UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

METODOS NUMERICOS PARA LA RESOLUCION DE MINIMOS CUADRADOS POR EL METODO DE GRAM-SCHMITO

TESIS PROFESIONAL

QUE PARA OBTENER EL TITULO DE:

A C T U A R I A

P R E S E N T A :

ANA MARIA HERNANDEZ MARTINEZ

Director de Tesis: José Luis Navarro Urrutia

MEXICO, D. F. 1990

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

I N T R O D U C C I O N

Un problema común que nos encontramos en el laboratorio de computación es el de encontrar las soluciones a problemas de Mínimos Cuadrados.

Estos problemas surgen en una gran variedad de áreas y en diferentes contextos. Los problemas de mínimos cuadrados son en particular difíciles de resolver dado que frecuentemente involucrán gran cantidad de datos y están mal condicionados por propia naturaleza. En este trabajo consideraremos métodos numéricos estables para manejar este problema. Nuestro instrumento básico es una descomposición de la matriz, basada en las trasformaciones ortogonales de GRAM-SCHMIDT.

Dado que un análisis general de la condición de los problemas de mínimos cuadrados está dada.

La influencia de los errores de redondeo se estudiará en detalle por una versión modificada de la ortogonalización de GRAM-SCHMIDT.

Para obtener una factorización A = QR de una matriz A de m X n dada, donde R es una matriz triangular superior

y $Q^{t} Q = I$. Sea x el vector el cual minimiza $\left\| b - Ax \right\|_2$

y r = b - Ax. Esto muestra que si los productos internos son acumulados en doble precisión entonces, los errores de las x y r calculadas son menores que los errores resultantes de alguna perturbación inicial simultanea δA, δb tal que :

$$\left\| \delta A \right\|_E \leq \left\| A \right\|_E 2^{-3t/2} \left\| \delta b \right\|_2 \leq \left\| b \right\|_2 2^{-t} n^{3/2}$$

unidades de maquina.

La no reortogonalización es necesaria y el resultado es independiente de la estrategia de pivoteo usada.

C A P I T U L O I

D E F I N I C I O N E S B A S I C A S

Sea A una matriz real de m x n, de rango n donde m ≥ n y b un vector real de m x 1. Entonces existe una única x la cual resuelve el problema de mínimos cuadrados a minimizar.

$$\min_{x} \| b - Ax \|_2$$

Y es bien sabido que la solución x satisface la condición.

$$A^T (b - Ax) = 0$$

es decir el vector residual r = b - Ax, es ortogonal a las columnas de A, entonces se sigue que podemos calcular x de la ecuaciones normales.

$$A^T Ax = A^T b$$

Ahora definimos siguiendo a Bauer, la condición de la matriz rectangular A.

Una definición natural de la condición de una matriz es el radio de su máximo y mínimo, con respecto a un par de normas.

Algunas veces, los renglones ó columnas de una matriz son grandes, solo para algunos factores, pero en cualquier caso escalonar es computacionalmente trivial, por lo que el problema de minimizar la condición por escalonamiento óptimo tiene varias aplicaciones, lo que se desea es determinar el máximo y el mínimo para poder derivar los límites superior e inferior y para ello es necesario conocer la dirección de escalonamiento al mínimo ó último que se le acerque.

Ahora bien, si queremos mostrar condiciones para que una matriz esté optimamente escalonada.
El problema está completamente resuelto por la condición subordinada a un par de normas máximas y así cada una de las matrices A y A^{-1} tienen una distribución de signo en bloques cuadriculares, (es decir son cuadradas), sin esta condición los últimos límites superior e inferior junto con factores apropiados de escalonamiento son obtenidos de la siguiente manera.

Sea B_1 y B_2, dos espacios vectoriales normados, con normas.

$$|| x ||_i \quad \text{para} \quad x \in B_i \quad \text{con } i = 1,2.$$

Para una función (mapeo) $A : B_1 \longrightarrow B_2$, el máximo, el cual es una norma, y el mínimo están definidos por

$$\text{máx}(A) = \sup_{x \neq 0} \frac{|| Ax ||}{|| x ||} \qquad \text{donde } || x || = 1$$

$$\text{mín}(A) = \inf_{x \neq 0} \frac{|| Ax ||}{|| x ||}$$

Entonces una condición subordinada a las normas se puede definir por

$$\text{Cond}(A) = \frac{\text{máx } (A)}{\text{mín } (A)}$$

si el mín (A) > 0, y si es cero entonces del cond (A) = ∞

el condicional puede ser totalmente finito para una A rectangular. Sin embargo con(A) = máx (A) máx (A^{-1}), si A^{-1} existe (asumiendo que A es una transformación rectangular).

Existen algunas propiedades básicas de la función condicional las cuales son validas independientemente de las normas fundamentales.

1).- Cond (A) > 1

2).- si el cond (A) = 1, entonces $|| Ax || = \delta || x ||$ para una cierta $\delta > 0 \in R$ donde $\delta || A ||$ y $|| A || = \sup(||Ax||$ tal que $x \in A$)

3).- $\text{cond}(A^{-1}) = \text{cond}(A)$

4).- $\text{cond}(AB) = \text{cond}(A)\,\text{cond}(B)$

La condición 1) se cumple, ya que $\text{cond}(A) = \dfrac{\text{máx}(A)}{\text{mín}(A)}$, y

puesto que el ínfimo no excede al supremo esto significa que

$$\text{máx}(A)\ \text{mín}(A) = \delta$$

$$\delta \qquad \| Ax \| = \delta \| x \| \quad y \quad \| Ax \| = \delta$$

por lo que tenemos que la condición de una matriz rectangular A será:

$$\text{cond}(A) = \frac{\text{máx} \| Ax \|}{\| x \| = 1} \div \frac{\text{mín} \| Ax \|}{\| x \| = 1}$$

En cuanto a la propiedad número 4 se cumple siempre y cuando

$$\| AB \| \leq \| A \| \| B \|$$

El número de condición para la norma L es denotado por $\chi(A)$ 2

En el capitulo VII se mostrará que bajo algunas restricciones $\chi(A)$ puede ser considerada como un número condición aproximado, para el problema de calcular x. De otro modo.

$$\chi(A^T A) = \chi^2(A)$$

Esto muestra que en general usando aritmetica binaria de t-digitos no podremos obtener totalmente una solución aproximada a

$$A^T Ax = A^T b$$

Dado que $x(A) \leq 2^{t/2}$

Ahora sea $B = AS$, donde S es cuadrada y no singular, (es decir tiene inversa), entonces de.

$$A^T (b - Ax) = 0$$

se sigue que $B^T (b - Ax) = 0$ y así las ecuaciones

$$B^T Ax = B^T b$$

pueden ser empleadas en lugar de las ecuaciones normales. Una cuestión natural es preguntarse si podemos elegir B de tal manera que

$$x(B^T A) = x(A)$$

Esto es realmente posible, dado que las columnas de A son linealmente independientes y tenemos una factorización de

A, tal que $A = QR$. Donde $Q^T Q = I$ además de que Q es ortogonal R es triangular superior, por otro lado cada renglón de R y cada columna de Q están determinados de manera única, salvo un factor escalar de modulo 1

Eligiendo $B = AR^{-1} = Q$, la matriz $B^T A = Q^T QR = R$ es triangular.
De esta manera las ecuaciones

$$B^T Ax = B^T b$$

se resuelven facilmente por sustitución hacia atrás. Además la condicion

$$x(B^T A) = x(A)$$

se satisface dado que,

$$\kappa(R) = \kappa(QR) = \kappa(A)$$

Esta factorizacion se puede obtener de dos maneras diferentes y han sido propuestas para resolver los problemas de minimos cuadrados. Un camino es eliminar los elementos de la subdiagonal de A por una serie de matrices hermitianas elementales.

$$H_k^{(k)} = \begin{pmatrix} I - 2w_k & w_k \end{pmatrix}^T, \text{ donde } k = 1,2,\ldots, n.$$

$$H_n \ldots H_2 H_1 A = Q A = R$$

Recordando una matriz Hermitiana, es aquella que es igual a su transpuesta conjugada, se denota com A^H un ejemplo tipico es :

$$A = \begin{bmatrix} 2 & 3 - 3i \\ 3 + 3i & 5 \end{bmatrix} = A^H$$

Notese que las entradas de la diagonal son números reales; no deben alterarse durante el proceso de conjugación. Cada entrada fuera de la diagonal corresponde a su reflejo al otro lado de la diagonal principal, y las dos són complejos conjugados entre si.

En cada caso $a_{ij} = a_{ji}$, y este ejemplo ilustrará claramente las cuatro propiedades de las matrices hermitianas.

se satisface dado que,

$$\kappa(R) = \kappa(QR) = \kappa(A)$$

Esta factorización se puede obtener de dos maneras diferentes y han sido propuestas para resolver los problemas de mínimos cuadrados. Un camino es eliminar los elementos de la subdiagonal de A por una serie de matrices hermitianas elementales.

$$H_k^{(k)} = \left(I - 2w_k w_k^T \right), \text{ donde } k = 1,2,\ldots, n.$$

$$H^{(n)} \ldots H^{(2)} H^{(1)} A = Q^T A = R$$

Recordando una Matriz Hermitiana, es aquella que es igual a su transpuesta conjugada, se denota con A^H un ejemplo tipico es :

$$A = \begin{bmatrix} 2 & 3 - 3i \\ 3 + 3i & 5 \end{bmatrix} = A^H$$

Notese que las entradas de la diagonal son números reales; no deben alterarse durante el proceso de conjugación. Cada entrada fuera de la diagonal corresponde a su reflejo al otro lado de la diagonal principal, y las dos son complejos conjugados entre si.

En cada caso $a_{ij} = a_{ji}$, y este ejemplo ilustrará claramente las cuatro propiedades de las matrices hermitianas.

Ahora al exponer esas cuatro propiedades, es necesario insistir otra vez en que igualmente se aplican a las matrices simétricas reales. Estas últimas representan un caso particular de matrices hermitianas, y el caso más importante.

Propiedad I.- Si $A = A^H$, entonces para todos los vectores complejos x, $x^H Ax$ es real.

Cada entrada de A contribuye a la expresión $x^H Ax$.

$$x^H Ax = \begin{bmatrix} \bar{u} & \bar{v} \end{bmatrix} \begin{bmatrix} 2 & 3 - 3i \\ 3 + 3i & 5 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

$$= 2\,\bar{u}\,u + 5\,\bar{v}\,v + (3 - 3i)\,\bar{u}\,v + (3 + 3i)\,u\,\bar{v}.$$

cada uno de los términos diagonales es real ya que

$2\,\bar{u}\,u = 2\,|u|^2$ y $5\,\bar{v}\,v = 5\,|v|^2$. Los términos fuera de la diagonal son complejos conjugados entre sí, así que se combinan para dar el doble de la parte real de $(3 - 3i)\,\bar{u}\,v$, por lo tanto, la expresión $x^H Ax$ es real.

Como una demostración general, podemos calcular $(x^H Ax)^H$. Obtendríamos el conjugado de la matriz $x^H Ax$ de 1 por 1, pero de hecho obtendríamos el mismo número :

$(x^H Ax)^H = x^H A^H (x^H)^H = x^H Ax$. Así el número debe ser real.

Propiedad II.- Cada valor propio de una matriz hermitiana es real.

Demostración, suponemos que r es un valor propio y que x es su vector propio correspondiente distinto de cero: $Ax = rx$.

El truco es multiplicar por x^H, $x^H Ax = \tau x^H x$. Por la propiedad uno, el lado izquierdo es real, y como $x \neq 0$, el lado derecho $x^H x = \|x\|^2$ es real y positivo por lo tanto, τ debe ser real. En nuestro ejemplo.

$$A - \tau I = \begin{bmatrix} 2-\tau & 3-3i \\ 3+3i & 5-\tau \end{bmatrix} = \tau^2 - 7\tau + 10 - (3-3i)^2$$

$$= \tau^2 - 7\tau - 8 = (\tau - 8)(\tau + 1)$$

Propiedad III.- Los valores propios de una matriz hermitiana, en caso de corresponder a valores propios diferentes, son ortogonales entre sí

Nuevamente la demostración comienza con la información dada, $Ax = \tau x$ y $Ay = \mu y$ con $\tau \neq \mu$, y requiere una pequeña treta. Normalmente la transpuesta hermitiana de $Ax = \tau x$ es $x^H \bar{A} = \tau x^H$, pero como $\bar{A} = A^H$ y τ es real, por la propiedad II, esto es en realidad $x^H A = \tau x^H$.

Multiplicando por y el lado derecho de esta ecuación y por x^H el lado izquierdo de la otra, tenemos que :

$$x^H Ay = \tau x^H y \qquad y \qquad x^H Ay = \mu x^H y$$

Por lo tanto, $\tau x^H y = \mu x^H y$, y como $\mu \neq \tau$, concluimos que $x^H y = 0$: x es ortogonal a y. En el ejemplo, con $Ax=8x$ y $Ay = -y$, los vectores propios se calculan de la manera usual, de.

$$(A - 8I)x = \begin{bmatrix} -6 & 3 - 3i \\ 3 + 3i & -3 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad x = \begin{bmatrix} 1 \\ 1 + i \end{bmatrix}$$

$$(A + I)y = \begin{bmatrix} 3 & 3 - 3i \\ 3 + 3i & 6 \end{bmatrix} \begin{bmatrix} y1 \\ y2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad y = \begin{bmatrix} 1 - i \\ -1 \end{bmatrix}$$

Estos dos vectores propios son ortogonales :

$$x^H y = \begin{bmatrix} 1 & 1 - i \end{bmatrix} \begin{bmatrix} 1 - i \\ -1 \end{bmatrix} = 0$$

Claramente cualesquiera múltiplos $x \div \alpha$ e $y \div \beta$ servirán igualmente como vectores propios. Si la A original es real y simétrica, entonces sus valores propios y vectores propios son reales; además de que una matriz simétrica real puede diagonalizarse mediante una matriz ortogonal Q. En analogía con las matrices ortogonales reales que satisfacen $Q^T Q = I$ y, por lo tanto, $Q^T = Q^{-1}$, la propiedad de tener columnas ortonormales se traduce en:

$$U^H U = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & x_1^H & - \\ - & x_2^H & - \\ & \vdots & \\ - & x_n^H & - \end{bmatrix} = I \quad o \quad U^H = U^{-1}$$

Como en el caso complejo necesitamos de un nuevo nombre y símbolo para una matriz de columnas ortonormales : le llamaremos matriz unitaria y se denotará por U.

Esto nos conduce a la última propiedad especial de las matrices hermitianas.

Propiedad IV.- Si $A = A^H$, entonces existe una matriz diagonalizante que además es unitaria S=U. Sus columnas son ortonormales y

$$U^{-1} A U = U^H A U = \Gamma$$

Por lo tanto cualquier matriz hermitiana puede descomponerse en :

$$A = U \Gamma U^H = \tau_1 \times_1 \times_1^H + \tau_2 \times_2 \times_2^H + \ldots + \tau_n \times_n \times_n^H$$

Esta descompocisión se conoce como teorema espectral. Expresa que A es una combinación de las proyecciónes unidimensionales $\times_i \times_i^H$ que son precisamente como las proyecciónes a A^T.

Descomponen cualquier vector b en sus componentes

$p_i = \times_i (\times_i^H b)$ en las direcciones de los vectores propios unitarios que constituyen un conjunto de ejes mutuamente perpendiculares. Estas proyecciónes individuales p están ponderadas por las $\times_i$ y reagrupadas para formar :

$$a b = \tau_1 \times_1 (\times_1^H b) + \ldots + \tau_n \times_n (\times_n^H b)$$

Si cada $\tau_i = 1$, hemos reagrupado b mismo, $A = U\,I\,U^H$ es la identidad. En cualquier caso podemos verificar los resultados anteriores directamente de la multiplicación de matrices :

$$Ab = U\,\Pi\,U^H\,b = \begin{bmatrix} & & \\ x_1 & \cdots & x_n \\ & & \end{bmatrix} \begin{bmatrix} \tau_1 & & \\ & \ddots & \\ & & \tau_n \end{bmatrix} \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_n \end{bmatrix} \begin{bmatrix} b \end{bmatrix} =$$

$$= \begin{bmatrix} & & \\ x_1 & \cdots & x_n \\ & & \end{bmatrix} \begin{bmatrix} \tau_1 x_1^H b \\ \vdots \\ \tau_n x_n^H b \end{bmatrix} = \tau_1 x_1 x_1^H b + \ldots + \tau_n x_n x_n^H b$$

Regresemos a nuestro ejemplo, ambos vectores propios tienen longitud $\sqrt{3}$ y una normalización produce la matriz que diagonaliza A :

$$U = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1-i \\ 1+i & -1 \end{bmatrix} \qquad U^{-1} A\,U = \begin{bmatrix} 8 & 0 \\ 0 & -1 \end{bmatrix}$$

entonces la descomposicion espectral.

$$U\,\Pi\,U^H = \tau_1 x_1 x_1^H + \tau_2 x_2 x_2^H \quad \text{se transforma en :}$$

$$\frac{8}{3}\left[\begin{array}{c} 1 \\ \\ i+i \end{array}\right]\left[\begin{array}{ccc} & 1, & 1-i \end{array}\right] - \frac{1}{3}\left[\begin{array}{c} i-i \\ \\ -1 \end{array}\right]\left[\begin{array}{ccc} & i+i, & -1 \end{array}\right],$$

$$= \left[\begin{array}{cc} 2 & 3-3i \\ 3+3i & 5 \end{array}\right] = A$$

Una estimación de error detallada para el algoritmo A = QR está dada por Wilkinson y la aplicación a problemas de mínimos cuadrados está dada por Golub.

Un segundo camino es obtener la factorización por medio del proceso de ortogonalizacion de Gram - Schimdt a las columnas de A y esto siempre se ha sabido que tiene poca estabilidad numérica. En estas notas haremos uso de un versión ligeramente modificada del proceso de Gram - Schmidt, la cual es equivalente a la eliminación con conbinación de renglones de peso, esto será aplicado a la solución de problemas de mínimos cuadrados lineales

La propagación de errores de redondeo será estudiada en detalle y en particular una estimación de la desviación de ortogonalidad de la Q calculada y se obtendrá una cota de error para la solución calculada.

