)^2} \cdot \frac{\rho_i}{\beta_i} \quad x_i$$

Thus, the error expansion for $x^{(1)}(\gamma\mu)$ is the same as that for $x(\gamma\mu)$ except the coeffient of x_i is diminished by the factor $\mu_i^2/(\mu_i^2 + \mu^2)$.

The calculation (4.1) is merely the first extrapolation in the following Richardson scheme:

```
For j=0,1,...
    x^(0)(γ^j μ) = x(γ^j μ)
    For k=1,2,...,j
        x^(k)(γ^j μ) = x^(k-1)(γ^j μ)
          + 1/(γ^2k - 1) * [x^(k-1)(γ^j μ) - x^(k-1)(γ^(j-1) μ)]
```

(4.2)

A messy but straightforward induction argument based on (2.10) can be used to show that

$$e^{(k)}(\gamma^j\mu) \equiv x^{(k)}(\gamma^j\mu) - x_{LSE} = \sum_{i=1}^{q} \varepsilon_i^{(k)}(j)\, x_i \tag{4.3}$$

where

$$\varepsilon_i^{(k)}(j) = \frac{\rho_i}{\beta_i} * \frac{\mu_i}{\mu_i^2+(\mu\gamma)^2} * \prod_{\ell=j-k}^{j-1} \left(\frac{\mu_i^2}{\mu_i^2 + \mu^2\gamma^{2\ell}}\right) \tag{4.4}$$

Other quantities of interest associated with $x^{(k)}(\gamma^j\mu)$ include the B-residual

$$(4.5) \qquad s^{(k)}(\gamma^j\mu) = d - Bx^{(k)}(\gamma^j\mu) = -\sum_{i=1}^{q} \beta_i \varepsilon_i^{(k)}(j) v_i$$

and the A-residual

$$(4.6) \qquad r^{(k)}(\gamma^j\mu) = b - Ax^{(k)}(\gamma^j\mu) = r_{LSE} - \sum_{i=1}^{q} \alpha_i \varepsilon_i^{(k)}(j) u_i$$

Using these formulae we have

Theorem 4.1

If $\theta = \mu_1^2/(\mu_1^2 + \mu^2\gamma^{2(j-k)})$ and $x^{(k)}(\gamma^j\mu)$ is generated via (4.2), then

(a) $$\left\| x^{(k)}(\gamma^j\mu) - x_{LSE} \right\|_2 \leq \frac{\Delta\sqrt{1+\mu_1^2}}{2\mu\gamma^j\sigma_n}\,\theta^k$$

(b) $$\| d - Bx^{(k)}(\gamma^j\mu) \|_2 \leq \frac{\Delta}{2\mu\gamma^j}\,\theta^k$$

(c) $$0 \leq \|r_{LSE}\|_2 - \|b - Ax^{(k)}(\gamma^j\mu)\|_2 \leq \frac{\mu_1\Delta}{2\mu\gamma^j}\,\theta^k.$$

Proof

Using (4.4) and the inequality

$$\frac{\mu_i}{\mu_1^2 + \mu^2\gamma^{2j}} \leq \frac{1}{2\mu\gamma^j}$$

it follows that $|\varepsilon_i^{(k)}(j)| \leq \frac{|\rho_i|}{\beta_i}\,\frac{\theta^k}{2\mu\gamma^j}$. Since $\|X\|_2 = 1$ we have by taking norms in (4.3) and using (2.7) that

$$\|e^{(k)}(\gamma^j\mu)\|_2 \leq \frac{\Delta}{2\mu\,\beta_1\,\gamma^j}\,\theta^k.$$

Inequality (a) follows since $\beta_1^{-2} = (1+\mu_1^2)/\sigma_n^2$. To prove (b) and the

upper bound in (c), take norms in (4.5) and (4.6) respectively and use the orthonormality of $\{u_1\}$ and $\{v_i\}$.

To establish the lower bound in (c), note from (4.6) and (2.6) that

$$r^{(k)}(\gamma^j\mu) = r_{LS} + \sum_{i=1}^{q} (\rho_i - \alpha_i \varepsilon_i^{(k)}(j))\, u_i.$$

Since $r_{LS} \in \text{span}\{u_1, \ldots, u_q\}^{\perp}$ it follows that

$$||r^{(k)}(\gamma^j\mu)||_2^2 = ||r_{LS}||_2^2 + \sum_{i=1}^{q} (\rho_i - \alpha_i \varepsilon_i^{(k)}(j))^2.$$

The desired bound follows from (2.7) and the easily verified inequality $\rho_i^2 \leq (\rho_i - \alpha_i \varepsilon_i^{(k)}(j))^2$. (Note that $||r^{(k)}(\gamma^j\mu)||_2$ increases monotonically to $||r_{LSE}||_2$ as μ increases.) □

