Least Squares Constrained by a Linear Equation

Given matrices F, g, C and d, we seek a vector u to minimize $\|F\cdot u - g\|^2 = (F\cdot u - g)'\cdot(F\cdot u - g)$ subject to the linear constraint $C\cdot u = d$. By introducing a vector v of *Lagrange Multipliers* we find that the minimizing vector $u = \hat{u}$ must satisfy the linear equations

$$\begin{bmatrix} F'\cdot F & C' \\ C & O \end{bmatrix}\cdot\begin{bmatrix} \hat{u} \\ v \end{bmatrix} = \begin{bmatrix} F'\cdot g \\ d \end{bmatrix}\ ; \qquad \textit{i.e.},\ \ C\cdot\hat{u} = d \ \text{ and } \ F'\cdot(g - F\cdot\hat{u}) = C'\cdot v\ . \qquad (*)$$

We shall show first that these equations have at least one solution $\hat{u}$ provided the constraint $C\cdot u = d$ is satisfiable, second that any such solution does minimize $\|F\cdot u - g\|^2$, and third that the solution $\hat{u}$ is unique just when the columns of $\begin{bmatrix} C \\ F \end{bmatrix}$ are linearly independent even if the rows of $[C\ \ d]$ are not. Fourth, we shall explore a way to compute $\hat{u}$ without solving equations (*). C will be assumed to have rather fewer rows than columns, and F to have more rows than columns.

Existence

To show that a solution $\hat{u}$ exists we shall invoke one of Fredholm’s criteria: $A\cdot x = b$ has at least one solution x if and only if $w'\cdot b = 0$ whenever $w'\cdot A = o'$. To this end we suppose that $[w'\ \ y']\cdot\begin{bmatrix} F'\cdot F & C' \\ C & O \end{bmatrix} = [o'\ \ o']$. This means that $w'\cdot F'\cdot F + y'\cdot C = o'$ and $w'\cdot C' = o'$. When the last equation is substituted into the second-last we find that $w'\cdot F'\cdot F\cdot w = 0$, whence follows in turn that $w'\cdot F' = o'$ and then $y'\cdot C = o'$. Consequently $[w'\ \ y']\cdot\begin{bmatrix} F'\cdot g \\ d \end{bmatrix} = w'\cdot F'\cdot g + y'\cdot d = 0$ provided the constraint $d = C\cdot u$ is satisfiable. Then equations (*) have at least one solution $\hat{u}$.

Minimization

Any solution $\hat{u}$ of (*) makes $\|F\cdot\hat{u} - g\|^2$ no bigger than $\|F\cdot u - g\|^2$ for any u that satisfies the constraint $C\cdot u = d$. Here is why:

$$\begin{aligned} \|F\cdot u - g\|^2 - \|F\cdot\hat{u} - g\|^2 &= (F\cdot(u-\hat{u}) + F\cdot\hat{u}-g)'\cdot(F\cdot(u-\hat{u}) + F\cdot\hat{u}-g) - (F\cdot\hat{u}-g)'\cdot(F\cdot\hat{u}-g) \\ &= \|F\cdot(u-\hat{u})\|^2 + 2(u-\hat{u})'\cdot F'\cdot(F\cdot\hat{u}-g) \\ &= \|F\cdot(u-\hat{u})\|^2 - 2(u-\hat{u})'\cdot C'\cdot v \qquad \text{from } (*) \\ &= \|F\cdot(u-\hat{u})\|^2 \qquad \text{because } C\cdot u = C\cdot\hat{u} = d \\ &\ge 0\ . \end{aligned}$$

Uniqueness

If more than one solution $\hat{u}$ exists their nonzero difference z must satisfy $\begin{bmatrix} F'\cdot F & C' \\ C & O \end{bmatrix}\cdot\begin{bmatrix} z \\ s \end{bmatrix} = o$ which means $F'\cdot F\cdot z + C'\cdot s = o$ and $C\cdot z = o$. As before, this implies $z'\cdot F'\cdot F\cdot z = 0$, whence follows $F\cdot z = o$ too, which is impossible if the columns of $\begin{bmatrix} C \\ F \end{bmatrix}$ are linearly independent, in which case solution $\hat{u}$ is determined uniquely by the linear equations (*) displayed above.

The Solution $\hat{u}$ as a Limit

When unique, this solution $\hat{u}$ turns out to be the limit, as $\mu \to \infty$, of the solution u of the unconstrained least-squares problem that chooses u to minimize $\| \begin{bmatrix} \mu \cdot C \\ F \end{bmatrix} \cdot u - \begin{bmatrix} \mu \cdot d \\ g \end{bmatrix} \|^2$. This problem should be solved numerically by QR factorization after some suitably huge μ has been chosen. However, to prove that $u \to \hat{u}$ as $\mu \to \infty$, we shall use the numerically dubious closed-form formula $u = (\mu^2 \cdot C' \cdot C + F' \cdot F)^{-1} \cdot (\mu^2 \cdot C' \cdot d + F' \cdot g)$ obtained from this least-squares problem's *Normal Equations*. By substitution from (*) we find that $u = \hat{u} + (\mu^2 \cdot C' \cdot C + F' \cdot F)^{-1} \cdot C' \cdot v$. Since v is determined by (*), though not uniquely if the rows of C are linearly dependent, our objective will be attained when we show that $\|(\mu^2 \cdot C' \cdot C + F' \cdot F)^{-1} \cdot C'\| \to 0$ as $\mu \to \infty$. This is barely unobvious enough to be worth proving thrice.