C A P I T U L O II

DESCRIPSION DEL ALGORITMO

Se tratará de explicar la importancia de la ortogonalidad al resolver problemas de mínimos cuadrados, ya que esta tiene una importancia y relación con la intuición que va más allá de los mínimos cuadrados.

Cada vez que se piensa en el plano x - y ó en el espacio tridimensional, la imaginación nos da la figura de un conjunto de ejes coordenados, nos proporciona un punto de referencia que llamamos origen, pero además los ejes coordenados son siempre ortogonales, al elegir un conjunto de ejes ortogonales lo que estamos haciendo es elegir una base para el plano x - y.

Si la idea es convertir cada construcción geométrica en un cálculo algebraico y es necesario una base ortogonal para que esos cálculos sean sencillos existe una especialización que hace esa base casi óptima.

Comensemos con un conjunto de vectores mutuamente ortogonales y normalizandolos los volveremos unitarios, esto es que cada v del conjunto se divida por su propia longitud y es remplazado por $v/\|v\|$. Con esto se logra que una base ortogonal se convierte en ortonormal.

Por lo que decimos que una base $v_1, \ldots, v_n$ es ortonormal si :

$$v_i^t v_j = \begin{cases} 0 & \text{si } i \neq j \text{ lo cual garantiza la ortogonalidad} \\ 1 & \text{si } i = j \text{ lo cual garantiza la normalización} \end{cases}$$

El ejemplo más importante es la base canónica donde los vectores en el plano x - y en R^n además de ser perpendiculares los ejes són unitarios los v_i.

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad \ldots \quad e_n = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}$$

Puesto que esta no es la unica base ortogonal ya que se pueden rotar el conjunto de ejes sin cambiar los angulos rectos que formen será necesario introducir matrices ortogonales ó de rotación. En el caso de los vectores canonicos e_i no se encuentran dentro del subespacio R^n considerado, no será tan claro que podamos encontrar una base ortonormal, por lo que se verá que siempre es posible encontrar una de tales bases ya que siempre existe y que es posible construirla a partir de cualquier base de una manera sencilla. Este procedimiento que logrará convertir cualquier conjunto de ejes oblicuos en uno perpendicular se le conoce con el nombre de GRAM-SCHMIDT.
Los temas basicos de este capitulo són :

1) .- La solución a Ax = b por minimos cuadrados cuando las columnas de A són ortonormales.

2).- Definición y propiedades de las matrices ortogonales.

3).- El proceso de Gram - Schimdt y su interpretación como una nueva factorización de matrices.

Proyecciones y Minimos Cuadrados : Caso Ortonormal

Dada una matriz A de m x n y suponiendo que sus columnas són ortonormales se ve claramente que estas columnas són independientes por lo que podemos conocer la matriz de proyección sobre el espacio columna y la solucion $\bar{x}$ en minimos cuadrados : $P = A (A^t A)^{-1} A^t$ y $\bar{x} = (A^t A)^{-1} A^t b$.

Estas fórmulas son fácilmente expresadas ya que la matriz $A^t A$ es la identidad y son válidas en el caso de columnas ortonormales.

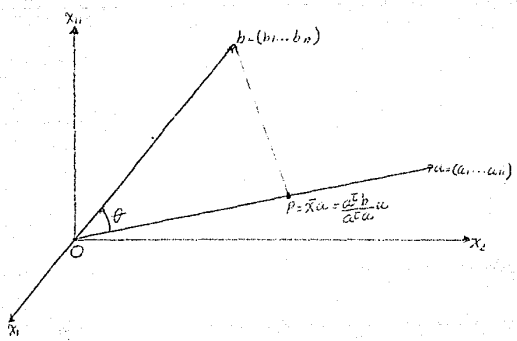
En el caso de columnas ortonormales para A se tiene :

$$\begin{bmatrix} \text{---}\ a_1^t\ \text{---} \\ \text{---}\ a_2^t\ \text{---} \\ \vdots \\ \text{---}\ a_n^t\ \text{---} \end{bmatrix} \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I$$

Con esto logramos que el algebra se simplifica ya que P y $\bar{x}$ se cambian por:

$P = AA^t$ y $\bar{x} = A^t b$.

Pero además es necesario simplificar la geometría y esto es posible gracias a que los ejes son perpendiculares, pues la proyección al espacio queda como proyección en cada eje.



La matriz de proyección es :

$$P = a_1 a_1^t + \ldots + a_n a_n^t$$

$$P = AA^t b = \begin{bmatrix} a_1 & a_2 & \ldots & a_n \end{bmatrix} \begin{bmatrix} a_1^t b \\ \cdot \\ \cdot \\ \cdot \\ a_n^t b \end{bmatrix} = a_1 a_1^t b + \ldots + a_n a_n^t b$$

Finalmente observamos que al hacer lo anterior el termino $(A^t A)^{-1}$ desaparece en el cálculo de P y de $\bar{x}$ y, P es la suma de las n-proyecciones por separado.

Hasta ahora se tienen a saber cinco ecuaciones que són :

1).- $Ax = b$ La ecuación dada probablemente inconsistente

2).- $A^t A\bar{x} = A^t b$ Las ecuaciónes normales para $\bar{x}$.

3).- $p = A\bar{x}$ La proyección de b sobre el espacio columna de A

4).- $P = A (A^t A)^{-1} A^t$ La matriz de proyección que nos da $p = P b$

5).- $\bar{x} = A^t b$ y $p = AA^t = a_1 a_1^t + \ldots + a_n a_n^t$ El caso en el que A tiene columnas ortonormales.

Matrices Ortogonales.

Se le llama a una matriz ortogonal a aquella que es

cuadrada con columnas ortonormales. Usando Q para denotar una matriz ortogonal y $q_1 \ldots q_n$ para denotar sus columnas, sabemos que $Q^t Q = I$ ya que es

$$Q^t Q = \begin{bmatrix} q_1^t \\ q_2^t \\ \vdots \\ q_n^t \end{bmatrix} \begin{bmatrix} q_1 & q_2 & \ldots & q_n \end{bmatrix} = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix} = I$$

La ecuación anterior es una replica de $A^t A$ solo que está última la matriz A no necesariamente es cuadrada.
La ventaja que se obtiene al tener una Q cuadrada con columnas independientes es que podemos asegurar que tiene el rango total, es decir r = n y además que es invertible por lo que si Q^t es una inversa izquierda, entonces es la inversa por lo que tambien es inversa derecha y $Q Q^t = I$
Las matrices ortogonales cuentan con las siguientes propiedades :

1.- $Q^t Q = I$

2.- $Q Q^t = I$

3.- $Q^t = Q^{-1}$

Q además de tener las columnas ortonormales sus filas tambien lo són por lo que se obtiene que si Q es ortogonal Q^t tambien lo es.

Lo anterior se sigue directamente de realizar $QQ^t = I$ puesto que lo que se está haciendo al realizar este producto es el producto interno de cada fila y como el resultado es la matriz identidad las filas son ortonormales no obstante las direcciones de las filas apunten en direcciones completamente diferentes de las columnas por lo que se puede afirmar que es suficiente con que las columnas sean perpendiculares para que automáticamente las filas también lo sean.

La Ortogonalización de Gram – Schmidt.

Hasta este momento se ha dicho que para resolver el problema de mínimos cuadrados $Ax = b$ se obtiene con mayores ventajas si las columnas són ortonormales ya que en cada caso se ha dicho que es fácilmente resuelto si las columnas son ortonormales por lo que a continuación daremos un método para hacerlas ortonormales.
Este método es posible realizarlo antes de cualquier aplicación y la manera más fácil de visualizarlo es cuando están involucrados solo dos vectores.
Lo que se busca es producir de dos vectores independientes dados a y b dos perpendiculares v_1 y v_2.

Lo primero que se debe hacer para lograr lo anterior es que el primer vector tenga la dirección de a esto es $v_1 = a$.

Nuestro problema es encontrar un segundo vector que sea perpendicular y para ello supongase que se tiene b un punto en el espacio n – dimencional y queremos encontrar la recta en la dirección del vector a, buscamos entonces el punto p

de la recta más cercana a b por lo que la recta que une b con p es perpendicular al vector original a. Lo que se necesita es encontrar la proyección del punto p, este debe ser un múltiplo es decir $p = \bar{x}a$ del vector a, por lo que el problema es el cálculo de $\bar{x}$ pero para realizar este cálculo
nos valdremos del hecho de que la recta desde b hasta el punto más cercano $p = \bar{x}a$ es perpendicular al vector a es

decir $\bar{x} = a\ b \div a\ a$ por lo que finalmente se obtiene que

$$p = \frac{a\ ba}{a\ a}$$

que es la proyección del punto b sobre la recta generada por el vertor a: (ver grática al final del capitulo)

El vector v - p es perpendicular al vector a por lo que el segundo eje irá en la dirección de :

$$v_2 = b - p = b - \frac{a\ b}{a\ a}\, a = b - \frac{v_1\ b\ v_1}{v_1\ v_1}$$

y además se verifica que:

$$v_1\ v_2 = v_1\ b - v_1\ b = 0$$

Para obtener directamente el proceso de Gram - Schmidt supongase que existe un tercer vector c independiente. Substraemos las componentes de c en las dos direcciones v_1 y v_2 que ya hemos encontrado.

$$v_3 = c - \frac{v_1\ c}{v_1\ v_1}\, v_1 - \frac{v_2\ c}{v_2\ v_2}\, v_2$$

Nuevamente se tiene que v_3 es perpendicular a v_1 y v_2

pues lo que se está haciendo es substraer de c sus componentes en el plano generado por a y b.
La proyección de c sobre el plano es la suma de sus proyecciones en los ejes v_1 y v_2 además con esta construcción se garantiza que v_3 no puede ser el vector cero puesto que esto haría que c estuviese dentro del plano formado por a y b contradiciendo la independencia lineal de a, b y c. Hasta ahora los vectores v_i son solo ortogonales

nos falta hacerlos otonormales esto es unitarios, para lograrlo dividiremos cada uno por su longitud.

$$q_1 = \frac{v_1}{\| v_1 \|}, \quad q_2 = \frac{v_2}{\| v_2 \|}, \quad q_3 = \frac{v_3}{\| v_3 \|}$$

Resumiendo el proceso anterior podemos decir que cualquier conjunto $a_1 \ldots a_n$ de vectores independientes es posible hacerlo un conjunto $v_1 \ldots v_n$ ortogonales mediante el método de Gram - Schmidt : primero $v_1 = a_1$ despues cada v_i es ortogonal a las $v_1, \ldots, v_{n-1}$ precedente por lo que

finalmente se obtiene que :

$$v_i = a_i - \frac{t}{\frac{v_1}{t}} v_1 - \ldots - \frac{t}{\frac{v_{i-1}}{t}} v_{i-1} \qquad (1)$$

Si observamos lo anterior se podrá ver que para cada v_j el subespacio que es generado por las $a_1, \ldots, a_j$ originales es generado también por $v_1, \ldots, v_j$ por lo que los vectores finales :

$$q_i = \frac{v_i}{\| v_i \|} \quad \text{son ortonormales.}$$

EJEMPLO : Supongase que los vectores originales son :

$$a_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad a_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad a_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

entonces $a_1 = v_1$ y v_2 se calcula así :

$$v_2 = b - p = b - \frac{a^t b}{a^t a} a = b - \frac{v1^t b}{v1^t v1} v_1$$

$$\frac{a2^t v1}{v1^t v1} = \frac{1}{2}, \qquad v2 = a2 - \frac{1}{2} v1 = \begin{bmatrix} ½ \\ -½ \\ 1 \end{bmatrix}$$

El tercer eje perpendicular viene de la :

$$v3 = c - \frac{v1^t C}{v1^t v1} v_1 - \frac{v2^t c}{v2^t v2} v2$$

Entonces :

$$\frac{a3^{t}\ v1}{v1^{t}\ v1} = \frac{1}{2}, \qquad \frac{a3^{t}\ v2}{v2^{t}\ v2} = \frac{1/2}{6/4} = \frac{1}{3},$$

$$v3 = a3 - \frac{1}{2} v1 - \frac{1}{3} v2 = \begin{bmatrix} -2/3 \\ 2/3 \\ 2/3 \end{bmatrix}$$

Los vectores ortonormales finales son:

$$q_1 = \frac{v1}{||\ v1\ ||} = \sqrt{1/2} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \qquad q_2 = \frac{v2}{||\ v2\ ||} = \sqrt{2/3} \begin{bmatrix} -1/2 \\ -1/2 \\ 1 \end{bmatrix},$$

$$q_3 = \frac{v3}{||\ v3\ ||} = \sqrt{3/4} \begin{bmatrix} -2/3 \\ 2/3 \\ 2/3 \end{bmatrix}$$

Lo cual implica que :

$$Q = \begin{bmatrix} \sqrt{\frac{1}{2}} & \frac{\sqrt{2/3}}{2} & \frac{2\sqrt{3/4}}{3} \\ \sqrt{\frac{1}{2}} & \frac{\sqrt{2/3}}{2} & \frac{2\sqrt{3/4}}{3} \\ 0 & \frac{\sqrt{2/3}}{2} & \frac{2\sqrt{3/4}}{3} \end{bmatrix}$$

y es posible recobrar las columnas originales de A a partir de los vectores q esto implica que deshaciendo las ecuaciones para v se tiene :

$$a_1 = v_1 \qquad a_1 = \sqrt{2}\, q_1$$

$$a_2 = 1/2\, v_1 + v_2 \qquad a_2 = \sqrt{1/2}\, q_1 + \sqrt{3/2}\, q_2$$

$$a_3 = 1/2\, v_1 + 1/3\, v_2 + v_3 \qquad a_3 = \sqrt{1/2}\, q_1 + \sqrt{1/6}\, q_2 + \sqrt{4/3}\, q_3$$

Y queda finalmente :

$$A = Q\, R$$

Donde R es una matriz trianguiar superior y Q una matriz ortonormal.

Por lo que finalmente se tiene :

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \sqrt{1/2} & \frac{\sqrt{2/3}}{2} & -\frac{2\sqrt{3/4}}{3} \\ \sqrt{1/2} & \frac{\sqrt{2/3}}{2} & \frac{2\sqrt{3/4}}{3} \\ 0 & \frac{\sqrt{2/3}}{2} & \frac{2\sqrt{3/4}}{3} \end{bmatrix}$$

$$* \begin{bmatrix} \sqrt{2} & \sqrt{1/2} & \sqrt{1/2} \\ 0 & \sqrt{3/2} & \sqrt{1/6} \\ 0 & 0 & \sqrt{4/3} \end{bmatrix}$$

El algoritmo de Gram - Schmidt es sencillo y directo, por lo que debe existir una forma sencilla de escribir el resultado final, a continuación se explicará la forma de hacerlo.

La situación es comparable con la eliminación gaussiana pues las operaciones se eligen de una manera natural conforme se van realizando. En este caso encontramos la manera correcta mediante la factorización A = LU por lo que el algoritmo de Gram - Schmidt se registra mediante una factorización diferente de la matriz A.
El truco consiste en saber como es posible recobrar las columnas originales a a partir de los vectores finales
i

q_i . Si observamos el siguiente ejemplo y despejamos las ecuaciones para v_i encontramos que :

$$a_1 = v_1 \qquad \acute{o} \qquad a_1 = \sqrt{2}\, q_1$$

$$a_2 = 1/2\, v_1 + v_2 \qquad \acute{o} \qquad a_2 = \sqrt{1/2}\, q_1 + \sqrt{3/2}\, q_2$$

$$a_3 = 1/2\, v_1 + v_2 + v_3 \qquad \acute{o} \qquad a_3 = \sqrt{1/2}\, q_1 + \sqrt{1/6}\, q_2 + \sqrt{4/3}\, q_3$$

Ahora si este conjunto de ecuaciones lo escribimos en forma matricial se tiene

$$\begin{bmatrix} & & \\ a_1 & a_2 & a_3 \\ & & \end{bmatrix} = \begin{bmatrix} & & \\ q_1 & q_2 & q_3 \\ & & \end{bmatrix} \begin{bmatrix} \sqrt{2} & \sqrt{1/2} & \sqrt{1/2} \\ 0 & \sqrt{3/2} & \sqrt{1/6} \\ 0 & 0 & \sqrt{4/3} \end{bmatrix}$$

La matriz original A está factorizada en una matriz ortogonal Q multiplicada por una matriz triangular superior R, esto es A = QR.

Las columnas de Q son los vectores ortonormales que queremos : la ecuación.

$$v_i = a_i - \frac{v_1^t a_i}{v_1^t v_1} v_1 - \ldots - \frac{v_{i-1}^t a_i}{v_{i-1}^t v_{i-1}} v_{i-1}$$

expresa a a_i como una combinación lineal de $v_1, \ldots, v_i$

y solo se tiene que reemplazar v_1 por $\|v_1\|\,q_1$, v_2 por $\|v_2\|\,q_2$ y así sucesivamente. Así a_i es una combinación de $q_1, \ldots, q_i$ no involucrando las restantes q : está es la razón por la que R es triangular superior, además de que los coeficientes de esta combinación están en la i-ésima columna de R en particular las entradas de la diagonal són distintas de cero.

En la ecuacion

$$v_i = a_i - \frac{v_1^t a_i}{v_1^t v_1} v_1 - \ldots - \frac{v_{i-1}^t a_i}{v_{i-1}^t v_{i-1}} v_{i-1}$$

v_i se remplaza por $\|v_i\|\,q_i$ de modo que la entrada de la diagonal es el coeficiente v_i diferente de cero.

R tiene entradas positivas en su diagonal y, por lo tanto está hace que sea invertible, está nos lleva al resultado más importante de este capituio.

Cualquier matriz A con columnas linealmente independientes puede ser factorizada en un producto A = QR. Donde las columnas de Q són ortonormales y R es triangular superior e invertible. Si A es cuadrada sus factores Q y R tambien lo són y entonces Q será ortogonal.

Comenzando con A = QR el problema de mínimos cuadrados Ax = b es más fácil de resolver pue sabemos que :

$$\bar{x} = (A^t A)^{-1} A^t b = (R^t Q^t Q R)^{-1} R^t Q^t b \quad \text{pero como}$$

$Q^t Q = I$ pues las columnas de Q son ortonormales tenemos que :

$$\bar{x} = (R^t R)^{-1} R^t Q^t b = R^{-1} Q^t b$$

aquí el cálculo de $\bar{x}$ sólo requiere de la multiplicación de una matriz - vector $Q^t b$ seguida de una substitución regresiva en el sistema triangular $R \bar{x} = Q^t b$. El haber ortogonalizado nos ayuda a no construir $A^t A$ y a no resolver las ecuaciones normales $A^t A \bar{x} = A^t b$ dandonos además está mayor estabilidad numérica. (Como fué mostrado en el ejemplo de la pagina No 20 de este mismo capitulo).