The main observation to make from the theorem is that the error, A-residual, and B-residual improve by a factor θ with each extrapolation. Fór example, in a small problem with $\mu_1 \approx 10^3$, we set $\mu = 100$, $\gamma = 2$ and found the following relative errors in $x^{(k)}(\gamma^j\mu)$:

$\gamma^j\mu$	k=0	k=1	k=2	k=3	k=4
100	10^0				
200	10^0	10^{-1}			
400	10^{-1}	10^{-2}	10^{-2}		
800	10^{-2}	10^{-3}	10^{-4}	10^{-4}	
1600	10^{-2}	10^{-4}	10^{-5}	10^{-6}	10^{-6}
3200	10^{-3}	10^{-5}	10^{-7}	10^{-8}	10^{-8}

In another example ($\mu_1 \cong 10^5$) we found the relative errors dimishing as follows

$\gamma^j\mu$	k=0	k=1	k=2	k=3
10^4	10^0			
10^6	10^{-1}	10^{-1}		
10^8	10^{-5}	10^{-6}	10^{-6}	
10^{10}	10^{-9}	10^{-11}	10^{-11}	10^{-11}

The corresponding entries for the relative B-residual $||d-Bx^{(k)}(\gamma^j\mu)||/||B||_2 * ||x_{LSE}||_2$ were smaller by a factor of 10^4.

Based on numerous small examples run with MATLAB we conclude

(a) The extrapolation cannot "take hold" until $\gamma^j \approx \mu_1$.

(b) In ill-conditioned examples, there is no significant reduction in the error beyond the second or third extrapolation.

(c) There is no point in extrapolating if $\gamma^j\mu >$ (macheps)$^{-1/2}$ when the unconstrained LS problem is in A-over-B form.

(d) Once the extrapolates "settle down", $||x^{(k)}(\gamma^j\mu) - x^{(k-1)}(\gamma^j\mu)||_2$ gives good estimates of the error in $x^{(k-1)}(\gamma^{j-1}\mu)$.

(e) Most time is spent in computing the Q-R decompositions for each $x(\gamma^j\mu)$. The extrapolates are "free".

5. The Iterative Improvement of $x(\mu)$

A routine Lagrange multiplier argument can be used to show that x_{LSE} and $r_{LSE} = b - Ax_{LSE}$ satisfy

$$(5.1) \qquad \begin{bmatrix} b \\ d \\ 0 \end{bmatrix} = \begin{bmatrix} I & 0 & A \\ 0 & 0 & B \\ A^T & B^T & 0 \end{bmatrix} \begin{bmatrix} r_{LSE} \\ \lambda_{LSE} \\ x_{LSE} \end{bmatrix}$$

Here, λ_{LSE} is the corresponding Lagrange multiplier.

The unconstrained LS problem (1.2) can also be posed as a linear equation problem:

$$\begin{bmatrix} b \\ d \\ 0 \end{bmatrix} = \begin{bmatrix} I & 0 & A \\ 0 & \mu^{-2}I & B \\ A^T & B^T & 0 \end{bmatrix} \begin{bmatrix} b - Ax(\mu) \\ \mu^2(d - Bx(\mu)) \\ x(\mu) \end{bmatrix}$$

(The last row in this equation is just the normal equation system (2.8).) This suggests the following iteration for improving $x(\mu)$:

Based on numerous small examples run with MATLAB we conclude

(a) The extrapolation cannot "take hold" until $\gamma^j \approx \mu_1$.

(b) In ill-conditioned examples, there is no significant reduction in the error beyond the second or third extrapolation.

(c) There is no point in extrapolating if $\gamma^j\mu >$ (macheps)$^{-1/2}$ when the unconstrained LS problem is in A-over-B form.

(d) Once the extrapolates "settle down", $||x^{(k)}(\gamma^j\mu) - x^{(k-1)}(\gamma^j\mu)||_2$ gives good estimates of the error in $x^{(k-1)}(\gamma^{j-1}\mu)$.

(e) Most time is spent in computing the Q-R decompositions for each $x(\gamma^j\mu)$. The extrapolates are "free".

5. The Iterative Improvement of $x(\mu)$

A routine Lagrange multiplier argument can be used to show that x_{LSE} and $r_{LSE} = b - Ax_{LSE}$ satisfy

$$(5.1) \qquad \begin{bmatrix} b \\ d \\ 0 \end{bmatrix} = \begin{bmatrix} I & 0 & A \\ 0 & 0 & B \\ A^T & B^T & 0 \end{bmatrix} \begin{bmatrix} r_{LSE} \\ \lambda_{LSE} \\ x_{LSE} \end{bmatrix}$$

Here, λ_{LSE} is the corresponding Lagrange multiplier.