The first and shortest proof uses a *Congruence* to diagonalize the two positive (semi-)definite matrices $C' \cdot C = L^{-1} \cdot M \cdot L'^{-1}$ and $F' \cdot F = L^{-1} \cdot W \cdot L'^{-1}$ simultaneously. Here L is some invertible matrix; M and W are nonnegative diagonal matrices whose respective elements' ratios m_{jj}/w_{jj} are generalized eigenvalues determined by a determinantal equation $\det(w_{jj} \cdot C' \cdot C - m_{jj} \cdot F' \cdot F) = 0$ in which every $m_{jj} + w_{jj} > 0$ because the columns of $\begin{bmatrix} C \\ F \end{bmatrix}$ are linearly independent and therefore $\begin{bmatrix} C \\ F \end{bmatrix}' \cdot \begin{bmatrix} C \\ F \end{bmatrix} = C' \cdot C + F' \cdot F = L^{-1} \cdot (M + W) \cdot L'^{-1}$ must be positive definite. Consequently

$$\begin{aligned} \|(\mu^2 \cdot C' \cdot C + F' \cdot F)^{-1} \cdot C'\|^2 &= \|(\mu^2 \cdot C' \cdot C + F' \cdot F)^{-1} \cdot C' \cdot C \cdot (\mu^2 \cdot C' \cdot C + F' \cdot F)^{-1}\| \\ &= \|L' \cdot M \cdot (\mu^2 \cdot M + W)^{-2} \cdot L\| \to 0 \text{ as } \mu \to \infty \end{aligned}$$

because every nonzero diagonal element $m_{jj}/(\mu^2 \cdot m_{jj} + w_{jj})^2 \to 0$. So ends the first proof.

A second proof starts from any orthogonalization of $\begin{bmatrix} C \\ F \end{bmatrix} = Q \cdot R$ in which R is invertible and the columns of Q provide an orthonormal basis ($Q' \cdot Q = I$) for $\mathrm{Range}(\begin{bmatrix} C \\ F \end{bmatrix})$. This subspace is separated by angles Θ from $\mathrm{Range}(\begin{bmatrix} I \\ O \end{bmatrix})$ in which the small identity matrix I and the diagonal square matrix Θ each has as many rows as C has, and $0 \le \Theta \le \pi/2$.

These angles Θ were exposed in "Some New Bounds on Perturbations of Subspaces" by C. Davis and W.M. Kahan in pp. 863-9 of *Bull. Amer. Math. Soc.* **75** #4 (July 1969), and explained in their "The Rotation of Eigenvectors by a Perturbation. III" in pp. 1-46 of *SIAM J. Numer. Anal.* **7** #1 (March 1970). C.C. Paige and M. Wei surveyed other applications in their "History and Generality of the CS Decomposition", pp. 303-326 in *Linear Algebra and Its Applications* **108/109** (1994). This second proof is an application instigated by C.F. Van Loan in "Generalizing the Singular Value Decomposition", pp. 76-83 of *SIAM J. Numer Anal.* **13** (1976).

Aptly chosen orthogonal $P' = P^{-1} = \begin{bmatrix} P'_C & O \\ O & P'_F \end{bmatrix}$ and $U' = U^{-1}$ turn Q into $P' \cdot Q \cdot U' = \begin{bmatrix} \cos(\Theta) & O \\ \sin(\Theta) & O \\ O & I \\ O & O \end{bmatrix}$

and decompose $C = P_C \cdot [\cos(\Theta)\ \ O] \cdot U \cdot R$ and $F = P_F \cdot \begin{bmatrix} \sin(\Theta) & O \\ O & I \\ O & O \end{bmatrix} \cdot U \cdot R$. Into the first proof we now substitute $L := U \cdot R'^{-1}$, $M := \begin{bmatrix} \cos^2(\Theta) & O \\ O & O \end{bmatrix}$ and $W := \begin{bmatrix} \sin^2(\Theta) & O \\ O & I \end{bmatrix}$ to find as before that

$$\|(\mu^2 \cdot C' \cdot C + F' \cdot F)^{-1} \cdot C'\|^2 \ = \ \|L' \cdot M \cdot (\mu^2 \cdot M + W)^{-2} \cdot L\| \to 0 \quad \text{as} \quad \mu \to \infty$$

because every nonzero diagonal element $m_{jj}/(\mu^2 \cdot m_{jj} + w_{jj})^2 = \big(\sec(\theta_j)/(\mu^2 + \tan^2(\theta_j)) \big)^2 \to 0$ at a rate determined by one of the angles $\theta_j < \pi/2$. So ends the second proof. However, what it tells us about how fast $u \to \hat{u}$ involves an obscure dependence of L and Θ upon C and F . For instance, replacing the constraint $C \cdot u = d$ by some equivalent constraint $B^{-1} \cdot C = B^{-1} \cdot d$ for an invertible B alters u, L and Θ without altering $\hat{u}$.