Gram - Schmidt Modificado:

Nuestro instrumento básico es una descomposición de la matriz A basada en la transformación de Householder; lo que nos proporciona un " paso de preparación " para el algoritmo QR modificado.

Sea A una matriz real de m x n, de rango r, y b un vector dado. Deseamos determinar un vector x tal que,

$$|| b - Ax || = \min$$

donde $|| \cdot ||$ indica la norma euclidiana.

Si m ≥ n y r < n, entonces no hay solución única. Bajo estas condiciones tenemos que,

$$|| x || = \min$$

La condición $|| b - Ax || = \min$, es una condición muy natural para problemas estadísticos y numéricos. Si m ≥ n y r = n, entonces sabemos que x satisface

$$A^t Ax = A^t b$$

Desafortunadamente, la matriz $A^t A$, frecuentemente está mal condicionada e influenciada en la mayoría de los casos por errores de redondeo.
El siguiente ejemplo ilustra bien este caso. Supongamos :

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ \epsilon & 0 & 0 & 0 & 0 \\ 0 & \epsilon & 0 & 0 & 0 \\ 0 & 0 & \epsilon & 0 & 0 \\ 0 & 0 & 0 & \epsilon & 0 \\ 0 & 0 & 0 & 0 & \epsilon \end{bmatrix}$$

Entonces

$$A^t A = \begin{bmatrix} 1+\epsilon^2 & 1 & 1 & 1 & 1 \\ 1 & 1+\epsilon^2 & 1 & 1 & 1 \\ 1 & 1 & 1+\epsilon^2 & 1 & 1 \\ 1 & 1 & 1 & 1+\epsilon^2 & 1 \\ 1 & 1 & 1 & 1 & 1+\epsilon^2 \end{bmatrix}$$

donde claramente vemos que $\epsilon \neq 0$, el rango de $A^t A$ es cinco, dado que, los eigen-valores de $A^t A$ son :

$$5 + \epsilon^2 \ , \ \epsilon^2 \ , \ \epsilon^2 \ , \ \epsilon^2 \ , \ \epsilon^2 \ .$$

Ahora si los elementos de $A^t A$ están calculados usando aritmética de doble precisión, entonces redondeamos a presición simple exacta. Ahora sea n el número más grande en la computadora, tal que $fl(1.0 + n) = 1.0$ donde $fl(\cdot)$ indica la computación de punto flotante. Entonces si

$\epsilon < \sqrt{n} \div 2$, el rango de la representación computada será 1.

Consecuentemente, no importa cuan exacta sea la solución de la ecuación lineal. Es imposible resolver las ecuaciones normales :

$$A^t A \, x = A^t b$$

Ahora veamos la descomposición de una matriz A.

Digamos que m ≥ n = r, dado que la norma de un vector es unitariamente invariante tenemos que,

$$|| b - Ax || = || c - QAx ||$$

donde c = Qb, matriz ortogonal, elegimos A tal que :

$$QA = R = \begin{bmatrix} \bar{R} \\ \hline \\ \hline \\ 0 \end{bmatrix} \Big\} \; (\; m - n \;) \; \geq \; n$$

Donde $\bar{R}$ es una matriz triangular superior, claramente

$$\hat{x} = \bar{R}^{-1} \bar{c}$$

donde $\bar{c}$ es el primero de los n componentes de C y por lo tanto

$$|| b - A\hat{x} || = \left(\sum_{j=n+1}^{m} c_j^2 \right)^{1/2}$$

dado que $\bar{R}$ es una matriz triangular superior y $\bar{R}^t \bar{R} = A^t A$

donde $\bar{R}^t \bar{R}$ es la descomposición de Choleski de $A^t A$.

Ahora bien como el camino que hemos elegido es vía las transformaciones de Hausseholder.

Sea $A = A^{(1)}$ y sean $A^{(2)}$ $A^{(3)}$ . . . $A^{(n+1)}$ definidas como sigue:

$$A^{(k+1)} = P^{(k)} A^{(k)} \quad (k = 1,n)$$

donde $P^{(k)}$ es una matriz simétrica ortogonal de la forma :

$$P^{(k)} = I - \frac{2\, v\, v^t}{\| v \|^2}$$

como v es ortogonal a v => $v^t v = 1$ a menudo se normaliza para tener una matriz ortonormal, entonces sea

$$w = \frac{v}{\| v \|}$$ Por lo tanto resulta que :

$P = P^t = P^{-1}$ por ser simétrica y ortogonal así :

$PP = (I - 2ww^t)(I - 2ww^t) = (I - 4ww^t + 4w(w^t w)w^t) = I$

y resulta que P es una matriz hermitiana elemental y como una matriz hermitiana tiene componentes reales en la diagonal

=> dado que cada entrada de P contribuye a la expresión $x^H Ax$ cada uno de los términos diagonales es real así

obteniendo el conjugado $\overline{x^H Ax}$ se tiene que :

$(x^H P x)^H = x^H P^H (x^H)^H = x^H P x$

<= si hacemos $Px = y$ donde y es numero real entonces, resulta que :

$$y = Pv = x - 2w(w^H x)$$

$$y^H y = x^H p^H p x = y^H x$$

y, $x^H y = x^H (Px)$

esto es que dadas x, y si :

$y = Px \Rightarrow x^H x = y^H y$ por lo tanto $x^H = y$, y es real

$$(k+1) \quad (k) \quad (k)$$
entonces por la definición de A $= P$ A tenemos que podemos descomponer a:

$$(n) \quad (n-1) \qquad (1)$$
Q en $P \; P \; \ldots \; P$, esto es en un producto de matrices hermitianas elementales y $Q\,A = R$ entonces por lo que tenemos que :

$$(n-1) \qquad (n) \; (n-1) \qquad (1) \qquad (k)$$
$R = A$, $Q = P \; P \; \ldots \; P$ donde P es simétrica y ortogonal, por lo tanto Q es una matriz ortonormal.

En el procedimiento de Gram - Schmidt modificado los elementos de R se calculan de un renglón en uno :

y definimos :

$$a_j^{(1)} = a_j \qquad j = 1, 2, \ldots, n$$

$$q_i = a_i^{(i)}$$

$$d_i^2 = q_i^T \, q_i$$

$$r_{ij} = \frac{a_j^{(i)T} \, q_i}{d_i^2} \qquad \text{donde } j = i+1, i+2, \ldots n$$

$$a_j^{(i+1)} = a_j^{(i)} - r_{ij} \, q_i$$

$\bar{A} = [\, A \;\vdots\; b \,]$; y lo que queremos obtener es la factorización :

$$\bar{A} = \bar{Q}\,\bar{R}$$

donde, $\bar{A}$ es una matriz de $m \times (n-1)$:

$r_{ij} = \dfrac{a_j^{(i)} q_i}{d_i^2}$ son los elementos de la matriz triangular superior $\bar{R}$; y $q_i = a_i^{(i)}$ son los elementos columna de la matriz $\bar{Q}_{m \times (n+1)}$ y demostraremos como :

$S_{(n+1)\times(n+1)}$ es una matriz diagonal cuyos elementos son los elementos de la diagonal:

Sean $d_i^2 = q_i^t q_i$ entonces la matriz $A^{(k)}$ es de la forma :

$$A^{(k)} = (q_1, \ldots, q_{k-1}, a_k^k, \ldots, a_n^k)$$

dado que $q_1 = a_1^{(i)}$

$\bar{A} = [\ A\ :\ b\ ]$; y lo que deseamos obtener es la factorización :

$$\bar{A} = \bar{Q}\,\bar{R}$$

donde, $\bar{A}$ es una matriz de m x (n + 1):

$$r_{ij} = \frac{(i)t_j}{d_i^2} \quad \text{son los elementos de la matriz triangular}$$

superior $\bar{R}$; y $q_i = a_i^{(i)}$ son los elementos columna de la matriz $\bar{Q}_{m \times (n+1)}$ y demostraremos como :

$S_{(n+1)\times(n+1)}$ es una matriz diagonal cuyos elementos son los elementos de la diagonal:

Sean $d_i^2 = q_i^t q_1$ entonces la matriz $A^{(k)}$ es de la forma :

$$A^{(k)} = (q_i , \ldots , q_{k-1} , a_k^k , \ldots , a_n^k)$$

dado que $q_1 = a_1^{(i)}$

y asumimos que :

$$\left.\begin{array}{l} \left(q_p \right)^t \; q_r = \delta_{pr} \\ \\ \left(q_p \right)^t \; a_j^{(k)} = 0 \end{array}\right| \begin{array}{l} 1 \le p, r \le k-1 \\ \\ > \\ \\ k \le j \le n \end{array}$$

en el k - ésimo paso tomamos $q_k = a_k^{(k)}$ y calculamos :

$$d_k = \left|\left| q_k \right|\right|_2^2$$

$$r_{kj} = \left(q_k \right)^t a_j^{(k)} \div d_k$$

$$a_j^{(k+1)} = a_j^{(k)} - r_{kj} q_k \qquad \text{donde } k+1 \le j \le n$$

y transformando el vector b del mismo modo nos da :

$$b = b^{(1)} \; . \; . \; . \; b^{(n+1)} = r$$

y en este paso k - ésimo calculamos para facilitar la notación :

$$y_k = \left(q_k \right)^t b^{(k)} \div d_k$$

$$b^{(k+1)} = b^{(k)} - y_k q_k$$

entonces :

$$a_k = \sum_{j=1}^{k} r_{jk} q_j$$

$$r = b - \sum_{j=1}^{n} y_j q_j$$

teniendo en cuenta que $r_{kk} = 1$.

y llegamos a la factorización $A = QR$, $b = Qy + r$.

donde $\overline{Q}^T Q = D = diag(d_1, \ldots, d_n)$

En notación matricial tenemos que :

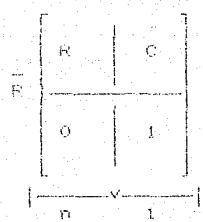
$$\underset{m \times n}{Q} \begin{bmatrix} \overline{D} \\ \hline 0 \end{bmatrix} \begin{matrix} n+1 \\ m-n-1 \end{matrix} = \underset{m \times (n+1)}{Q}$$

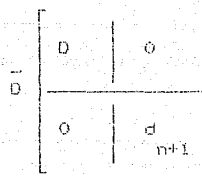
entonces :

$$\overline{A} = Q \begin{bmatrix} \overline{D} \\ \hline 0 \end{bmatrix} \overline{R}$$

$$= Q \begin{bmatrix} DR & DC \\ 0 & d_{n-1} \\ 0 & 0 \end{bmatrix}$$

tal que





entonces $||Ax - b||_2 = ||Q^t(Ax - b)||_2 = ||D(Rx - c)||_2 + d_{n+1}^2$

por lo tanto el mínimo valor de $||Ax - b||_2$ es $|d_{n+1}|$

el cual se obtiene por el vector $\bar{x}$ que satisface $R\bar{x} = c$.

notece que $x(A)$ y que $\hat{A}^t\hat{A}$ (redondeada) tiene rango 1 tal que las ecuaciones normales són singulares. El procedimiento modificado tiene tambien la ventaja de seguir facilmente una de las siguientes estrategias de pivoteo para ser usadas, podemos elegir como q' la columna k de la matriz $\hat{A}^{(k)}$ para la cual a_{kj}^{k} donde $k \le j \le n$ es

máximizada, entonces los elementos de $\bar{R}'$ deben satisfacer

si $r < 1$ si deseamos expresar a b en las menores columnas
jk
de A como sea posible podemos elegir como q_k la columna
para la cual P_{kj} sea maximizado donde:

$$P_{jk} = \left(a_j^{(k)} \right) \div b_j^{(k)} \qquad k \leq j \leq n$$

ambas $a_j^{(k)}$ y p_{kj} son fácilmente calculadas por iteración,

está es probablemente la razon por la cual el procedimiento modificado se prefiere en la practica.

Demostraremos que en la práctica la superioridad del procedimiento modificado debido a otros factores independientes de la estrategia de pivoteo, esto es un argumento con resultados experimentales.

Los procedimientos sin embargo tienen propiedades numericas totalmente diferentes cuando $n > 2$ si a está mal condicionada entonces con el clasico las columnas Q_k calculadas perderán su ortogonalidad, consecuentemente este procedimiento nunca deberá ser usado sin antes reortogonalizar lo cual incrementa grandemente la cantidad de evaluaciones.

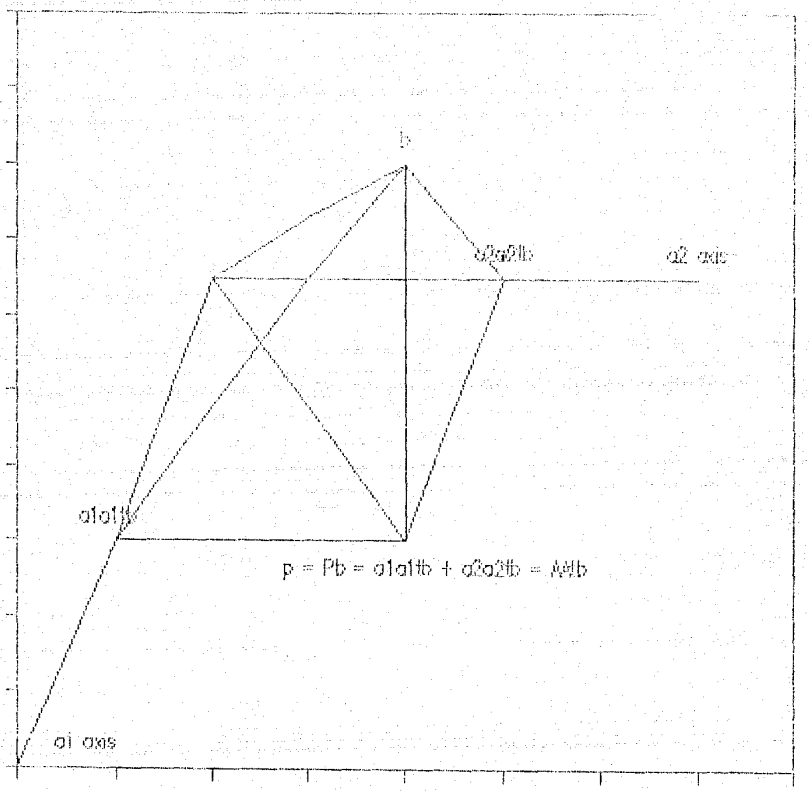
Como mostraremos la reortogonalización no es necesaria nunca cuando se usa el procedimiento modificado para resolver problemas de minimos cuadrados, el siguiente ejemplo dado nos ilustrará los diferentes aspectos de los dos precedimientos :

$$A = \begin{bmatrix} 1 & 1 & 1 \\ \epsilon & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & \epsilon \end{bmatrix}$$

Sea ϵ una cantidad pequeña tal que debido al redondeo de la operación $1 + \epsilon$ de como resultado la unidad y puede fácilmente verificarse que si no existen otros errores de redondeo entonces la desviación máxima de ortogonalidad de las columnas de Q' calculadas está dado por :

$$\text{clásico} \quad \frac{(q'_3)^t \, Q'_2}{q'_3 \, q'_2 \, q'_2} = \tfrac{1}{2} \qquad \text{modificado} \quad \frac{(q'_3)^t \, q'_1}{q'_3 \, q'_2 \, q'_1 \, q'_2} = 2/3$$

PROYECCION SOBRE UN PLANO



C A P I T U L O III

DEFINICIONES BASICAS DEL ANALISIS DE ERROR

El moderno análisis de error de redondeo originado en 1940, centró su atención en el algoritmo de eliminación Gaussiana para resolver sistemas lineales al mismo tiempo que los analistas numéricos contemplaban el inminente arribo de las maquinas computadoras electrónicas.Encarando la posibilidad de poder llevar a cabo computaciones que involucraban números muy grandes de operaciones aritmeticas y estaban muy interesados en los efectos acumulativos,que los errores de redondeo involucraban .Y centraron su atención en el algoritmo de Gauss, porque este es el problema computacional mas importante, y ademas, porque es matemáticamente muy simple y por consiguiente un punto de partida prometedor.

Muy pronto ensayaron para obtener un limite para el error en la solución calculada abandonando las mas pesimistas conclusiones. Esto fué un estímulo para tratar de investigar métodos alternativos en la espera de que ellos no sufrieran de la inestabilidad numérica de la eliminación gaussiana. Aguno de estos métodos alternativos han probado ser de gran importancia, aunque mucho menos estables que la eliminación gaussiana.

Desde el principio esos trabajos en análisis de error se encontraron a si mismos involucrados en las complejidades formales del tan llamado " acumulación de errores de redondeo " en la eliminación Gaussiana y esto tuvo un efecto adverso en la presentación del objeto de dicho análisis. Retomando la situación, dejemos por el momento el análisis de error detallado de algoritmos específicos y nos concentraremos en lugar de ello en las limitaciones inherentes al computo en presencia de errores de redondeo. Esto lleva a concluciones generales las cuales son muy ilustrativas cuando se aplican a la solución de sistemas lineales y por consiguiente seguiremos ese curso en este capitulo.

Las Operaciónes Aritméticas Básicas.

Muchos de los primeros análistas de error estaban formulados en términos de la aritmética de punto flotante, porque ninguna de las computadoras estaban bien preparadas para tener hardware de punto flotante. Esto de nuevo fué desafortunado, dado que los algoritmos en punto flotante casi invariablemente involucran un proceso continuo de

balanceo adecuado y esto complica la descripción del algoritmo computado por si mismo. Por extraño que parezca hubo una concepción popular en los años tempranos en que sería difícil poner el error de análisis en punto flotante en bases rigurosas. De hecho el análisis de error riguroso de computación en punto flotante es generalmente más simple que el de la computación de punto fijo. Dado que casi todo el software matemático es ahora expresado en términos de la aritmética de punto flotante y nos restringiremos a esto.