The unconstrained LS problem (1.2) can also be posed as a linear equation problem:

$$\begin{bmatrix} b \\ d \\ 0 \end{bmatrix} = \begin{bmatrix} I & 0 & A \\ 0 & \mu^{-2}I & B \\ A^T & B^T & 0 \end{bmatrix} \begin{bmatrix} b - Ax(\mu) \\ \mu^2(d-Bx(\mu)) \\ x(\mu) \end{bmatrix}$$

(The last row in this equation is just the normal equation system (2.8).) This suggests the following iteration for improving $x(\mu)$:

$x := x(\mu)$; $\lambda := \mu^2(d - Bx(\mu))$; $r := b - Ax(\mu)$

Repeat:

(5.3)

$$\begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} = \begin{bmatrix} b \\ d \\ 0 \end{bmatrix} - \begin{bmatrix} I & 0 & A \\ 0 & 0 & B \\ A^T & B^T & 0 \end{bmatrix} \begin{bmatrix} r \\ \lambda \\ x \end{bmatrix}$$

$$\begin{bmatrix} I & 0 & A \\ 0 & \mu^{-2}I & B \\ A^T & B^T & 0 \end{bmatrix} \begin{bmatrix} \Delta r \\ \Delta\lambda \\ \Delta x \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}$$

$x := x + \Delta x$; $\lambda := \lambda + \Delta\lambda$; $r := r + \Delta r$

In this scheme, if the matrix that defines the corrections Δr, $\Delta\lambda$, and Δx is replaced by the matrix of (5.1), then we obtain the iterative improvement scheme of Bjork and Golub [2]. These authors were interested in how to improve the computed solution $\hat{x}_{LSE}$ obtained via (1.1). We, however, are only interested in getting a solution approximately as good as $\hat{x}_{LSE}$. Consequently, we need not resort to multiple precision computation of the residuals δ_i.

The iteration (5.3) undergoes considerable simplification when we observe that δ_1 and δ_3 are always zero. (Use induction.) In light of this observation, it follows that Δx satisfies

$$\left\| \begin{pmatrix} \mu B \\ A \end{pmatrix} \Delta x - \begin{pmatrix} \mu\delta_2 \\ 0 \end{pmatrix} \right\|_2 = \min$$

where $\delta_2 = d - Bx$. Thus, (5.3) collapses to

$x(\mu, 0) := x(\mu)$

For $k = 0, 1, \ldots$

(5.4)

$\delta^{(k)} := d - Bx(\mu, k)$

Compute the solution $\Delta x^{(k)}$ to

$$\min_z \left\| \begin{pmatrix} \mu B \\ A \end{pmatrix} z - \begin{pmatrix} \mu\delta^{(k)} \\ 0 \end{pmatrix} \right\|_2$$

$x(\mu, k+1) = x(\mu, k) + \Delta x^{(k)}$

Note that only one Q-R factorization, that of $\begin{pmatrix} \mu B \\ A \end{pmatrix}$, is needed to execute (5.4).

The convergence of the method is established in the following result:

Theorem 5.1

In the notation of Section 2, the vectors $x(\mu,k)$ generated by (5.4) satisfy

$$x(\mu,k) = x_{LSE} + e(\mu,k)$$

where

$$e(\mu,k) = \sum_{i=1}^{q} \frac{\rho_i}{\beta_i} \frac{\mu_i}{\mu_i^2+\mu^2} \theta_i^k x_i$$

with $\theta_i = \mu_i^2/(\mu_i^2 + \mu^2)$.

Proof

We use induction observing from (2.10) that the theorem is true for $k=0$. Since $\delta^{(k)} = - Be(\mu,k)$ we have by induction that

$$\delta^{(k)} = - \sum_{i=1}^{q} \rho_i \frac{\mu_i}{\mu_i^2+\mu^2} \theta_i^k v_i.$$

Now $(A^TA+\mu^2B^TB)^{-1}B^T = X(D_A^TD_A + \mu^2D_B^TD_B)^{-1}D_B^TV^T$ and so

$$\Delta x^{(k)} = \mu^2(A^TA+\mu^2B^TB)^{-1}B^T\delta(k)$$

$$= -\mu^2 \sum_{i=1}^{q} \frac{\rho_i}{\beta_i} \frac{\mu_i}{\mu_i^2+\mu^2} \theta_i^{k+1} x_i .$$

The theorem follows directly from the equation $e(\mu,k+1) = e(\mu,k)+\Delta x^{(k)}$. □

Arguments very similar to those used in the proof of Theorem 4.1 can be used to establish the following inequalities

$$||x(\mu,k) - x_{LSE}||_2 \leq \frac{\Delta\sqrt{1+\mu_1^2}}{2\mu\sigma_n} \theta_1^k$$

$$||d - Bx(\mu,k)||_2 \leq \frac{\Delta}{2\mu} \theta_1^k$$

$$0 \leq ||r_{LSE}||_2 - ||b - Ax(\mu,k)||_2 \leq \frac{\Delta\mu_1}{2\mu}\,\theta_1^k.$$

These bounds closely correspond to those for the extrapolation method.