Though more computational, the third proof will help us assess how fast $u \to \hat{u}$ as $\mu \to \infty$. Let $\text{æ} := 1/\mu^2 \to 0+$ as $\mu \to \infty$. Then $\|(\mu^2 \cdot C' \cdot C + F' \cdot F)^{-1} \cdot C'\| = \text{æ} \cdot \|(C' \cdot C + \text{æ} \cdot F' \cdot F)^{-1} \cdot C'\|$, so our objective will be attained when we have shown that $\|(C' \cdot C + \text{æ} \cdot F' \cdot F)^{-1} \cdot C'\|$ stays bounded as $\text{æ} \to 0+$. Since the biggest-singular-value norm $\|\ldots\|$ is unitarily invariant, it is not changed after premultiplication and postmultiplication by orthogonal matrices that exhibit the singular value decomposition of C as $\begin{bmatrix} V & O \\ O & O \end{bmatrix}$ in which V is a square strictly positive diagonal matrix of the nonzero singular values of C . The second row $[O \quad O]$ is empty when the rows of C are linearly independent. After the same postmultiplication, F partitions into $[E, K]$ conformal with the partition of the columns of $\begin{bmatrix} V & O \\ O & O \end{bmatrix}$. Because the columns of $\begin{bmatrix} C \\ F \end{bmatrix}$ are linearly independent, so are the columns of $\begin{bmatrix} V & O \\ O & O \\ E & K \end{bmatrix}$ and then of K , whose pseudo-inverse is $K^\dagger = (K' \cdot K)^{-1} \cdot K'$. Then $C' \cdot C + \text{æ} \cdot F' \cdot F$ becomes $\begin{bmatrix} V^2 + \text{æ} \cdot E' \cdot E & \text{æ} \cdot E' \cdot K \\ \text{æ} \cdot K' \cdot E & \text{æ} \cdot K' \cdot K \end{bmatrix}$. Approximately $\begin{bmatrix} V^{-2} & -V^{-2} \cdot E' \cdot K'^\dagger \\ -K^\dagger \cdot E \cdot V^{-2} & K^\dagger \cdot K'^\dagger / \text{æ} \end{bmatrix}$ is its inverse because their product differs from an identity matrix by terms of order æ as $\text{æ} \to 0+$. Then $(C' \cdot C + \text{æ} \cdot F' \cdot F)^{-1} \cdot C'$ becomes approximately $\begin{bmatrix} V^{-1} & O \\ -K^\dagger \cdot E \cdot V^{-1} & O \end{bmatrix}$ to within terms of order æ as $\text{æ} \to 0+$. This stays bounded, as was asserted earlier. So ends the third proof.

Thus we see why $u - \hat{u} = O(1/\mu^2)$ as $\mu \to \infty$ whenever equation (*) determines $\hat{u}$ uniquely.

The third proof revealed something worth knowing about how the constant implicit in $O(1/\mu^2)$ depends upon the matrices C and F through V, E and K . That constant can be huge only if $\|V^{-1}\|$ and/or $\|K^\dagger \cdot E\|$ is huge. This can’t happen unless the rows of C and/or the columns of $\begin{bmatrix} C \\ F \end{bmatrix}$ are too nearly linearly dependent. In such cases the computation of u for ever increasing values of μ can appear to converge over a wide range of big values μ although the true limit $\hat{u}$ is not approached until μ gets very much bigger. In such cases a numerically satisfactory choice for μ can be difficult to ascertain. Here are examples:

Example 1: C has nearly linearly dependent rows.

Let $C := \begin{bmatrix} 1 & 0 & 0 \\ 0 & \eta & 0 \end{bmatrix}$, $d := \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $F' := \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}$ and $g' := [3\ \ 1\ \ 1\ \ 99\ \ 99]$. Now $\hat{u} = [0\ \ 0\ \ 1]'$ but as $\mu \to \infty$ the error $\hat{u} - u = [0\ \ {-2}\ \ 1]'/(1 + 2\eta^2 \cdot \mu^2/3) \to o$ at a rate determined by a parameter $\eta \ne 0$ that tells how near to linearly dependent the rows of C are. When η is very tiny the computed vector $u \approx [0\ \ 2\ \ 0]'$ for a wide range of huge values $\mu << 1/\eta$. To prevent acceptance of this plausible u in place of $\hat{u}$ we must know enough about η to choose $\mu >> 1/\eta$. Examples like this are vexatious also because a small change in data can alter the solution $\hat{u}$ drastically; for instance, changing d slightly from $[0\ \ 0]'$ to $[0\ \ \eta]'$ alters $\hat{u}$ from $[0\ \ 0\ \ 1]'$ to $[0\ \ 1\ \ 1/2]'$ and alters the error to $\hat{u} - u = [0\ \ {-1}\ \ 1/2]'/(1 + 2\eta^2 \cdot \mu^2/3)$. Changing η to 0 makes $\hat{u} = u = [0\ \ 2\ \ 0]'$. Examples like this will motivate remedial action against constraints [C d] with nearly linearly dependent rows when their near-redundancy is an accident due to a mathematical mishap easily dispelled.

Example 2: $\begin{bmatrix} C \\ F \end{bmatrix}$ has nearly linearly dependent columns.