Asumiremos que la computación es llevada a cabo en la base β con t dígitos en la mantisa y que la unidad aritmética en la computadora acepta números normalizados y produce números normalizados como la salida de las operaciones aritméticas.

Nos referimos a este conjunto de números que puede ser exactamente representados en la computadora con números digitales. Después será asunto nuestro el uso de números con 2t dígitos (números digitales en doble precisión), pero por el momento ignoramos está posibilidad. Aunque los problemas de overflow y underflow son importantes en consideración al robustecer algoritmos, los ignoraremos aquí, y damos por hecho que el exponente es el adecuado.

Usaremos la notación fl(a * b), donde * denota cualquier operación aritmética básica, +, - , x, /, para enotar el valor de a * b. Está implícito en el uso de esta notación que a y b son números y fl(a * b) es, por definición un número digital. Es el resultado establecido por la computadora después de cualquier redondeo, corte, etc. y la renormalización puede ser necesaria. Aunque se lo de gran importancia a los intentos que están hechos para asegurar que el futuro de las computadoras tendrá unidades aritméticas realmente satisfactorias, la naturaleza de los procesos de redondeo no serán de fundamental importancia en este capítulo. Meramente se asumirá por el momento que cada una de las operaciones básicas

$$fl(a * b) = (a * b)(1 + \epsilon) \quad , | \epsilon | < k \beta^{-t} \qquad (1)$$

Donde el orden de k es la unidad y es independiente de a y b.

En algunas computadoras la relación

$$fl(a * b) = (a * b)(1 + \epsilon)$$

No siempre se toma como adición y sustracción y debemos conformarnos con resultado de la forma

$$fl(a + b) = a(1 + \epsilon_1) + b(1 + \epsilon_2) \quad , \quad | \epsilon_i | < k \beta^{-t} \quad (2)$$

Donde en general no es posible tomar $\epsilon_1 = \epsilon_2$. Por consiguiente este RESULTADO MAS DEBIL no tendría serias repercusiones para los resultados dados.

La relación

$$fl(a + b) = a(1 + \epsilon_1) + b(1 + \epsilon_2)$$

implica que el resultado calculado de cada operación individual tiene un error relativo bajo. Sin embargo hay un medio alternativo de interpretar la relación

$$fl(a * b) = (a * b)(1 + \epsilon)$$

la cual por consiguiente parece trivial, tiene sorprendentes implicaciones, las cuales nos pueden ilustrar en el caso de la multiplicación. La relación

$$fl(a * b) = (a * b)(1 + \epsilon)$$

establece que el producto calculado es el producto exacto de $a(1 + \epsilon)$ y b ó de a y, $b(1 + \epsilon)$, ó más simétricamente,

de $a(1 + \epsilon)^{1/2}$ y $b(1 + \epsilon)^{1/2}$. Dado que ϵ es pequeña podemos decir que el resultado calculado es exacto para dos operandos los cuales difieren de los dados por errores relativos chicos; al presentar el resultado en esta forma estamos retomando los errores cometidos por errores equivalentes en los datos. Por supuesto hay un infinito de caminos en los cuales esto podría ser hecho pero muchos de ellos no son de interés; así por ejemplo podríamos tomar $a+2$ y $2b(1 + \epsilon)$ pero esto no involucra perturbaciones

relativas grandes en los datos.

Análisis de Error Inverso

La interpretación que hemos dado de la expresión,

$$fl(a * b) = (a * b)(1 + \epsilon) \qquad \text{donde } |\epsilon| < k\beta^{-t}$$

Usualmente es referido como un error de análisis inverso y en general puede ser descrito en los siguientes términos.

Consideremos cualquier algoritmo calculado con elementos x_i, con $i = 1, p$, y los elementos solución y_j con $j = 1..q$

Cuando los errores de redondeo son hechos en la ejecución de los algoritmos de elementos calculados y_j no serán exactos, pero en general, será un número infinito del conjunto de elementos $x_i(1 + \epsilon_i)$ para los cuales los y_j calculados son la solución exacta. En el análisis de error inverso se trata de probar la existencia de tales conjuntos de datos para los cuales todos los ϵ_i son pequeños.

Encontramos que para las operaciones aritméticas básicas los valores naturales de ϵ_i son sugeridos por las relaciones fundamentales sastifechas por los valores calculados; con muchos algoritmos de matrices las aplicaciones de la relación,

$$fl(a * b) = (a * b)(1 + \epsilon) \qquad \text{donde } |\epsilon| < k\beta^{-t}$$

para todas las operaciones involucradas también lleva de una manera equitativamente natural a un conjunto de datos apropiados.

$$x_i \quad (1 + \epsilon_i)$$

Ahora para todas las operaciones básicas en si mismas, estamos forzando para aceptar las perturbaciones relativas en el conjunto de datos (aquí son los operandos a, b) de

$-t$
arriba para kB , donde k es la constante en la expresión (i). Para un algoritmo muy complicado sería razonable esperar el poder garantizar guardar el ϵ bajo este nivel. i

Realmente si el número de operaciones involucradas en el algoritmo es np, es decir, un promedio de n operaciones por elemento de datos, un algoritmo para el cual podríamos $-t$
establecer límites a-priori de nkB , para la ϵ podría ser i
considerado como extremadamente estable con respecto a errores de redondeo.

Un algoritmo para el cual los límites existen sastifactoriamente para ϵ , se dice que es inversamente i
estable. Debemos distinguir entre la existencia de estabilidad inversa y nuestra habilidad para establecerla. Adicionalmente si tomamos dos algoritmos A y B para resolver el mismo problema el establecimiento riguroso de los límites más pequeños para las ϵ correspondiendo a A i
que aquel correspondiente a B, no necesariamente significa que sea más perceptivo para A que para B.
Sin embargo, ahora tenemos una experiencia suficientemente grande del análisis de error de algoritmos de matrices que en muchas situaciones podemos estar razonablemente confiados de que nuestros límites dan una estimación realista de ejecución cuando hemos hecho una apropiada concesión para la distribución estadística de los errores de redondeo. En la práctica será de nuestro interés la precisión de la solución calculada es decir, la distancia entre la y calculada y la y exacta correspondiente a la j j
x dada. La estabilidad inversa no garantiza que la i
solución calculada sea de gran precisión pero hasta ahora como no se hace esto es obviamente debido a la gran sensibilidad de la solución para perturbaciones en los datos.

Esta sensibilidad es una propiedad inherente del problema

calculado y no tiene nada que ver con la efectividad del algoritmo.

Hay tres características importantes del análisis de error inverso.

a).- Pone los errores hechos durante el curso de la solución en la misma posición, como errores en los datos. En la práctica los datos para ves serán exactos; comunmente involucrán errores de observación, si los errores relativos probables en las a_i dadas son más grandes que ϵ dada, por un análisis de error inverso entonces, los errores de redondeo, efectuados durante el curso del algoritmo son menos importantes que los errores iniciales en los datos. Cuando todas las a_i son conocidas exactamente pueden no ser números digitales: los errores serán entonces involucrados en las representaciones computacionales de a_i aunque los errores relativos introducidos de esta manera estarían limitados por $k\beta^{1-t}$.

b).- Para estimar los errores en la solución calculada podemos usar la teoría de las perturbaciones y podemos por consiguiente sacar una fuente muy rica de información.

c).- La tercera característica es la que parece brotar de la gran naturaleza del análisis de error inverso. Para muchos algoritmos matemáticos sigabusirna produce una simplicidad formal inesperada y muchas veces el efecto de los errores es puramente aditivo, cuando esto sucede se puede hacer un análisis de error para cubrir los grandes desatinos aritméticos, tanto como los meros errores de redondeo.

Ahora volviendo a la ortogonalización.

Sean X y Y vectores de dimención n donde $n * 2^{-t} \leq 0.1$ implica que los siguientes límites de error para los productos internos de X y Y computados son validos.

$$| \; fl(x^t y) - x^t y \; | \leq n \cdot 2^{-t} \; | x^t \; | \; | y \; | \qquad (3.1)$$

donde.

$$| \; fl(x^t y) - x^t y \; | \leq 2^{-t} \; | x^t \; y \; | + 3/2 \cdot n 2^{-2t2} \; |x^t| \; |y| \qquad (3.2)$$

$$t1 = t - \log_2(1.06), \qquad 2t2 = 2t - \log_2(1.06) \qquad (3.3)$$

Aquí $| x |$ denota un vector con componentes $| x_i |$ y fl ( ) aplicado al caso cuando el producto interno es realizado en doble presición y entonces redondeado, las cantidaes calculadas serán en lo sucesivo denotadas por barras asi, escribimos

$$\bar{A}^{(k)} = (\bar{q}_1, \ldots, \bar{q}_{k-1}, \bar{a}^{(k)}_k, \ldots, \bar{a}^{(k)}_n)$$

Y las formulas

$$d_k = \left\| q^{(k)} \right\|_2^2 \qquad r'_{kj} = (q^{(k)}_k)^t \; a^{(k)}_j \div d_k$$

$$a^{(k+1)}_j = a^{(k)}_j - r'_{kj} \; q^{(k)}_k \qquad k+1 \leq j \leq n$$

$$y'_k = (q^{(k)}_k)^t \; b^{(k)} \div d_k, \qquad b^{(k+1)} = b^{(k)} - y'_k \; q^{(k)}_k$$

Se convierten en:

$$\bar{d}_k = fl\left(\left\|\bar{q}'_k\right\|_2^2\right), \qquad \bar{r}'_{kj} = fl\left(\left(\bar{q}'_k\right)^t \bar{a}_j^{(k)} \div \bar{d}_k\right) \tag{3.4}$$

$$\bar{a}_j^{(k+1)} = fl\left(\bar{a}_j^{(k)} - \bar{r}'_{kj}\,\bar{q}'_k\right) \qquad k+1 \le j \le n$$

y

$$\bar{y}'_k = fl\left(\left(\bar{q}'_k\right)^t \bar{b}^{(k)} \div \bar{d}_k\right), \qquad \bar{b}^{(k+1)} = fl\left(\bar{b}^{(k)} - \bar{y}'_k\,\bar{q}'_k\right) \tag{3.5}$$

este algoritmo no se rompe a menos que $\bar{d}_k = 0$ para alguna k esto puede pasar debido a la totalidad de errores de redondeo cuando $r(A) = n$ de acuerdo a

$$\left| fl\left(x^t y\right) - x^t y \right| \le n \cdot 2^{-t} \left| x^t \right| \left| y \right| \tag{3.1}$$

y

$$\left| fl\left(x^t y\right) - x^t y \right| \le 2^{-t} \left| x^t y \right| + 3/2 \cdot n2^{-2t} \|x\|\|y\| \tag{3.2}$$

$\bar{d}_k = 0$ implica $\bar{q}'_k = 0$ para este caso añadimos la regla trivial que cuando $\bar{q}'_k = 0$ tomamos :

$$\bar{r}'_{kj} = 0, \quad k+1 \le j \le n, \quad \bar{y}'_k = 0 \tag{3.6}$$

entonces $\bar{R}'$ y $\bar{Y}'$ están tambien unicamente definidas.

Para convenir en notación tambien introducimos las cantidades normalizadas :

$$\bar{q}_k = d_k^{-\frac{1}{2}}\,\bar{q}'_k, \qquad \bar{r}_{kj} = d_k^{\frac{1}{2}}\,\bar{r}'_{kj} \qquad \bar{y}_k = d_k^{\frac{1}{2}}\,\bar{y}'_k \tag{3.7}$$

donde :

$$d_k^{\frac{1}{2}} = \begin{cases} \left\| \bar{q}'_k \right\|_2, & \bar{q}'_k \neq 0 \\ \left\| \bar{q}'_k \right\|_2, & \bar{q}'_k = 0 \end{cases}$$

Notese que esas cantidades nunca son calculadas y así

$$d_k^{\frac{1}{2}} = \begin{cases} \left\| \bar{q}'_k \right\|_2, & \bar{q}'_k \neq 0 \\ \left\| \bar{q}'_k \right\|_2, & \bar{q}'_k = 0 \end{cases}$$

Son relaciones exactas.

C A P I T U L O IV

ERRORES EN UNA PROYECCION ELEMENTAL

Si $\bar{q}_k$ =/= 0 entonces, en el k - ésimo paso calculamos los vectores $\bar{a}_j^{(k+1)}$ con j = k + 1, . . . , n

Como la proyección de $\bar{a}_j^{(k)}$ en el subespacio complementario a $\bar{q}_k$. Si esto se ejecuta sin errores de redondeo el resultado es :

$$\bar{a}_j^{(k+1)} = (I - \bar{q}_k \bar{q}_k^T) \bar{a}_j^{(k)} = \bar{a}_j^{(k)} - r_{kj} \bar{q}_k$$

donde r_{kj} es el multiplicador exacto correspondiente a la $\bar{q}_k$ calculada y $\bar{a}_j^{(k)}$. Si se usa el multiplicador calculado $\bar{r}_{kj}$ la sustracción se lleva a cabo exactamente, entonces el resultado es :

$$\bar{a}_j^{(k)} = \bar{r}_{kj} \bar{q}_k = \bar{a}_j^{(k)} - \bar{r}_{kj} \bar{q}_k$$

y definimos los errores $\alpha_j^{(k)}$ y $\beta_j^{(k)}$ por :

$$\bar{a}_j^{(k+1)} = \bar{a}_j^{(k)} - \bar{r}_{kj} q_k + \alpha_j^{(k)}$$

$$\bar{a}_j^{(k+1)} = (I - \bar{q}_k \bar{q}_k^T) \bar{a}_j^{(k)} + \beta_j^{(k)}$$

en el caso singular cuando $\bar{\eta} = 0$ estas relaciones se satisfacen con :

$$\bar{a}_j^{(k+1)} = \bar{a}_j^{(k)}$$

$$\bar{\alpha}_j^{(k)} = \bar{\beta}_j^{(k)} = 0$$

En el caso no singular probaremos las siguientes entidades las cuales son básicas para nuestro análisis.

$$\left\| \bar{\alpha}_j^{(k)} \right\|_2 \leq 1.45 \cdot 2^{-t} \left\| \bar{a}_j^{(k)} \right\|$$

$$\left\| \bar{\beta}_j^{(k)} \right\|_2 \leq \begin{cases} 3.23 \cdot 2^{-t} \left\| \bar{a}_j^{(k)} \right\|_2 \\ (2m+3) \cdot 2^{-t} \left\| \bar{a}_j^{(k)} \right\|_2 \end{cases}$$

donde $\left\| \bar{\alpha}_j^{(k)} \right\|_2 \leq 1.45 \cdot 2^{-t} \left\| \bar{z}_j^{(k)} \right\|_2$

es valida tambien cuando se usa precisión simple. Para simplificar la notación omitiremos los indices j y k y escribimos.

$$\bar{a}_j^{(k+1)} = x$$

$$\bar{a}_j^{(k)} = y$$

$$\bar{q}_{k} = x$$

Entonces las relaciones siguientes.

$$\bar{a}_{k} = \int \left| \left[\left| \left| \bar{q}_{k}' \right| \right|^{2} \right|_{2} \right]$$

$$\bar{r}_{kj}' = \int \left| \left(\bar{a}_{j}^{(k)} - r_{kj}' \, q_{k}' \right) \right.$$

con $k + 1 \leq j \leq n$.

$$\bar{a}_{j}^{(k+1)} = \bar{a}_{j}^{(k)} - \bar{r}_{kj} \, \bar{q}_{k} + \bar{\Omega}_{j}^{(k)}$$

y

$$\bar{a}_{j}^{(k+1)} = \left(I - \bar{q}_{k} \, \bar{q}_{k} \right) \bar{a}_{j}^{(k)} + \bar{\Omega}_{j}^{(k)}$$

se convierten en .

$$\bar{r}_{\ }' = \int \left| \left(\left(x' \right)^{T} y \div \left(x' \right)^{T} x' \right) \right.$$

$$Z = \int \left| \left(y - \bar{r}' x' \right) \right.$$

$$r = x^{T} y$$

$$z = y - \bar{r} \, x + \bar{\Omega} \quad ; \quad z = y - r \, x + \hat{\beta}$$

respectivamente.

Ahora de la definición de operación en punto flotante y de las cantidades normalizadas

$$\bar{q}_k = d_k^{-\frac{1}{2}} \bar{q}'_k ,$$

$$\bar{r}_{kj} = d_k^{\frac{1}{2}} \bar{r}'_{kj}$$

$$\bar{Y}_k = d_k^{\frac{1}{2}} \bar{y}'_k$$

donde $d_k^{\frac{1}{2}} = \begin{cases} \| \bar{q}'_k \|_2 , & \bar{q}'_k \neq 0 \\ 1 , & \bar{q}'_k = 0 \end{cases}$

se sigue que

$$\bar{z}_i = (y_i - \bar{r}_1 x_1 (1 + E_1)) (1 + E_2)$$

$$| E_1 | \leq 2^{-t} , \quad | E_2 | \leq 2^{-t}$$

usando esto para eliminar y_i de la definición de $\hat{\rho}$ tenemos.

$$\hat{\rho} = \frac{E_2}{1 + E_2} \| \bar{z} \|_2 - E_1 \cdot r_1 x_1 ,$$

y de aquí dado que $\| x \|_2 = 1$

$$\left\| \hat{\alpha} \right\|_2 \leq \frac{2^{-t}}{1 - 2^{-t}} \left\| \bar{z} \right\|_2 + 2^{-t} \left| \bar{r} \right|$$

inmediatamente tenemos de las relaciones

$$r = x^T y , \quad \bar{z} = y - \bar{r}\, x + \hat{\alpha} ; \quad \bar{z} = y - r\, x + \beta$$

$$\left\| \beta \right\|_2 \leq \left\| \hat{\alpha} \right\|_2 + \left| r - \bar{r} \right|$$

Para estimar el error en el multiplicador $\bar{r}$ usamos la definición de la operación de punto flotante y le restamos.

$$\left| fl\left(x^T y \right) - x^T y \right| \leq 2^{-t} \left| x^T y \right| + 3/2 \cdot n 2^{-2t} \left| x \right| \left| y \right|$$

y la identidad.

$$\frac{fl\left(x^T y \right)}{fl\left(x^T x \right)} - r = \frac{x^T x}{fl\left(x^T x \right)} \left[\frac{fl\left(x^T y \right) - x^T y}{x^T x} - r \frac{fl\left(x^T x \right) - x^T x}{x^T x} \right]$$

si calculamos $\bar{r}'$ como..

$$\bar{d} = fl_2\left(\left(x' \right)^T x' \right)$$

$$\bar{r}' = fl_2\left(\left(x' \right)^T y \Big| \bar{d} \right)$$

acumulando productos internos en doble precisión y dividiendo $\bar{d}$ entre la mantisa de doble precisión después de redondear, entonces

$$|\bar{r} - r| \leq \frac{2^{-t}(2 + 2/3 \cdot m \cdot 2^{-t})|r| + 2/3 \cdot m \cdot 2^{-t} ||y||_2}{1 - 2^{-t}(|r| + 2/3 \cdot m \cdot 2^{-t})}$$

Usando precisión simple tenemos

$$|\bar{r} - r| \leq \frac{(m+1)2^{-t}|r| + m(1 + 2^{-t})2||y||_2}{1 - m \cdot 2^{-t}}$$

Para simplificar estas últimas desigualdades haremos enseguida las siguientes observaciones razonables.

$$m \geq 2$$

$$2(m+1)2^{-t} < 0.01$$

entonces ciertamente tenemos.

$$|\bar{r} - r| < \begin{cases} (2.01 \cdot |r| + 0.01 ||y||_2) 2^{-t} \\ ((m+1) \cdot |r| + m ||y||_2) 2^{-t} \end{cases}$$

donde los hemos tomado en cuenta que de las desigualdades

$$m \geq 2^{t}, \quad 2(m+1)2^{-t} < 0.01$$

el factor 1.01 en 2^{-t} es ahora suficientemente generoso para también para los factores $1 - m \cdot 2^{-t}$ y $1 + 2^{-t}$.