To examine the effectiveness of the iteration, we applied it to a problem in which μ_1=5000. The relative error of $x(\mu,k)$ is tabulated in the following table:

μ	k=0	k=1	k=2	k=3	k=4
10^3	10^0	10^0	10^0	10^0	10^0
10^4	10^{-1}	10^{-2}	10^{-3}	10^{-3}	10^{-4}
10^5	10^{-3}	10^{-6}	10^{-8}	10^{-11}	10^{-14}
10^8	10^{-9}	10^{-14}	10^{-14}	10^{-14}	10^{-14}

Note that the iteration cannot substantially improve the accuracy of $x(\mu,0)$ unless $\mu_1^2/(\mu_1^2+\mu^2)$ is somewhat less than unity.

If the A-over-B formulation to compute $x(\mu,0)$ is used, then improvements in the iterates should not be expected if $\mu >$ (macheps)$^{-\frac{1}{2}}$.

In practice, $||\Delta x^{(k)}||_2$ is a very good estimate of the error in $x(\mu,k)$ once convergence begins to set in.

6. Conclusions

Much more work is needed before the ideas in this paper can take on a practical form. Some topics for future research include the following.

(i) A closer examination of ill-conditioning is needed. Does the weighting method merely fail to compute the "unstable" components of x_{LSE}?

(ii) It is not necessary to re-compute the Q-R factorization of $\begin{pmatrix}\mu B\\ A\end{pmatrix}$ for each different μ when extrapolating. Instead, the Q-R factorization of A can be updated [1]. How could this idea be implemented with the George-Heath algorithm [6]?

(iii) What are good values for γ and μ when extrapolating? How many extrapolation steps can we "afford"?

(iv) Are there ways to implement (5.4) without having to store the complete Q-R factorization of $\begin{pmatrix} \mu B \\ A \end{pmatrix}$?

It is our intention to continue investigating these questions.

Acknowledgements

I would like to thank Gene Golub, Per-Åke Wedin, and the referee for their constructive comments. In addition, the support of the National Science Foundation and the Swedish Natural Science Research Council is gratefully acknowledged. This paper was produced while I was a visitor at the University of Umeå.

REFERENCES

1. Å. BJÖRK (1981), "A general updating algorithm for constrained linear least squares problems," Report LiTH-MAT-R-81-18, Department of Mathematics, University of Linkoping, Sweden.

2. Å. BJÖRK AND G.H. GOLUB (1967), "Iterative refinement of linear least squares solutions by Householder transformation," BIT, 7, 327-337.

3. Å. BJÖRK AND I.S. DUFF (1980), "A direct method for the solution of sparse linear least squares problems," Lin. Alg.&Applic., 34, 43-67.

4. P. BUSINGER AND G.H. GOLUB (1965), "Linear least squares solutions by Householder transformations," Numer. Math. 7, 169.

5. L. ELDEN (1980), "Perturbation theory for the least squares problem with linear equality constraints", SIAM J. Numer. Anal., 17, 338-350.

6. A. GEORGE AND M. HEATH (1980), "Solution of sparse linear least squares problems using Givens rotations," Lin. Alg.& Applic. 34, 69-84.

7. C.L LAWSON AND R.J. HANSON (1974), "Solving least squares problems, Prentice-Hall, Englewood Cliffs NJ.

8. M.J.D. POWELL AND J.K. REID, (1969), "On applying Householder's method to linear least squares problems," Proc. IFIP Congress, 1968.

9. G.W. STEWART (1977)," On the perturbation of pseudo-inverses, projections and linear least squares problems," SIAM Review, 19, 634-662.

10. G.W. STEWART (1982), " A Method for Computing the Generalized Singular value decomposition," this volume.

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

12. P-Å WEDIN (1979), "Notes on the constrained least squares problem. A new approach based on generalized inverses," Report UMINF 75.79, Institute of Information Processing, University of Umeå, Sweden.

13. C.B. MOLER, (1980), "MATLAB- An Interactive matrix Laboratory," Dept of Computer Science, University of New Mexico, Albuquerque, New Mexico.