Let $C := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$, $d = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $F' := \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \eta & 0 \end{bmatrix}$ and $g' := [\xi\ \ 3\ \ 1\ \ \eta\ \ 99]$. Now $\hat{u} = [0\ \ 0\ \ 1\ \ 1]'$ but as $\mu \to \infty$ the error $\hat{u} - u = [-\xi\ \ {-3}\ \ 0\ \ \xi/\eta]'/(1 + \mu^2) \to o$ at a rate determined by two parameters: One is $\eta \ne 0$ that tells how near to linearly dependent the columns of F are. The other, ξ , exerts no influence upon $\hat{u}$ but, when η is very tiny, affects the computed vector u drastically unless $\mu^2 >> \xi/\eta$. This is how μ must be chosen to prevent acceptance of a plausible u in place of $\hat{u}$, and this choice depends upon g as well as C and F . Again, $\hat{u}$ can be altered drastically by a small change in the data; for instance, changing g' slightly to $g' := [\xi\ \ 3\ \ 1\ \ 0\ \ 99]$ alters $\hat{u}$ to $[0\ \ 0\ \ 1\ \ 0]'$ without changing the error $\hat{u} - u$. This kind of hypersensitivity to perturbation can be difficult to dispel by solely numerical means when it arises from a redundancy injected into the least-squares problem at an early stage of its mathematical formulation. It happens often.

Misbehavior ostensibly similar to the foregoing two examples can occur when the first few columns of C are too nearly linearly dependent even though the rows of C are not. This misbehavior is caused by roundoff. Increasing μ worsens the misbehavior. Here is an example:

Example 3: The first two columns of C are linearly dependent.

Let $C := \begin{bmatrix} 4 & \frac{4}{10} & 1 \\ 3 & \frac{3}{10} & -1 \end{bmatrix}$, $d := \begin{bmatrix} 3 \\ -3 \end{bmatrix}$, $F' := \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 2 & -1 & 1 & 1 \\ 1 & \frac{2}{3} & 1 & -\frac{4}{3} & 0 \end{bmatrix}$ and $g' := [7\ \ {-18}\ \ 12\ \ {-15}\ \ {-9}]$. Now $\hat{u} = \begin{bmatrix} 1 \\ -10 \\ 3 \end{bmatrix}$.

Error $\hat{u} - u = [15100\mu^2{+}29500\ \ \ {-48550\mu^2}{-}8500\ \ \ {-256125\mu^2}{-}30000]'/(301203\mu^4{+}614370\mu^2{+}23700) \to o$ like $1/\mu^2$ in the absence of roundoff, so any choice $\mu > 2^{26} \approx 6.7 \cdot 10^7$ should keep the error $\hat{u} - u$ below the uncertainty inherited when the data C and F were rounded to 53 sig. bits.

Let $C := \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \end{bmatrix}$, $d := \begin{bmatrix} 28 \\ 16 \end{bmatrix}$, $F' := \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$ and $g' := [4\ \ 8\ \ 12\ \ 16]$. Now $\hat{u} = \begin{bmatrix} 23 \\ -1 \\ 6 \end{bmatrix}$. This example was constructed by C.F. Van Loan. Its error $\hat{u} - u = [36\ \ 0\ \ 36]'/(\mu^2+4) \rightarrow o$ like $36/\mu^2$ in the absence of roundoff, so any choice $\mu > 2^{27} \approx 1.34 \cdot 10^8$ should keep the error $\hat{u} - u$ comparable with what we might expect to inherit from arithmetic rounded to 53 sig. bits.

For other treatments of the linearly constrained least squares problem see Ch. 5.1 of Åke Björck’s book *Numerical Methods for Least Squares Problems* (1996, S.I.A.M, Philadelphia), and §17 and §22 of C.L. Lawson & R.J. Hanson’s book *Solving Least Squares Problems* (1974, Prentice-Hall, New Jersey).

Charles Van Loan’s Example:

$F := \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, $g := \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$, $C := \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \end{bmatrix}$, $d := \begin{bmatrix} 7 \\ 4 \end{bmatrix}$, $\hat{u} = \begin{bmatrix} 23 \\ -1 \\ 6 \end{bmatrix}/4$ computed to only half-precision when $\mu := 1/\sqrt{eps}$ without full column pivoting. He had used the Generalized SVD via the CS-decomposition $Q' \cdot F = [\sin(\text{angles})\text{ on diag}] \cdot Z$, $P' \cdot C = [\cos(\text{angles})\text{ on diag}] \cdot Z$, and $Z' \cdot Z = (F' \cdot F + C' \cdot C)$ to analyze the algorithm’s behavior, which resembles Gaussian Elimination if μ is big enough.

If the rows of C as well as the columns of $\begin{bmatrix} C \\ F \end{bmatrix}$ are linearly independent, then the constraint $C \cdot u = d$ is satisfiable for every d by $u = C' \cdot (C \cdot C')^{-1} \cdot d$ among other things. Also the equations (*) have a unique solution $\begin{bmatrix} \hat{u} \\ v \end{bmatrix}$. In this case the equations (*) are equivalent to equations

$$\begin{bmatrix} O & C' & F' \\ C & O & O \\ F & O & -I \end{bmatrix} \cdot \begin{bmatrix} \hat{u} \\ v \\ r \end{bmatrix} = \begin{bmatrix} o \\ d \\ g \end{bmatrix}, \text{ namely } C \cdot \hat{u} = d,\ \ r = F \cdot \hat{u} - g,\ \text{ and }\ F' \cdot r + C' \cdot v = o, \qquad \binom{*}{*}$$

whose matrix is nonsingular. Here is a proof that it is nonsingular:

If $C\cdot u = o$, $F\cdot u - r = o$ and $F\cdot r + C'\cdot v = o$, then $0 = u'\cdot F'\cdot r + u'\cdot C'\cdot v = u'\cdot F'\cdot F\cdot u + 0$, whence follows $F\cdot u = o$ and therefore $u = o$ because the columns of $\begin{bmatrix} C \\ F \end{bmatrix}$ are linearly independent. Also $r = o$ and therefore $C'\cdot v = o$, whence $v = o$ because the rows of C are linearly independent. In other words, the matrix of the equations $\binom{*}{*}$ has only o in its nullspace so it must be nonsingular, as claimed.

The numerically dubious closed-form formula $u = (\mu^2\cdot C'\cdot C + F'\cdot F)^{-1}\cdot(\mu^2\cdot C'\cdot d + F'\cdot g)$ turns out to be equivalent (in the absence of roundoff) to equations

$$\begin{bmatrix} O & C' & F' \\ C & -I/\mu^2 & O \\ F & O & -I \end{bmatrix}\cdot\begin{bmatrix} u \\ w \\ s \end{bmatrix} = \begin{bmatrix} o \\ d \\ g \end{bmatrix} \text{ , namely } C\cdot u - w/\mu^2 = d \text{ , } \quad s = F\cdot u - g \text{ , and } F'\cdot s + C'\cdot w = o \text{ , } \qquad \binom{\#}{\#}$$

whose matrix (also nonsingular) differs from the matrix of equations $\binom{*}{*}$ above by the term $-I/\mu^2$ in the middle. If this term is tiny enough, negligible compared with the smallest singular value of the latter matrix, the two systems of linear equations will have negligibly different solutions u and $\hat{u}$. Though that smallest singular value is positive because of our linear independence hypotheses, its value can change when the constraint $C\cdot u = d$ is replaced by some equivalent constraint $B^{-1}\cdot C = B^{-1}\cdot d$ that leaves $\hat{u}$ unaltered. What does this approach to our problem imply?

Generating Test Data:

Given ostensibly “simple” F, C and $\hat{u}$, we set $d := C\cdot\hat{u}$ and $g := F\cdot\hat{u} - r$ wherein r need only be a not-too-complicated solution of $[C' \;\; F']\cdot\begin{bmatrix} v \\ r \end{bmatrix} = o$ for any arbitrary (perhaps complicated) v .

By reducing $[C' \;\; F']$ to its row-reduced echelon form, we can choose simple values for all but the first few elements of r almost arbitrarily subject to the constraint that those first few elements now determined by the foregoing equation be not too complicated. Whether this is feasible will depend upon how the first few rows of F correlate with C . Try MATLAB’s null([C' F'],’r’) .

How Roundoff Can Ruin the Computation of $\hat{u}$

Why was the formula $u = (\mu^2 \cdot C' \cdot C + F' \cdot F)^{-1} \cdot (\mu^2 \cdot C' \cdot d + F' \cdot g)$ called “numerically dubious”? It suffers intolerably from roundoff because of two phenomena. One occurs as μ gets very big: Digits of $F' \cdot F$ get rounded away when it is added to $\mu^2 \cdot C' \cdot C$. After μ becomes big enough, the computed value of $\mu^2 \cdot C' \cdot C + F' \cdot F$ becomes just the computed value of $\mu^2 \cdot C' \cdot C$, as if F were O . The same phenomenon can lose $F' \cdot g$ from the sum $\mu^2 \cdot C' \cdot d + F' \cdot g$. After these losses the computation of u , if not aborted, becomes an accident of roundoff divorced from the data F and g .

A second more subtle phenomenon can afflict the computation of $F' \cdot F$ and $F' \cdot g$ in both the numerically dubious formula for u and equation (*) that defines $\hat{u}$. This second phenomenon can arise when the columns of F are too nearly linearly dependent even though the columns of $\begin{bmatrix} C \\ F \end{bmatrix}$ are amply independent enough to determine $\hat{u}$ sharply.

Backward Error in $F' \cdot F$

Suppose $A := F' \cdot F$ rounded; actually $A = F' \cdot F + \Delta A$. How small a perturbation ΔF can satisfy $A = (F+\Delta F)' \cdot (F+\Delta F)$ exactly? Extra-precise arithmetic may be needed to compute ΔA ; usually we expect $\|\Delta A\| \approx \text{æ} \cdot \|A\| \approx \text{æ} \cdot \|F\|^2$ wherein $\text{æ} << 1$ is a tiny roundoff threshold (like MATLAB’s eps) . But if F has singular values so disparate that condition number $\kappa(F) = \|F\| \cdot \|F^\dagger\| >> 1$, which occurs often, the smallest perturbation ΔF compatible with ΔA can have a surprisingly big norm $\|\Delta F\| \approx \text{æ} \cdot \|F\| \cdot \kappa(F) >> \text{æ} \cdot \|F\|$ though $\|\Delta F\|$ can almost never exceed $\sqrt{\text{æ}} \cdot \|F\|$.