Dado que $(x - \beta)$ es ortogonal a x se sigue que :

$$\| z \|_2^2 \leq (\| y \|_2^2 - r^2)^{1/2} + \| \beta \|_2$$

sustituyendo esto y

$$\| \Omega \|_2 \leq \frac{2^{-t}}{1 - 2^{-t}} \| z \|_2 + 2 \left| \bar{r} \right|$$

en

$$\| \beta \|_2 \leq \| \Omega \|_2 + \left| r - \bar{r} \right|$$

sustituyendo el valor $\| \Omega \|_2$ nos queda

$$\| \beta \|_2 \leq \frac{2^{-t}}{1 - 2^{-t}} \| z \|_2 + 2^{-t} \left| \bar{r} \right| + \left| r - \bar{r} \right|$$

Ahora sustituimos, el valor de $\| \quad \|_2$

$$\| \beta \|_2 \leq \frac{2^{-t}}{1 - 2^{-t}} \left(\| y \|_2^2 - r^2 \right)^{1/2} + \| \beta \|_2 + 2^{-t} |r| + |r - \bar{r}|$$

resolviendo para $\| \beta \|_2$ y despues de algunos arreglos nos queda.

$$(1 - 2 \cdot 2^{-t}) \| \beta \|_2 \leq 2^{-t} \left(\| y \|_2^2 - r^2 \right)^{1/2} + |r| + |r - \bar{r}|$$

de aquí usando.

$$|\bar{r} - r| < \begin{cases} (2.01 \cdot |r| + 0.01 \| y \|_2) 2^{-t} \\ ((m+1) \cdot |r| + m \| y \|_2) 2^{-t} \end{cases}$$

nos queda

$$(1-2 \cdot 2^{-t}) \| \beta \|_2 \leq \begin{cases} \left(\| y \|_2^2 - r^2 \right)^{1/2} + 3 |r| + \ldots \\ \quad + 0.02 \| y \|_2 \big) 2^{-t} \\ \left(\| y \|_2^2 - r^2 \right)^{1/2} + (m+2) |r| + \ldots \\ \quad + m \| y \|_2 \big) 2^{-t+1} \end{cases}$$

Maximizando sobre $|r|$, donde

$0 \le |r| \le \|y\|_2$ tenemos

$$\left(\|y\|_2^2 - r^2 \right)^{\frac{1}{2}} + k\,|r| \le \left(1 + k^2 \right)^{\frac{1}{2}} \|y\|_2$$

y finalmente :

$$\left\| \beta_{ij}^{(k)} \right\|_2 \le \begin{cases} 3 \cdot 2^{3} \cdot 2^{-t} \left\| a_j^{(k)} \right\|_2 \\ (2m+3) \cdot 2^{-t1} \left\| a_j^{(k)} \right\|_2 \end{cases}$$

se sigue de :

$$(1-2 \cdot 2^{-t}) \|\beta\|_2^2 \le \begin{cases} \left(\|y\|_2^2 - r^2 \right)^{\frac{1}{2}} + 3\,|r| + \ldots \\ \quad + 0.02 \|y\|_2^2 \cdot 2^{-t} \\ \left(\|y\|_2^2 - r^2 \right)^{\frac{1}{2}} + (m+2)\,|r| + \ldots \\ \quad + m \|y\|_2^2 \cdot 2^{-t1} \end{cases}$$

al retomar

$$m \ge 2, \quad 2(m+1)2^{-t} < 0.01$$

Para la estimación de $\| \delta \|$ las relaciones.

$$\| \delta \|_2 \leq \frac{2^{-t}}{1 - 2^{-t}} \| r \|_2 + 2^{-t} | \bar{r} |$$

y

$$\| z \|_2 \leq (\| y \|_2^2 - r^2)^{1/2} + \| \beta \|_2$$

nos llevan a

$$\| \hat{x} \|_2 \leq \frac{2^{-t}}{1 - 2^{-t}} ((\| y \|_2^2 - r^2)^{1/2} + \| \beta \|_2 + | r - \bar{r} |)$$

Usando la relación :

$$(\| y \|_2^2 - r^2)^{1/2} + k | r | \leq (1 + k^2)^{1/2} \| y \|_2$$

para los primeros dos términos y los límites en precisión simple para los últimos dos: tenemos.

$$\| \hat{x} \|_2 \leq \frac{2^{-t}}{1 - 2^{-t}} (2 + 4 (m + 1) \cdot 2^{-t1})^{1/2} \| y \|_2$$

y retomando las relaciones :

$$m \geq 2; \quad 2(m + 1)2^{-t1} < 0.01$$

y nuevamente esto prueba que :

$$\left\| \bar{a}_j^{(k)} \right\|_2 \leq 1.45 \cdot 2^{-t} \left\| \bar{a}_j^{(k)} \right\|_2$$

se cumple.

En un calculo exacto, ciertamente se tiene :

$$\left\| \bar{a}_j^{(k+1)} \right\|_2 \leq \left\| \bar{a}_j^{(k)} \right\|_2$$

y esto no necesariamente tiene que ser cierto cuando son tomados en cuenta los errores de redondeo.

Estimando $\| \beta \|_2$ de la relación.

$$(1-2 \cdot 2^{-t}) \| \beta \|_2 \leq \begin{cases} ((\| Y \|_2^2 - r^2)^{1/2} + 3 |r| + \ldots \\ \quad + 0.02 \| Y \|_2)2^{-t} \\ ((\| Y \|_2^2 - r^2)^{1/2} + (m+2) |r| + \ldots \\ \quad + m \| v \|_2)2^{-ti} \end{cases}$$

en

$$\| z \|_2 \leq (\| y \|_2^2 - r^2)^{1/2} + \| \beta \|_2$$

y otra vez maximizando el resultado por la derecha como una función de $|r|$, resulta despues de algunos cálculos, que

$$\left\| \underline{a}_j^{(k+1)} \right\|_2 \leq \begin{cases} \left(1 + 1.05 \cdot 2^{-t} \right) \left\| \underline{a}_j^{(k)} \right\|_2 \\ \left(1 + 1.01 \left(m + 2 \right) \cdot 2^{-t} \right) \left\| \underline{a}_j^{(k)} \right\|_2 \end{cases}$$

Notese que estos limites solo pueden ser aprovechados cuando $\underline{a}_j^{(k)}$ es tambien ortogonal a $\bar{q}$.

C A P I T U L O V

ERRORES EN LA FACTORIZACION

Usando el análisis de error de la proyección elemental podemos ahora derivar límites para los errores de la factorización de A y b.

Mientras podemos mostrar que el error ($\bar{Q}\bar{R} - A$) es pequeño independientemente de $\kappa(A)$, esto quiza no se deba al error ($\bar{R} - \bar{Q}^{-1} A$) está nos lleva a introducir las matrices:

$$\tilde{Q} = (\tilde{q}_1 , \tilde{q}_2 , \ldots , \tilde{q}_n), \quad \hat{Q} = (\hat{q}_1 , \hat{q}_2 , \ldots \hat{q}_n)$$

$$\tilde{q}_k = (I-\bar{q}_1 \bar{q}_1^t) (I - \bar{q}_2 \bar{q}_2^t)\ldots(I - \bar{q}_{k-1} \bar{q}_{k-1}^t)\bar{q}_k ,$$

$$\tilde{q}_k = (I-\bar{q}_n \bar{q}_n^t) (I-\bar{q}_{n-1} \bar{q}_{n-1}^t)\ldots(I-\bar{q}_{k-1} \bar{q}_{k-1}^t)\bar{q}_k , \qquad (5.1)$$

cuando $\bar{q}_j$ =/- 0 la matriz $p^{(j)} = (I-\bar{q}_j \bar{q}_j^t)$ donde p es una matriz unitaria, es la proyección del subespacio complementario a $\bar{q}_j$, de otra manera es igual a la matriz unitaria y se sigue que :

$$\left\| p^{(j)} \right\|_2 = 1, \quad \left\| \tilde{q}_k \right\|_2 \leq 1, \quad \left\| \hat{q}_k \right\|_2 \leq 1 \qquad (5.2)$$

Primero probaremos el Lema simple pero importante concerniente a $\tilde{Q}$ y $\hat{Q}$.

LEMA (5.1)

Sea U la matriz triangular superior con elementos $u_{ij} = \bar{q}_i^t \bar{q}_j$ donde $j > i$, $u_{ij} = 0$ cuando $j \leq i$ entonces

$$\tilde{Q} = Q(I + U), \quad \hat{Q} = Q(I + U^t)$$

Primero probemos por inducción sobre k para $k = 1$

$$(I-\bar{q}_1\bar{q}_1^t)\ldots(I-\bar{q}_k\bar{q}_k^t) = I-\tilde{q}_k\bar{q}_k^t - \ldots \tilde{q}_1\bar{q}_1^t ;$$

esto es verdad para $k = 1$ y tenemos :

$$(I-\bar{q}_1\bar{q}_1^t)\ldots(I-\bar{q}_k\bar{q}_k^t)(I-\bar{q}_{k+1}\bar{q}_{k+1}^t) =$$

$$= \tilde{q}_{k+1}\bar{q}_{k+1}^t + (I-\bar{q}_1\bar{q}_1^t)\ldots(I-\bar{q}_k\bar{q}_k^t),$$

lo cual prueba la hipotesis.

Usando este resultado tenemos de la definicion de q_k

$$\tilde{q}_k = (I-\hat{q}_{k-1}\hat{q}_{k-1}^t - \ldots - \tilde{q}_1\bar{q}_1^t)\bar{q}_k ,$$

y se sigue que para k = 1, . . . n

$$\bar{q}_k = q_k + (\bar{q}_k \bar{q}_{k-1}^t)q_{k-1} + \ldots + (\bar{q}_k \bar{q}_1^t)q_1$$

y está es la k-esima columna de la igualdad $\bar{Q} = Q(I+U)$ y la primera parte del lema está probado, la segunda parte es tambien obvia por la simetria de la definición $\tilde{Q}$ y $\tilde{\tilde{Q}}$ y se prueba igual.

Al tomar la transpuesta de la hipotesis de inducción usada en la prueba del lema obtenemos el siguiente

COLORARIO: si $\tilde{Q}$ está definida por (5.1)

$$\tilde{Q} = (\tilde{q}_1, \tilde{q}_2, \ldots \tilde{q}_n), \quad \tilde{\tilde{Q}} = (\tilde{\tilde{q}}_1, \tilde{\tilde{q}}_2, \ldots \tilde{\tilde{q}}_n)$$

$$\tilde{q}_k = (I-\bar{q}_1\bar{q}_1^t)(I-\bar{q}_2\bar{q}_2^t)\ldots(I-\bar{q}_{k-1}\bar{q}_{k-1}^t)\bar{q}_k,$$

$$\tilde{\tilde{q}}_k = (I-\bar{q}_n\bar{q}_n^t)(I-\bar{q}_{n-1}\bar{q}_{n-1}^t)\ldots(I-\bar{q}_{k-1}\bar{q}_{k-1}^t)\bar{q}_k,$$

tenemos la identidad :

$$I - \bar{Q}\bar{Q}^t = (I-\bar{q}_n\bar{q}_n^t)\ldots(I-\bar{q}_2\bar{q}_2^t)(I-\bar{q}_1\bar{q}_1^t)$$

y ahora definimos :

$$E_1 = \bar{Q}\bar{R} - A, \qquad e_1 = \bar{Q}\bar{Y} + \bar{b}^{(m+i)} - b \qquad (5.3)$$

$$E_2 = (I - \bar{Q}\,\bar{Q}^{t})A, \qquad e_2 = (I - \bar{Q}\,\bar{Q}^{t})b - \bar{b}^{(n+1)} \tag{5.4}$$

$$E_3 = \bar{R} - \bar{Q}^{t}A, \qquad e_3 = \bar{Y} - \bar{Q}^{t}b. \tag{5.5}$$

Ahora probaremos los siguientes estimadores los cuales son validos si los productos internos son acumulados en doble presición:

$$\| \epsilon_1 \|_2 \le 1.5(n-1)\cdot 2^{-t} \| A \|_E \tag{5.6}$$

$$\| E_1 \|_E \le 1.5n \cdot 2^{-t} \| b \|_2$$

$$\| \epsilon_2 \|_2 \le 3.25(n-1)\cdot 2^{-t} \| A \|_E \tag{5.7}$$

$$\| E_2 \|_E \le 3.25n \cdot 2^{-t} \| b \|_2$$

$$\| \epsilon_3 \|_2 \le 1.9(n-1)^{\frac{1}{2}} n \cdot 2^{-t} \| A \|_E \tag{5.8}$$

$$\| E_3 \|_E \le 1.9\, n^{\frac{1}{2}} (n+1)\cdot 2^{-t} \| b \|_2$$

Sumando :

$$\underline{a}_j^{(k+1)} = \underline{a}_j^{(k)} - \bar{r}_{kj}\,\bar{q}_k + \delta_j^{(k)} \qquad \text{para } k=1,\ldots,j-1$$

y usando :

$$\underline{a}_j^{(j)} = \bar{r}_{jj}\,\bar{q}_j \;,\; \underline{a}_j^{(1)} = a_j \qquad \text{tenemos :}$$

$$\sum_{k=1}^{j} \bar{r}_{kj}\,\bar{q}_k = a_j \;=\; \sum_{k=1}^{j-1} \delta_j^{(k)} + \underline{a}_j^{(j)}$$

de :

$$\left\| \delta_j^{(k)} \right\|_2 \le 1.45 \cdot 2^{-t} \left\| \underline{a}_j^{(k)} \right\|_2$$

se sigue : $\left\| \delta_j \right\|_2 \le 1.45 \cdot 2^{-t} \sum_{k=1}^{j-1} \left\| \underline{a}_j^{(k)} \right\|_2$

usando el límite de punto flotante fl ( ) en doble precisión en :

$$\left\| \underline{a}_j^{(k+1)} \right\|_2 < \begin{cases} (1+1.05 \cdot 2^{-t}) \left\| \underline{a}_j^{(k)} \right\|_2 \\ (1+1.01(m+2) \cdot 2^{-t}) \left\| \underline{a}_j^{(k)} \right\|_2 \end{cases}$$

tenemos : $\| a_j^{(k)} \|_2 < (1 + 1.05 \cdot 2^{-t(k-1)}) \| a_j \|_2$

y así recordando : $m \geq 2, \quad 2(m+1)2^{-t1} < 0.01$

se tiene que :

$$\| a_j^{(k)} \|_2 < (1+1.05\, m\, 2^{-t1}) \| a_j \|_2 < 1.006 \| a_j \|_2 \qquad (5.9)$$

está desigualdad se toma complementariamente cuando se usa presicion simple si

$$m \geq 2, \quad 2(m+1)2^{-t1} < 0.01$$

es sustituida por la condicion más fuerte:

$2n(m+2)2^{-t1} < 0.01$ y entonces se tiene

$$\| \delta_j \|_2 < 1.5 \cdot (j-1) 2^{-t} \| a_j \|_2 \qquad (5.10)$$

y se tiene :

$$\| E_1 \|_E = \left[\sum_{j=1}^{n} \| \delta_j \|_2^2 \right]^{1/2} < 1.5 \cdot (n-1) 2^{-t} \left[\sum_{j=1}^{n} \| a_j \|_2^2 \right]^{1/2}$$

Lo cual prueba la primera parte de :

$$\| E_1 \|_E \leq 1.5(n-1) \cdot 2^{-t} \| A \|_E$$

dado que el lado derecho de p es tratado igual que las columnas de Q, la segunda parte se sigue analogamente. enseguida resolveremos la ecuación diferencia.

$$\bar{a}_j^{(k+1)} = (I - \bar{q}_k \bar{q}_k^{t}) \bar{a}_j^{(k)} + \Gamma_j^{(k)}$$

para $\bar{a}_j^{(k)}$ y tomamos para $k \leq j$

(5.11)

$$\bar{a}_j^{(k)} = \bar{a}_j^{(k)} + p^{(k-1)} \ldots p^{(2)} p^{(1)} \Gamma_j^{(k+1)} + \ldots + p^{(k-1)} \Gamma_j^{(k-1)} + \Gamma_j^{(k)}$$

donde: $\bar{a}_j^{(k)} = p^{(k-1)} \ldots p^{(2)} p^{(1)} \bar{a}_j$

de aquí que $\bar{a}_j^{k}$ es el vector que obtenemos si ejecutamos la computación exactamente con la $\bar{q}_k$

Dado que la $\bar{q}_j = \bar{a}_j^{(j)}$ tenemos $(I - \bar{q}_j \bar{q}_j^{t}) \bar{a}_j^{(j)} = 0$ consecuentemente

$$\bar{a}_j^{(k)} = \bar{a}_j^{(k)} + p^{(k-1)} \ldots p^{(2)} p^{(1)} \Gamma_j^{(k+1)} + \ldots + p^{(k-1)} \Gamma_j^{(k-1)} + \Gamma_j^{(k)}$$

toma tambien para $j + 1 \leq k \leq n + 1$ con :

$$\bar{a}_j^{(k)} = 0, \quad \Gamma_j^{(j)} = \ldots = \Gamma_j^{(k-1)} = 0.$$

Tomando normas en :

$$\bar{a}_j^{(k)} = a_j^{(k)} + p_j^{(k+1)} + p_j^{(3)} r_j^{(2)} + r_j^{(1)} + \ldots + p_j^{(k+1)} r_j^{(k-2)} r_j^{(k-1)}$$

y usando :

$$\left\| p_j^{(j)} \right\|_2 = 1, \quad \left\| \tilde{q}_k \right\|_2 \leq 1, \quad \left\| \hat{q}_k \right\|_2 \leq 1$$

llegamos a:

$$\left\| \bar{a}_j^{(k)} - a_j^{(k)} \right\|_2 \leq \sum_{i=1}^{s-1} \left\| r_j^{(i)} \right\|_2 \qquad s = \min(j, k)$$

De esto se sigue usando el fl ( ) de :
2

$$\left\| r_j^{(k)} \right\|_2 \leq \begin{cases} 3.23 \cdot 2^{-t} \left\| \bar{a}_j^{(k)} \right\|_2 \\ (2m+3) \cdot 2^{-t1} \left\| \bar{a}_j^{(k)} \right\|_2 \end{cases}$$

$$\left\| \bar{a}_j^{(k)} - a_j^{(k)} \right\|_2 \leq 3.23 \cdot 2^{-t} \sum_{i=1}^{s-i} \left\| \bar{a}_j^{(i)} \right\|_2$$

$$\leq 3.25 \, (s - 1) \, 2^{-t} \left\| a_j \right\|_2 \qquad (5.12)$$

Usando el Corolario del Lema anterior se puede escribir para k = n + 1 como :

$$\left\| \begin{bmatrix} I - \hat{Q}\,\hat{Q}^{t} & \\ & \end{bmatrix} a_j \right\|_2 \leq 3.25\,(j-1)\cdot 2^{-t} \left\| a_j \right\|_2$$

Esto y una relación similar para b prueba:

$$\left\| E_2 \right\|_E \leq 3.25(n-1)\cdot 2^{-t} \left\| A \right\|_E$$

Usando la identidad :

$$\tilde{q}_k^{t}\, a_j = \bar{q}_k^{t}\,(I-\bar{q}_{k-1}\bar{q}_{k-1}^{t})\ldots(I-\bar{q}_1\bar{q}_1^{t})\, a_j = \bar{q}_k^{t}\, a_j^{(k)}$$

podemos escribir el componente (k, j) de $\bar{R} - \tilde{Q}^{t} A$

$$\tilde{r}_{k,j} - \tilde{q}_k^{t} a_j = (\bar{r}_{k,j} - r_{k,j}) + \bar{q}_k^{t}\,(a_j^{(k)} - a_j^{(k)}),\quad k < j$$

se sigue que :

$$\left\| a_j^{(k)} - a_j^{(k)} \right\|_2 \leq 3.23\cdot 2^{-t} \sum_{i=i}^{s-1} \left\| a_j^{(i)} \right\|_2$$

$$\leq 3.25\,(s-1)\,2^{-t} \left\| a_j \right\|_2$$

y

$$\left| \bar{r} - r \right| < \begin{cases} (2.01 \cdot |r| + 0.01 \, ||y||_2) 2^{-t} \\ ((m+1) \cdot |r| + m \, ||y||_2) 2^{-t1} \end{cases}$$

Dan la norma L_2 de la j-ésima columna en E_3, que está

así mismo limitado por lo cual finaliza la prueba de :

$$||E_3||_E \leq 1.9(n-1)^{1/2} n \cdot 2^{-t} ||A||_E$$

Notese que cuando A está mal condicionada la cancelación

ocurrira tal que norma de $||\bar{a}_j^{(k)}||_2 \leq ||a_j||_2$

generalmente para un valor pequeño de k. Entonces al

$$||a_j^{(k)}||_2 < (i+1.05 \, n \, 2^{t1}) ||a_j||_2 < 1.006 ||a_j||_2$$

estimado el cual hemos usado en :

$$||\delta_j||_2 < 1.5 \cdot (j-1) 2^{-t} ||a_j||_2 \quad y$$

$$\left\| \bar{a}_j^{(k)} - a_j^{(k)} \right\|_2 \leq 3.23 \cdot 2^{-t} \sum_{i=1}^{s-1} \left\| \bar{a}_j^{(i)} \right\|_2$$

$$\leq 3.25\,(s-1)\,2^{-t} \left\| a_j \right\|_2$$

es muy bajo, y si restamos (5.6) de (5.8), esto es :

$$\left\| E_1 \right\|_E \leq 1.5(n-1) \cdot 2^{-t} \left\| A \right\|_E \quad \text{menos}$$

$$\left\| E_3 \right\|_E \leq 1.9(n-1)^{\frac{1}{2}} n \cdot 2^{-t} \left\| A \right\|_E$$

el error será considerablemente sobreestimado.

Si usamos presición simple de :

$$\left\| E_1 \right\|_E \leq 1.5(n-1) \cdot 2^{-t} \left\| A \right\|_E$$

tomados sin cambio pero en

$$\left\| E_2 \right\|_E \leq 3.25(n-1) \cdot 2^{-t} \left\| A \right\|_E \quad Y$$

$$\left\| E_3 \right\|_E \leq 1.9(n-1)^{\frac{1}{2}} n \cdot 2^{-t} \left\| A \right\|_E$$

los límites deben ser incrementados por un factor de : (3/2m + 1).

A menos que una a ó b varíen, el vector q determinará
 k   k   i
completamente el total de los vectores p y q
                                       i   i
Consecuentemente q podría ser elegido tal que el
                  i
algoritmo sea inestable en presencia del error de redondeo a no ser que bajo la reortogonalización, las líneas sugeridas por WILKINSON sean usadas. Esto es que el gran total de la ortogonalidad de los vectores generados por usar el método de GRAM-SCHMIDT para reortogonalizar cada uno de los vectores nuevamente generados p ó q a los vectores generados previamente
          i   i
p ó q , respectivamente
 i   i

Con el trabajo extra involucrado en la reortogonalización, el algoritmo es notablemente más lento, exceptuando la posibilidad de que A sea una matriz hueca.

C A P I T U L O V I

ORTOGONALIZACION DE LOS VECTORES CALCULADOS

Todos los límites de error dados en el capítulo 5 se aplican también en el caso singular cuando $\bar{q}'_k = 0$ para alguna k, es decir cuando el rango de $\bar{A} = \bar{Q}\bar{R}$ es $< n$. Ahora derivemos una condición suficiente para $\bar{A}$, para tener $r = n$.

$$E_1 = \bar{Q}\bar{R} - A, \qquad e_1 = \bar{Q}\bar{Y} + \bar{b}^{(n+1)} - b$$

podemos escribir :

$$\bar{A}^t \bar{A} = (A + E_1)^t (A + E_1) = \bar{R}^t (I + F_1) \bar{R} \qquad (6.1)$$

donde :

$$F_1 = (\bar{Q} E_1 \bar{R}^{-1})^t + \bar{Q} E_1 \bar{R}^{-1} + (E_1 \bar{R}^{-1})^t E_1 \bar{R}^{-1} \qquad (6.2)$$

Tomando normas en lo anterior tenemos :

$$\|F_1\|_2 \leq 2 \|E_1\|_E \|R^{-1}\|_2 + \|E_1\|_E^2 \|R^{-1}\|_2^2 \qquad (6.3)$$

y así la norma $\|F_1\|_2 < 1$ si :

$$\|E_1\|_E \|R\|_2^{-1} < \sqrt{2} - 1$$

usando :

$$\left\| \bar{E}_1 \right\|_E \leq 1.5(n-1) \cdot 2^{-t} \left\| \bar{A} \right\|_E$$

$$\left\| \bar{e}_i \right\|_2 \leq 1.5\, n \cdot 2^{-t} \left\| b \right\|_2$$

se sigue que $\bar{A}^t \bar{A}$ es no singular y $\bar{A}$ tiene rango n si:

$$1.5\,(n-1)\,2^{-t} \left\| A \right\|_E \left\| R^{-1} \right\|_2 < \sqrt{2} - 1 \qquad (6.4)$$

Asumimos en lo siguiente que la formula anterior satisface la ortogonalidad de los vectores calculados $\bar{q}_1, \bar{q}_2, \ldots, \bar{q}_n$ pueden ser medidos por la $\| \ \|$ de matriz $(I - \bar{Q}^t \bar{Q})$ ahora :

$$I - \bar{Q}^t \bar{Q} = -(U + U^t) \qquad (6.5)$$

donde U es la matriz introducida en el LEMA 5.1.

Si resumimos :

$$\bar{a}_j^{(k+1)} = \bar{a}_j^{(k)} - \bar{r}_{kj} \bar{q}_k + \delta_j^{(k)}$$

para $k = i+1, i+2, \ldots, j-1$ tenemos :

$$\bar{a}_j^{(i+1)} = \sum_{k=i+1}^{j} \bar{r}_{kj} \bar{q}_k - \sum_{k=i+1}^{j-1} \delta_j^{(k)} \qquad (6.6)$$

$$\text{de : } \bar{a}_j^{(k+1)} = (1 - \bar{q}_k \bar{q}_k^t) \bar{a}_j^{(k)} + \bar{r}_j^{(k)}$$

$$\text{implica : } \bar{q}_i^t \bar{a}_j^{(i+1)} = \bar{q}_i^t \bar{r}_j^{(i)}$$

y de aqui multiplicando :

$$\bar{a}_j^{(i+1)} = \bar{r}_j^{(i)} - \sum_{k=i+1}^{j-1} \bar{r}_{kj}^{(k)} \bar{q}_k \quad \text{por} \quad \bar{q}_i^t$$

tenemos :

$$S_{ij}^{(k)} = \sum_{k=i+1}^{j} \bar{r}_{kj}^{(i)} (\bar{q}_i^t \bar{q}_k) = \bar{q}_i^t (\bar{r}_j^{(i)} - \sum_{k=i+1}^{j-1} S_{ij}^{(k)}) \qquad (6.7)$$

donde S_{ij} es la componente (i, j) de la matriz $S = \bar{Q}\bar{R}$

usando el fl el límite de : $m \geq 2^{-t+1} (m+1) 2^{-t} < 0.01$

$$| \bar{r} - r | < \begin{cases} (2.01 \cdot || r ||_2 + 0.01 || y ||_2) 2^{-t} \\ ((m+1) \cdot || r ||_2 + m || y ||_2) 2^{-t+1} \end{cases}$$

y

$$|| \bar{a}_j^{(k)} ||_2 < (1+1.05\, n\, 2^{-t+1}) || a_j ||_2 < 1.006 || a_j ||_2$$

tenemos :

$$\left| S_{ij} \right| \leq \left(8.25 + 1.5 \left(j - i - 1 \right)^{1/2} \right)^{-t} \left\| a_j \right\|_2$$

y dado que :

$$\sum_{i=1}^{j-1} (j - i + 7/6)^2 \leq \sum_{i=1}^{n-1} (i+7/6)^2 = 1/3(n-1)((n+3/2)^2 + 11/6)$$

$$< \frac{1.0001}{3} n^2 (n+1)^2$$

Podemos limitar $\| \|_{L_2}$ de cada columna de S por :

$$0.87 \cdot n^{1/2} (n+1)2^{-t} \left\| a \right\|_2$$

despues de una estimación similar usando los limites para precisión simple obtenemos.

$$\left\| U \bar{R} \right\|_E < \begin{cases} 0.87 \cdot n^{1/2} (n+1) 2^{-t} \left\| A \right\|_E \\ 0.87 \cdot n^{1/2} (n+1+2.5\,m) 2^{-t} \left\| A \right\|_E \end{cases}$$

de :

$$I - \bar{Q}^t \bar{Q} = - \left(U + U^t \right)$$

se sigue que cuando se acumulan productos internos en doble presición tenemos la estimación aposteriori :(6.8)

$$\left\| I - \bar{Q}^t \bar{Q} \right\|_2 \leq 2 \left\| U \right\|_2 \leq 1.74\, n^{1/2} (n+1) 2^{-t} \left\| A \right\|_E \left\| \bar{R}^{-1} \right\|_2$$

para estimar $\left\| \bar{R}^{-1} \right\|_2$ en :

$$\left\| I - \bar{Q}^t \bar{Q} \right\|_2 \leq 2 \left\| U \right\|_2 \leq 1.74\, n^{\frac{1}{2}} (n+1) 2^{-t} \left\| A \right\|_E \left\| \bar{R}^{-1} \right\|_2$$

usamos la identidad :

$$\bar{R}^t \bar{R} = (A + E_1)^t (A + E_1) - \bar{R}^t (U + U^t) \bar{R}$$

la cual puede ser derivada de :

$$E_1 = \bar{Q} \bar{R} - A, \qquad e_1 = \bar{Q} \bar{Y} + \bar{b}^{(n+1)} - B$$

y :

$$I - \bar{Q}^t \bar{Q} = - (U + U^t)$$

si escribimos :

$$\bar{R}^t R = R^t (I + F_2) R \qquad (6.9)$$

donde :

$$F_2 = F_1 - R^{-t} (U \bar{R}) \bar{R}^t R^{-1} - R^{-t} \bar{R}^t (U \bar{R}) R^{-1}$$

y por :

$$\bar{R}^t R = R^t (I + F_2) R$$

se sigue :

$$\left\| \bar{R}^{-1} R \right\|_2^2 \leq 1 + \left\| F_2 \right\|_2 \qquad (6.10)$$

y de aquí usando :

$$\left\| F_1 \right\|_2 \leq 2 \left\| E_1 \right\|_E \left\| R^{-1} \right\|_2 + \left\| E_1 \right\|_E^2 \left\| R^{-1} \right\|_2^2$$

tenemos la desigualdad :

$$\left\| F_2 \right\|_2 \leq 2 \left\| E_1 \right\|_E \left\| R^{-1} \right\|_2 + \left\| E_1 \right\|_E^2 \left\| R^{-1} \right\|_2^2$$

$$+ 2 \left\| U \bar{R} \right\|_E \left\| R^{-1} \right\|_2 (1 + \left\| F_2 \right\|_2)^{1/2}$$

usando los estimadores derivados para :

$$\left\| E_1 \right\|_E \quad y \quad \left\| U \bar{R} \right\|_E \quad y \text{ la desigualdad :}$$

$(n - 1) \leq 0.3 n^{1/2} - n + 1$ tenemos asumiendo que :

$$\left\| F_2 \right\|_2 \leq 2 \cdot 0.3 \cdot 1.5 c + (0.3 \cdot 1.5 c)^2 +$$

$$+ 2 \cdot 0.87 c (1 + \left\| F_2 \right\|_2)^{1/2}$$

donde :

$$c = N^{1/2} (n + 1) \cdot 2^{-t} \left\| A \right\|_E \left\| R^{-1} \right\|$$

se sigue que : $\left\| F_2 \right\|_2 \leq b$ donde b es la raiz cuadrada positiva de la ecuación

$$b = 0.9\,c + (0.45\,c)^2 + 1.74\,c\,(1 + b)^{1/2}$$

dado que $b = 1$ correspondiente a $c = 1/2$ (0.419...) es realmente verificado que aquí si requerimos:

$$\beta = 3.42c = 3.42\,n^{1/2}\,(n+1)2^{-t}\,\|A\|_E\,\|R^{-1}\|_2 < 1 \qquad (6.12)$$

la cual es más fuerte que la condición :

$$1.5\,(n-1)\,2^{-t}\,\|A\|_E\,\|R^{-1}\|_2 < \sqrt{2} - 1$$

Entonces $\|F_2\|_2 < \beta$ y de :

$$\bar{R}^t \bar{R} = R^t (I+F) R$$

$$\|\bar{R}^{-1}\|_2 \le \frac{1}{1-\beta}\,\|R^{-1}\|_2^2 \qquad (6.13)$$

está desigualdad se toma para presición simple y definimos a β por :

$$\beta = 3.42\,n^{1/2}\,(n+1+2.5\,m)\,2^{-t}\,\|A\|_E\,\|R^{-1}\|_2$$

finalmente de :

$$\|I - \bar{Q}^t \bar{Q}\| \le 2\,\|U\| \le 1.74\,n^{1/2}\,(n+1)2^{-t}\,\|A\|\,\|\bar{R}^{-1}\|$$

y

$$\|\bar{R}^{-1}\|_2 \le \frac{1}{1-\beta}\,\|R^{-1}\|_2^2$$

Tenemos un estimados apriori : (6.14)

$$\left\| I - \bar{Q}^{T} \bar{Q} \right\|_2 \leq \frac{1.74}{d1-\beta} n^{3/2} (n+1) \cdot 2^{-t} \left\| H \right\|_F \left\| R^{-1} \right\|_2$$

Extensivos experimentos computacionales reportan que cuando se usa el método modificado de GRAM-SMIT los resultados de pivoteo se persiven pero son muy pequeños en la ortogonalidad de los vectores calculados.