A simple example:

Choose $\kappa >> 1$ while keeping $1 \pm \text{æ} \cdot \kappa^2 \geq 0$, and then set $F := \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix}$ and $\Delta A := \pm \text{æ} \cdot \kappa^2 \cdot \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ so that $A := F' \cdot F + \Delta A = \begin{bmatrix} \kappa^2 & 0 \\ 0 & 1 \pm \text{æ} \cdot \kappa^2 \end{bmatrix} = (F+\Delta F)' \cdot (F+\Delta F)$ for $\Delta F := \pm \text{æ} \cdot \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \cdot \kappa^2 / (1 + \sqrt{(1 \pm \text{æ} \cdot \kappa^2)})$.

No smaller ΔF is compatible with ΔA . Now $\|\Delta A\|/\|A\| = \text{æ}$, the roundoff threshold, but $\|\Delta F\|/\|F\| = \text{æ} \cdot \kappa / (1 + \sqrt{(1 \pm \text{æ} \cdot \kappa^2)}) >> \text{æ}$. Still, $\|\Delta F\|/\|F\| \leq \sqrt{\text{æ}}$ for all condition numbers κ that satisfy the assumed constraint $1 \pm \text{æ} \cdot \kappa^2 \geq 0$.

This simple example foreshadows what can happen in general: The roundoff term ΔA can induce a perturbation ΔF that obscures almost as many as the last half of the significant digits stored in F . Why nothing worse can ne expected to happen requires a lengthy explanation:

Roundoff's contribution ΔA must be small enough that $F'\cdot F + \Delta A = A = (F+\Delta F)'\cdot(F+\Delta F)$ is still positive (semi-)definite if ΔF exists. This obliges us to assume that F has linearly independent columns, and that $1 > \text{æ}\cdot\kappa(F)^2 \approx \|(F'\cdot F)^{-1}\|\cdot\|\Delta A\|$ to be sure that ΔF exists. Even so, ΔF cannot be determined uniquely by the previous sentence's equation alone since any one solution ΔF can be turned into another, $\text{Ø}'\cdot(F+\Delta F) - F$, by any orthogonal $\text{Ø}' = \text{Ø}^{-1}$. We can try to determine ΔF uniquely by restricting it somehow, say by trying to minimize $\|\Delta F\|$ while satisfying $F'\cdot\Delta F + \Delta F'\cdot F + \Delta F'\cdot\Delta F = \Delta A$. Our first estimate $\Delta_1 F$ will neglect the second-order term $\Delta F'\cdot\Delta F$. Our second estimate $\Delta_2 F$ will take this term into account but its $\|\Delta F\|$ will be somewhat bigger than minimal. Later a third estimate $\Delta_3 F$ will minimize not $\|\Delta F\|$ but the *Frobenius* norm $\|\Delta F\|_f := \sqrt{(\mathrm{Trace}(\Delta F'\cdot\Delta F))}$. (It is MATLAB's norm(ΔF, 'fro') .)

The singular-value decomposition of $F = Q\cdot\begin{bmatrix}V\\O\end{bmatrix}\cdot P'$ has orthogonal $Q' = Q^{-1}$, $P' = P^{-1}$ and a positive diagonal matrix V of singular values so ordered that the first $v_1 = \|F\|$ is the biggest and the last $v_n = 1/\|F^\dagger\|$ is the smallest. Let $\Delta H := P'\cdot\Delta A\cdot P = \Delta H'$ and $\begin{bmatrix}\Delta V\\\Delta K\end{bmatrix} := Q'\cdot\Delta F\cdot P$ noting that ΔV need not be diagonal. Substituting these into the equation ΔF must satisfy turns it into an equation $V\cdot\Delta V + \Delta V'\cdot V + \Delta V'\cdot\Delta V + \Delta K'\cdot\Delta K = \Delta H$ that ΔV and ΔK must satisfy. Note that $\|\Delta H\| = \|\Delta A\| \approx \text{æ}\cdot\|F\|^2$ and $\|\Delta F\| = \left\|\begin{bmatrix}\Delta V\\\Delta K\end{bmatrix}\right\|$. A first estimate $\Delta_1 F$ that ignores second-order terms must set $\Delta K := O$ to minimize $\|\Delta F\|$ after deducing that $\Delta V_{ij} = \Delta V_{ji} \approx \Delta H_{ij}/(v_i + v_j)$. Then a first estimate is $\Delta_1 F = Q\cdot\begin{bmatrix}\Delta V\\O\end{bmatrix}\cdot P'$. Its $\|\Delta_1 F\|$ can get roughly as big as $\frac{1}{2}\text{æ}\cdot\|F\|^2\cdot\|F^\dagger\| \approx \frac{1}{2}\text{æ}\cdot\|F\|\cdot\kappa(F)$, enormously bigger than $\text{æ}\cdot\|F\|$, in case $\Delta V_{nn} \approx \frac{1}{2}\Delta H_{nn}/v_n \approx \frac{1}{2}\|\Delta H\|\cdot\|F^\dagger\|$, which case cannot be ruled out. Though this first-order estimate loses its validity when $\|\Delta_1 F\|$ gets too big for its square to be ignored, it still allows $\Delta_1 F$ to get big enough to obscure almost half the significant digits stored in F when $\kappa(F) \approx 1/\sqrt{\text{æ}}$. Then, because $\Delta_1 F$ need not correlate with the rounding errors in $F'\cdot g$, roundoff can corrupt results computed from A almost as badly as if the last half of the significant digits stored in the data F and g had been disregarded.