Cuando m = n un análisis previo indica que la solución calculada satisface la ecuación :

$$(A + H^{(1)}) \cdot x^{(1)} = b + e$$

Obviamente si A está mal condicionada $A + H^{(1)}$ podria ser singular mostramos que si:

$$27\, n^{3/2} \cdot 2^{-t} \left\| A \right\|_2 \left\| A^{-1} \right\|_2 \cdot 2^{-p} \qquad (p = 0)$$

entonces : $\left\| x - x^{(1)} \right\|_2 \div \left\| x \right\|_2 \cdot 2^{-p} \div (1 - 2^{-p-1})$

donde : $x = A^{-1} b$ es la solución real.

de la ecuación :

$$(A + H^{(1)}) \cdot x^{(1)} = b + e$$

$$x^{(1)} = (A + H^{(1)})^{-1} (b + e)$$

Dado que $A + H^{(1)}$ es no singular, dando:

$$\left\| x^{(1)} - x \right\|_2 = \left(\left\| A^{-1} \right\|_2 \left\| H^{(1)} \right\|_2 \left\| x \right\|_2 + \left\| A^{-1} \right\|_e \right) \div \left(1 - \left\| A^{-1} \right\|_2 \left\| H^{(1)} \right\|_2 \right)$$

Escribiendo a $12.5n\,2^{-t}$ y recordando que $\bar{B}_F = \bar{U}_F$ tenemos

$$\left\| H^{(1)} \right\|_F = a \left\| A \right\|_F + (2^{-t} + 3/2\, n\, 2^{-2t})(1 + a) \left\| A \right\|_F$$

y dado que : $\left\| A \right\|_2 \left\| A^{-1} \right\|_2 = 1$, se sigue de :

$$27\, n^{3/2}\, 2^{-t} \left\| A \right\|_2 \left\| A^{-1} \right\|_2 2^{-p} \qquad (p = 0)$$

que $H^{(1)}$ ciertamente satisface el límite :

$$\left\| H^{(1)} \right\|_F = 13.5\, n\, 2^{-t} \left\| A \right\|_F$$

usando las relaciones $\left\| A \right\|_F = n^{1/2} \left\| A \right\|_2$

(una desigualdad muy debil para la mayoria de A).

$$\left\| H^{(1)} \right\|_2 = \left\| H^{(1)} \right\|_F \quad y \quad \left\| e \right\|_2 =$$

$$= 12.5\, n\, 2^{-t} \left\| A \right\|_2 \left\| x \right\|_2$$

tenemos finalmente :

$$\frac{x^{(1)} - x}{\left\| x \right\|_2} = \frac{13.5\, n^{3/2}\, 2^{-t} \left\| A^{-1} \right\|_2 \left\| A \right\|_2}{1 - 13.5\, n^{3/2}\, 2^{-t} \left\| A^{-1} \right\|_2 \left\| A \right\|_2} + \frac{12.5\, n^{-t}\, 2 \left\| A^{-1} \right\|_2 \left\| A \right\|_2}{1 - 13.5\, n^{3/2}\, 2^{-t} \left\| A^{-1} \right\|_2 \left\| A \right\|_2}$$

$$= \frac{26\, n^{3/2}\, 2^{-t} \times (A)}{1 - 13.5\, n^{3/2}\, 2^{-t} \times (A)}$$

Donde x (A) es el número de condición espectral usual. La condición.

$$27\, n^{3/2}\, 2^{-t} \left\| A \right\|_2 \left\| A^{-1} \right\|_2 2^{-p} \qquad (p = 0)$$

ahora da el resultado requerido.

Dado que $x^{(1)}$ está limitado uniformemente por todos los lados derechos esto muestra que la A calculada no está demasiado mal condicionada, se puede tener la certeza de que la solución calculada tendra algunas figuras correctas. El límite :

$$27\, n^{3/2}\, 2^{-t} \left\| A \right\|_2 \left\| A^{-1} \right\|_2 2^{-p} \qquad (p = 0)$$

es extremadamente conservador. La experiencia sugiere que :

$$\left\| x - x^{(1)} \right\|_2 : \left\| x \right\|_2 2^{-p} \div (1 - 2^{-p-1})$$

es verdad cuando p está definido por algunas relaciones tales que :

$$n^{1/2}\, 2^{-t} \left\| A \right\|_2 \left\| A^{-1} \right\|_2 = 2^{-p}$$

dado que A está demasiado mal condicionada el procedimiento del refinamiento iterativo es ciertamente un trabajo con culquier lado derecho, dado que los errores cometidos en calcular el residual no tiene importancia.

Realmente si el residual estaba calculado exactamente, el análisis garantiza que la corrección s - ésima calculada $e^{(s)}$ satisface la relación :

$$\left\| e^{(s)} - (x - x^{(s)}) \right\|_2 \div \left\| x - x^{(s)} \right\|_2 = 2^{-p} \div (1 - 2^{-p-1})$$

si, además, ningun error fue cometido al sumar

$e^{(s)}$ a $x^{(s)}$ tenemos :

$$\left\| x^{(s+1)} - x^{(s)} \right\|_2 = \left\| x - x^{(s+1)} \right\|_2 \leq$$

$$\leq \left\| x - x^{(s)} \right\|_2 \cdot 2^{-p} \cdot \left(1 - 2^{-p-1} \right)$$

En el promedio $x^{(s)}$ se gana aproximadamente el mismo número de dígitos en cada iteración bajo este correcto para el trabajo preciso. Realmente por una modificación sutil del procedimiento es generalmente posible obtener una $x^{(s)}$ de cualquier precisión usando la factorización original de A. En la práctica es adecuado calcular $r^{(s)}$ el residual s - ésimo, usando acumulación de productos internos, dado que si $\bar{r}^{(s)}$ es el residual calculado y $r^{(s)}$ el verdadero residual entonces $\bar{r}^{(s)} = r^{(s)} + f$ donde :

$$\left\| f \right\|_2 \leq \left\| 2^{-t} \cdot \bar{r}^{(s)} \right\|_2 + 3/2\, n^{3/2} \cdot 2^{-2t} \left\| A \right\|_2 \left\| x^{(s)} \right\|_2$$

Este resultado es consecuencia inmediata del análisis de error dado por : WILKINSON.
La acumulación de productos internos no es vital en cualquier otro paso del procedimiento. El fracaso de usarlo donde quiera meramente deja un límite de condición :

$$27\, n^{3/2} \cdot 2^{-t} \left\| A \right\|_2 \left\| A^{-1} \right\|_2 \cdot 2^{-p} \qquad (p = 0)$$

requiriendo el rango de numeros condición para el cual el refinamiento iterativo converderá, y tambien baja la proporción de mejoramiento en algun grado. Dado que $x^{(s)}$ satisface el limite requerido y $x^{(s-1)}$ finalmente alcanzados por lo cual.

$$x^{(s)} - x^{(s-1)} \Rightarrow x = 2$$

a dicho paso x es " correcto para el trabajo preciso".

C A P I T U L O V I I

CONDICION DEL PROBLEMA
DE
MINIMOS CUADRADOS

Antes de estimar los errores en la solución calculada, estudiaremos la sensitividad para perturbaciones en A y b del problema de mínimos cuadrados, el cual siguiendo a GOLUB, escribimos.

$$\begin{bmatrix} r + Ax \\ A^t r \end{bmatrix} = \begin{bmatrix} I & A \\ A^t & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} - \begin{bmatrix} b \\ 0 \end{bmatrix} \qquad (7.1)$$

Sea el sistema perturbado.

$$\begin{bmatrix} I & A + \delta A \\ A^t + \delta A & 0 \end{bmatrix} \begin{bmatrix} r + \delta r \\ x + \delta x \end{bmatrix} = \begin{bmatrix} b + \delta b \\ 0 \end{bmatrix} \qquad (7.2)$$

y sea $\tilde{A} = A + \delta A = \tilde{V} \tilde{R}$, $\tilde{V}^t \tilde{V} = I$

sea la factorización de la matriz perturbada, entonces :

$$\begin{bmatrix} \delta r \\ \delta x \end{bmatrix} = \begin{bmatrix} I & \tilde{A} \\ \tilde{A}^t & 0 \end{bmatrix}^{-1} \left[- \begin{bmatrix} 0 & \delta A \\ \delta A^t & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} + \begin{bmatrix} \delta b \\ 0 \end{bmatrix} \right]$$

Notese que deseamos la solución de mínimos cuadrados del problema original sin turbaciones y frecuentemente se desearía comparar este vector con la solución final x. Este vector z, por supuesto sería calculado por los procedimientos de ortogonalización discutidos anteriormente.

Usando :

$$\begin{bmatrix} I & A + \delta A \\ A^t + \delta A^t & 0 \end{bmatrix} \begin{bmatrix} r + \delta r \\ x + \delta x \end{bmatrix} = \begin{bmatrix} b + \delta b \\ 0 \end{bmatrix}$$

la inversa puede ser calculada rapidamente y tenemos:

$$\begin{bmatrix} \delta r \\ \tilde{R}\ \delta x \end{bmatrix} = \begin{bmatrix} I - \tilde{V}\tilde{V}^t & \tilde{V} \\ \tilde{V}^t & -I \end{bmatrix} \begin{bmatrix} -\delta A x + \delta b \\ -\tilde{R}^{-t} A^t r \end{bmatrix} \qquad (7.3)$$

Los eingenvalores de la matriz simétrica $(I - \tilde{V}\tilde{V}^t)$ són todos iguales a 0 ó 1 de aquí $\left\| I - \tilde{V}\tilde{V}^t \right\|_2 \leq 1$ y

$$\begin{bmatrix} \delta r \\ \tilde{R}\ \delta x \end{bmatrix} = \begin{bmatrix} I - \tilde{V}\tilde{V}^t & \tilde{V} \\ \tilde{V}^t & -I \end{bmatrix} \begin{bmatrix} -\delta A x + \delta b \\ -\tilde{R}^{-t} A^t r \end{bmatrix}$$

nos lleva a:

$$\left\| \begin{bmatrix} \delta r \\ \tilde{R}\ \delta x \end{bmatrix} \right\|_2 \leq \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \left\| \begin{bmatrix} \left\| \tilde{R}^{-t} \right\|_2 \left\| \delta A \right\|_2 \left\| r \right\|_2 \\ \left\| \delta A \right\|_2 \left\| x \right\|_2 \left\| \delta b \right\|_2 \end{bmatrix} \right\|_2 \qquad (7.4)$$

Para estimar $\left\| \tilde{R}^{-1} \right\|_2$ sea la factorización de A, $A = QR$

entonces podemos escribir :

$$\tilde{R}^t \tilde{R} = \tilde{A}^t \tilde{A} = (A + \delta A)^t (A + \delta A) = R^t (I + k) R$$

donde :

$$k = (Q^t \delta A R^{-1})^t + Q^t \delta A R^{-1} + (\delta A R^{-1})^t \delta A R^{-1}$$

Ponemos :

$$\alpha = (\sqrt{2} + 1) \left\| R^{-1} \right\|_2 \left\| \delta A \right\|_2 \qquad (7.5)$$

y asumimos que $\alpha < 1$, entonces, es facilmente mostrar que :

$$\left\| k \right\|_2 \leq 2 \left\| \tilde{R}^{-1} \right\|_2 \left\| \delta A \right\|_2 + \left\| R^{-1} \right\|_2^2 \left\| \delta A \right\|_2^2 < \alpha < 1,$$

y de aquí :

$$\left\| \tilde{R}^{-1} \right\|_2^2 \leq \left\| R^{-1} \right\|_2^2 \left\| (I+k)^{-1} \right\|_2 \left\| R^{-t} \right\|_2 < \frac{1}{1-\alpha} \left\| R^{-1} \right\|_2^2$$

De aquí por : $\kappa(R) = \kappa(QR) = \kappa(A)$

$$\left\| A \right\|_2 \left\| \tilde{R}^{-1} \right\|_2 \leq (1-\delta)^{-\frac{1}{2}} \left\| R \right\|_2 \left\| R^{-1} \right\|_2 = (1-\alpha)^{-\frac{1}{2}} \kappa(A)$$

Dado que $\left\| \delta x \right\|_2 \leq \left\| \tilde{R}^{-1} \right\|_2 \left\| \tilde{R}\, \delta x \right\|_2$ se sigue ahora de

$$\begin{bmatrix} \left\| \delta r \right\|_2 \\ \left\| \tilde{R}\, \delta x \right\|_2 \end{bmatrix} \leq \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \left\| \tilde{R}^{-1} \right\|_2 \left\| \delta A \right\|_2 \left\| r \right\|_2 \\ \left\| \delta A \right\|_2 \left\| x \right\|_2 \left\| \delta b \right\|_2 \end{bmatrix}$$

que :

(7.6)

$$\begin{Vmatrix} \left\| \delta r \right\|_2 \\ \left\| A \right\|_2 \left\| \delta x \right\|_2 \end{Vmatrix} \leq \begin{bmatrix} \dfrac{\kappa(A)}{\sqrt{1-\alpha}} & 1 \\ \dfrac{\kappa^2(A)}{1-\alpha} & \dfrac{\kappa(A)}{\sqrt{1-\alpha}} \end{bmatrix} *$$

$$* \begin{bmatrix} \left\| r \right\|_2 & \dfrac{\left\| \delta A \right\|_2}{\left\| A \right\|_2} \\ \left\| A \right\|_2 \left\| x \right\|_2 & \dfrac{\left\| \delta A \right\|_2}{\left\| A \right\|_2} + \left\| \delta b \right\|_2 \end{bmatrix}$$

Es posible construir ejemplos donde esos límites sean cercanamente obtenidos.

El resultado de :

$$\frac{\|\delta r\|_2}{\|A\|_2 \|\delta x\|_2} \leq \begin{bmatrix} \frac{\kappa(A)}{\sqrt{1-\alpha}} & 1 \\ \frac{2\,\kappa(A)}{1-\alpha} & \frac{\kappa(A)}{\sqrt{1-\alpha}} \end{bmatrix} * \begin{bmatrix} \frac{\|r\|_2}{\|A\|_2} & \frac{\|\delta A\|_2}{\|A\|_2} \\ \|A\|_2 \|x\|_2 & \frac{\|\delta A\|_2}{\|A\|_2} + \|\delta b\|_2 \end{bmatrix}$$

tiene las siguientes interpretaciones interesantes.

notese que : $x = R^{-1} Rx = R^{-1} Q^t Ax$

y así :

$$\|A\|_2 \|x\|_2 + \kappa(A) \|r\|_2 \leq \kappa(A) \left(\|Ax\|_2 + \|r\|_2 \right)$$

$$\leq \sqrt{2}\, \kappa(A) \|b\|_2$$

Entonces del primer renglon de :

$$\begin{Vmatrix} \|\delta r\|_2 \\ \|A\|_2 \|\delta x\|_2 \end{Vmatrix} \leq \begin{bmatrix} \dfrac{\kappa(A)}{\sqrt{1-\alpha}} & 1 \\ \dfrac{\kappa^2(A)}{1-\alpha} & \dfrac{\kappa(A)}{\sqrt{1-\alpha}} \end{bmatrix} *$$

$$\cdot \left[\begin{matrix} \|r\|_2 & \dfrac{\|\delta A\|_2}{\|A\|_2} \\ \|A\|_2 \|x\|_2 \dfrac{\|\delta A\|_2}{\|A\|_2} & + \|\delta b\|_2 \end{matrix} \right]$$

Se sigue que :

$$\frac{\|\delta r\|_2}{\|b\|_2} \leq \sqrt{2}\,\frac{\kappa(A)}{\sqrt{1-\alpha}}\left(\frac{\|\delta A\|_2}{\|A\|_2}+\frac{\|\delta b\|_2}{\|b\|_2}\right) \qquad (7.7)$$

De aquí como número condición para la descomposición de b en componentes ortogonales Ax y r podemos tomar

$$\sqrt{2}\,(1-\alpha)^{-\frac{1}{2}}\,\kappa(A)$$

Si asumimos que $\|x\|_2 \neq 0$, entonces podemos también derivar un número condición para la determinación de x. Del segundo renglon de :

$$\begin{Vmatrix} \|\delta r\|_2 \\ \|A\|_2 \quad \|\delta x\|_2 \end{Vmatrix} \leq \begin{bmatrix} \dfrac{\kappa(A)}{\sqrt{1-\alpha}} & 1 \\ \dfrac{\kappa^2(A)}{1-\alpha} & \dfrac{\kappa(A)}{\sqrt{1-\alpha}} \end{bmatrix} *$$

$$* \begin{bmatrix} \|r\|_2 & \dfrac{\|\delta A\|_2}{\|A\|_2} \\ \|A\|_2 \|x\|_2 & \dfrac{\|\delta A\|_2}{\|A\|_2} + \|\delta b\|_2 \end{bmatrix}$$

se sigue :

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq \frac{\kappa(A)}{\sqrt{1-\alpha}} \left(\frac{1+\kappa(A)}{\sqrt{1-\alpha}} \frac{\|r\|_2}{\|A\|_2 \|x\|_2} \right) \frac{\|\delta A\|_2}{\|A\|_2} +$$

$$+ \frac{\kappa(A)}{\sqrt{1-\alpha}} \frac{\|\delta b\|_2}{\|A\|_2 \|x\|_2}$$

Notese en particular la importancia de la proporción

$$\frac{\|r\|_2}{\|A\|_2 \|x\|_2}$$

si las ecuaciones són cercanamente compatibles en el sentido de que : $\|r\|_2 < \|A\|_2 \|x\|_2$, entonces la situación es mucho muy parecida cuando $m > n$ como cuando $m = n$ y $\kappa(A)$ es un número condición aproximado para la determinación de x.

Sin embargo, si $\|r\|_2$ es de la misma magnitud de $\|A\|_2 \|x\|_2$ entonces el número condición (en un caso mal condicionado) es más que $\kappa^2(A)$.

C A P I T U L O V I I I

ERRORES EN LA SOLUCION CALCULADA

Frecuentemente se desea determinar una x tal que
b - Ax sea minimizado, sujeto a la condición de que
Ax = b.