Our second estimate $\Delta_2 F$ is not much bigger than the first-order estimate but takes second-order terms fully into account. Like the first-order estimate, our second $\Delta_2 F := Q\cdot\begin{bmatrix}\Delta V\\O\end{bmatrix}\cdot P'$ wherein $\Delta V = \Delta V'$, but now $\Delta V := \sqrt{(V^2 + \Delta H)} - V$. Here $\sqrt{\dots}$ is the positive (semi)definite square root of a positive (semi)definite symmetric matrix. Now $(F+\Delta_2 F)'\cdot(F+\Delta_2 F) = A = F'\cdot F + \Delta A$ exactly. To gauge how big this second $\|\Delta_2 F\|$ cannot get we need the following inequality:

If M and W are symmetric positive definite, their positive definite square roots differ by at most $\|\sqrt{M} - \sqrt{W}\| \le \|M-W\|/(\,1/\sqrt{\|M^{-1}\|} + 1/\sqrt{\|W^{-1}\|}\,)$. To prove this inequality use the formula $\sqrt{M} = (2/\pi)\cdot\int_0^\infty (\text{ß}^2 I + M)^{-1}\cdot M\cdot d\text{ß}$. The inequality becomes equality when matrices are 1-by-1 .

Applying this inequality to Δ_2F and assuming $\|\Delta A\| \le \text{æ}\cdot\|F\|^2 \le 1/\|F^\dagger\|^2$ we find

$$\begin{aligned}\|\Delta_2F\| = \|\Delta V\| &\le \|\Delta H\|/\left(1/\sqrt{\|(V^2 + \Delta H)^{-1}\|} + 1/\|V^{-1}\| \right)\\ &\le \|\Delta H\|/\left(1/\sqrt{(1/(1/\|F^\dagger\|^2 - \|\Delta H\|))} + 1/\|F^\dagger\| \right)\\ &= \|\Delta A\|/\left(\sqrt{(1/\|F^\dagger\|^2 - \|\Delta A\|)} + 1/\|F^\dagger\| \right) \qquad \text{because } \|\Delta H\| = \|\Delta A\|\\ &\le \kappa(F)\cdot\text{æ}\cdot\|F\|/\left(1 + \sqrt{(1 - \text{æ}\cdot\kappa(F)^2)} \right) ,\end{aligned}$$

less than twice as big as the first-order estimate Δ_1F and yet valid no matter how big it gets. It cannot exceed $\sqrt{\text{æ}}\cdot\|F\|$. But this second estimate Δ_2F need not minimize $\|\Delta F\|$ though the two estimates Δ_1F and Δ_2F become indistinguishable when their squares are negligible. Whatever ΔF is minimal, its $\|\Delta F\| \le \|\Delta_2F\|$, so its $\|\Delta_2F\| \le \sqrt{\text{æ}}\cdot\|F\|$ too.

A third estimate Δ_3F will minimize $\|\Delta F\|_f^{\,2}$. Recourse to *Lagrange Multipliers* implies that a minimizing ΔF is $\Delta F = F\cdot\Delta X$ for some symmetric array $\Delta X = \Delta X'$ derived from the Lagrange multipliers and satisfying $(I+\Delta X)\cdot F'\cdot F\cdot(I+\Delta X) = A$. Let $Y := \sqrt{F'\cdot F}$ be the positive definite square root so that $\|\Delta F\| = \|Y\cdot\Delta X\|$ and $\|\Delta F\|_f = \|Y\cdot\Delta X\|_f$; then $\Delta X := Y^{-1}\cdot(\sqrt{Y\cdot A\cdot Y} - F'\cdot F)\cdot Y^{-1}$ must satisfy the previous sentence’s equation for every matrix square root $\sqrt{Y\cdot A\cdot Y}$. We choose the positive (semi)definite square root because it is easily proved to minimize $\|\Delta F\|_f = \|Y\cdot\Delta X\|_f$. Thus can the smallest $\Delta_3F = F\cdot\Delta X$ be computed, though only with extravagantly extra-precise arithmetic. Arithmetic less extravagantly extra-precise suffices to compute it via the coordinate system provided by the aforementioned singular value decomposition of F : Starting from $F = Q\cdot V\cdot P'$ and $\Delta H := P'\cdot\Delta A\cdot P$ set $\Delta_3F := Q\cdot\left(\sqrt{\left(V\cdot(V^2 + \Delta H)\cdot V \right)} - V^2\right)\cdot V^{-1}\cdot P'$.

How big can this third estimate Δ_3F not get? Recalling that $\|\Delta H\| = \|\Delta A\| \le \text{æ}\cdot\|F\|^2$, we find

$$\begin{aligned}\|\Delta_3F\|_f^{\,2} = \| \sqrt{\left(V\cdot(V^2 + \Delta H)\cdot V \right)}\cdot V^{-1} - V \|_f^{\,2} &= \mathrm{Trace}\left(2V^2 + \Delta H - 2\sqrt{\left(V\cdot(V^2 + \Delta H)\cdot V \right)} \right)\\ &\le \mathrm{Trace}\left(2V^2 + \text{æ}\cdot\|F\|^2\cdot I - 2\sqrt{\left(V\cdot(V^2 - \text{æ}\cdot\|F\|^2\cdot I)\cdot V \right)} \right) \quad \text{because } \sqrt{\ldots} \text{ is monotonic}\\ &= n\cdot\text{æ}\cdot\|F\|^2 + 2\,\mathrm{Trace}\left(V^2 - \sqrt{\left(V^4 - \text{æ}\cdot\|F\|^2\cdot V^2 \right)} \right) \quad \text{wherein } n := \#\text{ columns}(F)\\ &= n\cdot\text{æ}\cdot\|F\|^2 + 2\text{æ}\cdot\|F\|^2\cdot\mathrm{Trace}\left(\left(I + \sqrt{\left(I - \text{æ}\cdot\|F\|^2\cdot V^{-2} \right)} \right)^{-1} \right)\\ &\le n\cdot\text{æ}\cdot\|F\|^2\cdot\left(1 + 2/\left(1 + \sqrt{(1 - \text{æ}\cdot\kappa(F)^2)} \right)\right) .\end{aligned}$$

Though this bound seems grossly pessimistic, it does keep $\|\Delta_3F\|_f/\|F\|$ below something of the order of $\sqrt{\text{æ}}$, so Δ_3F can affect at most the last half of the significant digits stored in $F+\Delta_3F$.

If roundoff is so gross that $\text{æ} > 1/\kappa(F)^2$, the perturbations Δ_2F and Δ_3F may fail to exist or, less likely, may exceed substantially the bounds derived for them above. These bounds are unlikely to be approached closely in any event unless rounding errors conspire or are contrived to that end.

Example:

Suppose arithmetic carries six sig. dec., so $\text{æ} = 0.000005$, and consider the example

$$F' := \begin{bmatrix} -945202 & 862444 & -892315 & -790042 & 1000000 & 249247 \\ 1000000 & -678443 & 1000000 & 514413 & -1000000 & 44972 \\ 911012 & -1000000 & 814980 & 1000000 & -1000000 & -467316 \\ 107073 & 676717 & 293322 & -985317 & 70215 & 1000000 \end{bmatrix}$$

. $||F|| \approx 3571141.5$ and the condition number $\kappa(F) := ||F||\cdot||F^\dagger|| \approx 17286.27$ rather exceeds $1/\sqrt{æ} \approx 447.2$. Rounded to six sig. dec.,

$$A := F'\cdot F + \Delta A = \begin{bmatrix} 4119730 & -3817830 & -4357270 & 1318590 \\ -3817830 & 3726930 & 3897830 & -590822 \\ -4357270 & 3897830 & 4712520 & -1862970 \\ 1318590 & -590822 & -1862970 & 2531230 \end{bmatrix}\cdot 10^6; \quad \Delta A = \begin{bmatrix} -2961938 & 4833954 & 2354176 & -3095486 \\ 4833954 & 1880398 & -1864848 & -214448 \\ 2354176 & -1864848 & 491600 & -1351436 \\ -3095486 & -214448 & -1351436 & 1942184 \end{bmatrix}.$$

Here $||\Delta A||/||A|| \approx 0.00000059 < æ/8$. However, computed perturbations

$$\Delta_2 F' \approx \begin{bmatrix} -231.915425 & 300.864252 & 11.992624 & 593.638098 & -102.423853 & 407.711362 \\ -427.847825 & 516.747158 & 22.895860 & 1019.553015 & -199.132283 & 708.358292 \\ 208.207541 & -229.537071 & -12.315385 & -454.953451 & 103.918353 & -320.402987 \\ 173.534835 & -205.714729 & -9.151738 & -407.061221 & 81.629918 & -282.693126 \end{bmatrix} \quad \text{and}$$

$$\Delta_3 F' \approx \begin{bmatrix} -1.257943 & 224.569396 & -6.460172 & 417.283503 & 63.964507 & 254.555622 \\ -229.720992 & 534.988777 & 4.589035 & 1022.628296 & -31.856636 & 671.600182 \\ 236.136074 & -326.106231 & -11.906222 & -637.618372 & 98.586326 & -437.690460 \\ 121.380572 & -232.240881 & -3.709921 & -448.040219 & 31.309494 & -297.615466 \end{bmatrix}$$

have $||\Delta_3 F||_f/||F||_f \approx 0.00046 < ||\Delta_3 F||/||F|| \approx 0.00050 < ||\Delta_2 F||/||F|| \approx 0.00052$, none of them much smaller than $\sqrt{(||\Delta A||/||A||)} \approx 0.00077$, so both $F+\Delta_2 F$ and $F+\Delta_3 F$ differ from F in almost half of its last six sig. dec. Thus the bounds for $||\Delta F||$ derived above turn out to be approachable.

The computations of $\Delta_2 F$ and $\Delta_3 F$ above were carried out in MATLAB 5.2 on an Apple Macintosh Quadra 950, and confirmed in MATLAB 6.5 on a Wintel PC; but the same program run in MATLAB 5.2 on an Apple Power Mac and in MATLAB 7.1 on that Wintel PC produced utterly inaccurate estimates of $\Delta_3 F$ for lack of extra-precisely accumulated matrix products. This lack has been discussed in http://www.cs.berkeley.edu/~wkahan/MxMulEps.pdf and …/Mindless.pdf .