Como la solución calculada tomamos :

$$\bar{r} = \bar{b}^{(n+1)} \qquad , \qquad \bar{x} = fl((\tilde{\bar{R}}')^{-1}\,\bar{y}') \qquad (8.1)$$

Si definimos R tal que x es la solución exacta de :

$$\tilde{\bar{R}}\bar{x} = (\bar{\bar{R}} + \delta\bar{\bar{R}})\,\bar{x} = \bar{y}$$

entonces usando las definiciónes :

$$E_2 = (I - \tilde{\bar{Q}}\bar{Q}^{t})\,A, \qquad \bar{e}_2 = (I - \tilde{\bar{Q}}\bar{Q}^{t})\,b - \bar{b}^{(n+1)}$$

$$E_3 = \bar{\bar{R}} - \tilde{Q}^{t}\,A, \qquad \bar{e}_3 = y - \tilde{Q}^{t}\,b$$

podemos introducir :

$$\bar{r} = \bar{b}^{(n+1)} \qquad , \qquad \bar{x} = fl((\tilde{\bar{R}}')^{-1}\,\bar{y}')$$

como :

$$\bar{r} = (I - \tilde{\bar{Q}}\bar{Q}^{t})\,b - \bar{e}_2 \,, \qquad \bar{x} = \tilde{R}^{-1}(Q^{t}\,b + \bar{e}_3) \qquad (8.2)$$

sustituyendo b = r + Ax en :

$$\bar{r} = (I - \tilde{\bar{Q}}\bar{Q}^{t})\,b - \bar{e}_2 \,, \qquad \bar{x} = \tilde{R}^{-1}(Q^{t}\,b + \bar{e}_3)$$

Tenemos :

$$\begin{bmatrix} \bar{r} \\ \bar{x} \end{bmatrix} = \begin{bmatrix} I - \bar{Q}\hat{Q}^{t} \left(I - \bar{Q}\tilde{Q}^{t} \right) A \\ \tilde{R}^{-1}\hat{Q}^{t} \quad \tilde{R}^{-1}\tilde{Q}^{t} A \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} + \begin{bmatrix} -e_2 \\ \tilde{R}^{-1} e_3 \end{bmatrix} \qquad (8.3)$$

Ahora dado que $A_b r = 0$ tenemos de :

$$E_1 = \bar{\bar{Q}}\bar{R} - A, \qquad e_2 = \bar{\bar{Q}}y + b^{(n+1)} - b$$

que :

$$\bar{Q}^{t} r = \bar{R}^{t} \left(A + E_1 \right)^{t} r = \bar{R}^{-t} E_1^{t} r$$

y del LEMA (5.1) ver capitulo 5.
Se sigue que :

$$\tilde{Q}^{t} = \left(I + U \right)^{t} \bar{Q}^{-t}, \qquad \tilde{\bar{Q}} = \bar{Q} \left(I + U \right)^{t^{-1}}$$

Así los errores en la solución calculada pueden ser escritos como :

$$\begin{bmatrix} \bar{r} - r \\ \bar{x} - x \end{bmatrix} = \begin{bmatrix} -\tilde{\bar{Q}} \bar{R}^{-t} E_1^{t} & \left(I - \bar{Q}\tilde{Q}^{t} \right) A \\ \left(\bar{R}^{-t} (I+U)^{t} R \right) E_1^{t} & \tilde{R}^{-1} \left(\tilde{Q}^{t} A - R \right) \end{bmatrix} * \begin{bmatrix} r \\ x \end{bmatrix} + \begin{bmatrix} -e_2 \\ \tilde{R}^{-1} e_3 \end{bmatrix} \qquad (8.4)$$

Ahora estimaremos la Norma L de las submatrices
aparecidas en : 2

$$\begin{vmatrix} \bar{r} - r \\ \bar{x} - x \end{vmatrix} = \begin{vmatrix} \hat{Q}\,\bar{R}^{-t}\,E & (I - \bar{Q}\,\bar{Q}^{t})\,A \\ (\bar{R}^{t}(I+U)\,R)^{-1}\,E & R^{-1}\,(Q^{t}\,A - R) \end{vmatrix} * \begin{vmatrix} r \\ x \end{vmatrix} + \begin{vmatrix} -e \\ \tilde{R}^{-1}\,e \end{vmatrix}$$

Asumimos para abreviar que el modo fl ( ) de computacion es usada en ambos, en la descomposición y en la sustitución hacia atras. Los estimadores relevantes cuando usamos presición simple no son muy diferentes y pueden ser derivados de manera similar.
Dado que de acuerdo a :

$$\left\| p^{(j)} \right\|_2 = 1 \qquad \left\| \tilde{q} \right\|_2 \leq 1 \qquad \left\| \hat{q} \right\|_2 \leq 1$$

$\left\| \hat{Q} \right\|_E < n^{1/2}$ tenemos inmediatamente de :

$$\left\| E_1 \right\|_E \leq 1.5\,(n-1) \cdot 2^{-t} \left\| A \right\|_E$$

$$\left\| e_1 \right\|_2 \leq 1.5\,n \cdot 2^{-t} \left\| b \right\|_2$$

$$\left\| E_2 \right\|_E \leq 3.25\,(n-1)\cdot 2^{-t} \left\| A \right\|_E$$

$$\left\| e_2 \right\|_2 \leq 3.75\,n\cdot 2^{-t} \left\| b \right\|_2$$

(8.5)

$$\left\| -Q\bar{R}^{-t}\begin{bmatrix} \tilde{E}^{t} \\ 1 \end{bmatrix} \right\|_2 \leq 1.5\,n^{\frac{1}{2}}\,(n-1)\cdot 2^{-t} \left\| \bar{R}^{-1} \right\|_2 \left\| A \right\|_E$$

(8.6)

$$\left\| (I - \bar{Q}\tilde{Q}^{t})\,A \right\|_2 \leq 3.25\cdot(n-1)\cdot 2^{-t} \left\| A \right\|_E$$

Escribimos el residuo de dos submatrices en :

$$\begin{vmatrix} \bar{r} - r \\ \bar{x} - x \end{vmatrix} = \begin{vmatrix} -\tilde{Q}\,\bar{R}^{-t}\begin{bmatrix} \tilde{E}^{t} \\ 1 \end{bmatrix} & (I - \bar{Q}\,\tilde{Q}^{t})\,A \\ (\bar{R}^{-t}(I+U^{t})\,\bar{R})\,E & \tilde{R}^{-1}(\tilde{Q}^{t}A - \bar{R}) \end{vmatrix} * \begin{vmatrix} r \\ x \end{vmatrix} + \begin{vmatrix} -e_2 \\ \tilde{R}^{-1}e_3 \end{vmatrix}$$

como

$$(\bar{R}^{t}(I + V^{t})\bar{R})\,E = \bar{R}^{t}\bar{R}\,\bar{R}^{-1}(I + V^{t})\bar{R})\,E \qquad (8.7)$$

$$\tilde{R}^{-1}(\tilde{Q}^{t}A - \bar{R}) = \bar{R}\,\bar{R}^{-1}(\tilde{Q}^{t}A - \bar{R}) - \bar{R}^{-1}\delta\bar{R}) \qquad (8.8)$$

Ahora de :

$$\bar{A}^{t}\bar{A} = (A + E_1)^{t}(A + \bar{E}_1) = R^{t}(I + F_1)R$$

$$I - \bar{Q}^{t}\bar{Q} = -(U + U^{t}).$$

se sigue :

$$\bar{R}(I + U^{t})\bar{R} = \bar{A}^{t}\bar{A} - \bar{R}^{t}U\bar{R} = R^{t}(I + F_3)R$$

donde :

$$F_3 = F_1 - R^{-t}\bar{R}^{t}(U\bar{R})R^{-1}$$

comparando esto con :

$$\bar{R}^{t}\bar{R} = R^{t}(I + F_3)R$$

y $$\left\| \bar{R}^{-t} \right\|_2^2 \leq \frac{1}{1 - \beta} \left\| R^{-1} \right\|_2^2$$

observamos que ciertamente $\left\| F_3 \right\| \leq \beta$

y $$\left\| (\bar{R}^{-t}(I + U^{t})\bar{R})^{-1} \right\|_2 \leq \frac{1}{1 - \beta} \left\| R^{-1} \right\|_2^2$$

si ponemos :

$$\left\| \tilde{R}^{-1}\bar{R} \right\|_2 = \left\| (I + \bar{R}^{-1}\delta\bar{R}^{-t})^{-1} \right\|_2 = 1 + r \qquad (8.9)$$

Entonces usando :

$$\left\| E_1 \right\|_E \leq 1.5\,(n-1)\cdot 2^{-t} \left\| A \right\|_E$$

$$\left\| e_1 \right\|_2 \leq 1.5\,n\cdot 2^{-t} \left\| b \right\|_2$$

tenemos de :

$$(\bar{R}^t (I + V^t)R)^{-1} E = R^{-1} \bar{R}^{-t} R^{-t} (I + V^t)\bar{R}^t E$$

$$\left\| (\bar{R}^t (I + U^t)R^{-1})^t E \right\|_2 \leq \frac{1}{1-\beta} 1.5\,(n-1)\,\beta$$

$$* 2^{-t} \left\| R^{-1} \right\|_2^2 \left\| A \right\|_E \qquad (8.10)$$

y de :

$$E_3 = \bar{R} - \tilde{Q}^t A, \qquad \bar{e}_3 = y - \tilde{Q}^t b$$

y de :

$$\tilde{R}^{-1} (\tilde{Q}^t A - \tilde{R}) = \tilde{R}^{-1} \bar{R}^{-1} (\tilde{Q}^t A - \bar{R}) - \tilde{R}^{-1} \delta\bar{R}$$

$$(8.11)$$

$$\left\| \tilde{R}^{-1} (\tilde{Q}^t A - \tilde{R}) \right\|_2 \leq (n+1)\left(\left\| E_3 \right\|_E \left\| R^{-1} \right\|_2 + \left\| \tilde{R}^{-1} \delta\bar{R} \right\|_2 \right)$$

WILKINSON muestra que si la acumulación en doble precisión es usada entonces,

$$\left\| \delta R' \right\|_E \leq (1.001 \cdot 2^{-t} + 3/2\,(n+1)2^{-2t2}) \left\| R' \right\|_E$$

$$\leq 2^{-t1} \left\| \bar{R}' \right\|_E \qquad (8.12)$$

dado que $\left\| \bar{R} \right\|_E \leq n^{1/2} \left\| \bar{R}' \right\|_2$

tenemos :

$$\left\| \bar{R}^{-1} \delta\bar{R} \right\|_2 = \left\| (\bar{R}')^{-1} \delta\bar{R}' \right\|_2 \leq n^{1/2} \cdot 2^{-t1} \kappa(\bar{R}') *$$

$$* \left\| \bar{R} \right\|_2 \leq (1+\beta)^{1/2} \left\| R \right\|_2 \qquad (8.13)$$

cuando $\beta < 1$, se sigue :

$$\left\| \bar{R}^{-1} \delta\bar{R} \right\|_2 \leq n^{1/2} 2^{-t1} \kappa(\bar{R}) \leq (1+\beta)\, n^{1/2} 2^{-t1} \left\| \bar{R}^{-1} \right\|_2 \left\| A \right\|_E \qquad (8.14)$$

sustituyendo en :

$$\left\| \tilde{R}^{-1} (\tilde{Q}^t A - \tilde{R}) \right\|_2 \leq (n+\Gamma)\left(\left\| E_3 \right\|_E \left\| \bar{R}^{-1} \right\|_2 + \left\| \tilde{R}^{-1} \delta\bar{R} \right\|_2 \right)$$

de :

$$\left\| E \right\|_E \leq 1.9 \left(n - 1 \right)^{3/2} n \cdot 2^{-t} \left\| A \right\|_E$$

$$\left\| e \right\|_2 \leq 1.9 \, n^{3/2} \left(n + 1 \right) \cdot 2^{-t} \left\| b \right\|_2$$

y :

$$\left\| \bar{R}^{-1} \delta \bar{R} \right\|_2 \leq n^{3/2} \, 2^{-t1} \, \kappa(\bar{R}) \leq (1 + \beta)^{3/2} \, n^{3/2} \, 2^{-t1} \left\| \bar{R}^{-1} \right\|_2 \left\| A \right\|_E$$

podemos demostrar que :

$$\left\| \tilde{R}^{-1} \left(Q^t A - \tilde{R} \right) \right\|_2 \leq (1 + r) \cdot 1.9 \, n^{3/2} (n + 1) \, \beta \\ * 2^{-t} \left\| \bar{R}^{-1} \right\|_2 \left\| A \right\|_E \qquad (8.15)$$

Para la estimación de :

$$\left\| \tilde{R}^{-1} \bar{R} \right\|_2 = \left\| \left(I + \bar{R}^{-1} \delta \bar{R} \right)^{-1} \right\|_2 = 1 + r$$

nos lleva a:

$$r \leq \left\| \bar{R}^{-1} \delta \bar{R} \right\|_2 \div \left(1 - \left\| \bar{R}^{-1} \delta \bar{R} \right\|_2 \right) \qquad (8.16)$$

de :

$$\left\| \bar{R}^{-1} \delta \bar{R} \right\|_2 \leq n^{3/2} \, 2^{-t1} \, \kappa(\bar{R}) \leq (1 + \beta)^{3/2} \, n^{3/2} \, 2^{-t1} \left\| \bar{R}^{-1} \right\|_2 \left\| A \right\|_E$$

$$\left\| \bar{R}^{-1} \right\|_2 \leq \frac{1}{1-\beta} \left\| R^{-1} \right\|_2$$

y

$$\beta = 3.42\, c = 3.42\, n^{\frac{1}{2}} (n + 1)\, 2^{-t} \left\| A \right\|_2 \left\| R^{-1} \right\|_2 < 1$$

tenemos :

$$\left\| \bar{R}^{-1} \delta R \right\|_2 \leq 1.06 \left(\frac{1+\beta}{1-\beta} \right) \cdot n^{\frac{1}{2}}\, 2^{-t} \left\| R^{-1} \right\|_2 \left\| E \right\|_E \leq$$

$$\leq \frac{1.06\, \beta}{3.42 (n + 1)} \left(\frac{1+\beta}{1-\beta} \right)^{\frac{1}{2}}$$

y si hácemos la suposición adicional $\beta \leq 0.9$ entonces,

$$\left\| \bar{R}^{-1} \delta \bar{R} \right\|_2 \leq 1.22 / (n + 1) \leq 0.41$$

usando esto para estimar el denominador en :

$$r \leq \left\| \bar{R}^{-1} \delta \bar{R} \right\|_2 \div \left(1 - \left\| \bar{R}^{-1} \delta \bar{R} \right\|_2 \right)$$

podemos mostrar que :

$$r \leq 1.8\, n^{\frac{1}{2}}\, 2^{-t}\, \varkappa (\bar{R}') \qquad (8.17)$$

$$r \leq \frac{0.53}{n+1} \left(\frac{1+\beta}{1-\beta} \right)^{\frac{1}{2}} \beta \qquad (8.18)$$

Aunque no generalmente es verdad que $\varkappa(\bar{R}') \leq \varkappa (A)$

ocasionalmente $\varkappa(\bar{R}') < (\bar{R}') < \varkappa(\bar{R}) \approx \varkappa(A)$.
En efecto, para muchos sistemas muy mal condicionados

$\varkappa(\bar{R}')$ serán de orden unitario.

Entonces :

$$\left\| \bar{R}^{-1} \delta\bar{R} \right\|_2 \leq n^{1/2} \, 2^{-t+1} \, \kappa(\bar{R}) \leq (1+\beta)^{1/2} \, n^{1/2} \, 2^{-t+1} \left\| \bar{R}^{-1} \right\|_2 \left\| A \right\|_E$$

y consecuentemente de :

$$r \leq \frac{0.53}{n+1} \left(\frac{1+\beta}{1-\beta} \right)^{1/2} \beta$$

se sobreestimará considerablemente el error en la sustitución hacia atras, el cual siempre es de poca importancia.

Ahora tomaremos normas en :

$$\begin{vmatrix} \bar{r} - r \\ \bar{x} - x \end{vmatrix} = \begin{vmatrix} \tilde{Q}^{t} \, \delta\bar{R}^{t} \, \bar{E}_1 & (I - \bar{Q} \tilde{Q}^{t}) A \\ (\bar{R}^{t} (I+U)^{t} \bar{R}) \, \bar{E}_1 & \tilde{R}^{-1} \, (\tilde{Q}^{t} A - \tilde{R}) \end{vmatrix} *$$

$$* \begin{vmatrix} r \\ x \end{vmatrix} + \begin{vmatrix} -e_2 \\ \tilde{R}^{-1} e_1 \end{vmatrix} \qquad (8.4)$$

y usando :

(8.5)

$$\left\| \begin{matrix} \tilde{Q}^{t} \, \delta\bar{R}^{t} \, \bar{E}_1 \\ \bar{E}_1 \end{matrix} \right\|_2 \leq 1.5 \, n^{1/2} \, (n-1) \cdot 2^{-t} \left\| \bar{R}^{-1} \right\|_2 \left\| A \right\|_E$$

(8.6)

$$\left\| (I - \bar{Q} \tilde{Q}^{t}) A \right\|_2 \leq 8.25 \cdot (n - 1) \cdot 2^{-t} \left\| A \right\|_E$$

$$\left\| \left(\bar{R}^{t} \left(I + U^{t} \right) \tilde{R}^{-1} \right) E^{t} \right\|_2 \leq \frac{1}{1-\beta} 1.5 (n-1) *$$

$$* 2^{-t} \left\| R^{-1} \right\|_2^2 \left\| A \right\|_E \qquad (8.10)$$

y

$$\left\| \hat{R}^{-1} \left(Q^{t} A - \tilde{R}^{t} \right) \right\|_2 \leq (1 + \Gamma) \cdot 1.9 n^{\frac{1}{2}} (n+1) *$$

$$* 2^{-t} \left\| \bar{R}^{-1} \right\|_2 \left\| A \right\|_E \qquad (8.15)$$

junto con los estimadores para $\left\| e_2 \right\|_2$ y $\left\| e_3 \right\|_2$

en :

$$\left\| E_2 \right\|_E \leq 3.25 (n-1) \cdot 2^{-t} \left\| A \right\|_E$$

$$\left\| e_2 \right\|_2 \leq 3.25 n \cdot 2^{-t} \left\| b \right\|_2$$

y

$$\left\| \bar{R}^{-1} \right\|_2 \left\| A \right\|_2 \leq (1-\beta)^{-\frac{1}{2}} \times (A)$$

eliminando : $\left\| \bar{R}^{-1} \right\|_2$ por medio de la desigualdad :

$$\left\| \bar{R}^{-1} \right\|_2 \left\| A \right\|_2 \leq (1 - \beta)^{-\frac{1}{2}} \cdot \varkappa(A)$$

y retomando las desigualdades :

$$n - 1 \leq 0.3n^{\frac{1}{2}} (n + 1), \qquad n \leq 2/3\, n^{\frac{1}{2}} (n + 1)$$

y llegamos al resultado :

$$\left\| \begin{matrix} \left\| \bar{r} - r \right\|_2 \\ \left\| A \right\|_2 \left\| \bar{x} - x \right\|_2 \end{matrix} \right\| \leq \begin{bmatrix} \dfrac{0.79 \cdot \varkappa(A)}{\sqrt{1 - \beta}} & 0.81 \\ \dfrac{0.24\, \varkappa^2(A)}{1 - \beta} & \dfrac{\varkappa(A)}{\sqrt{1 - \beta}} \end{bmatrix} *$$

$$* \begin{bmatrix} \left\| r \right\|_2 \dfrac{\left\| A \right\|_E}{\left\| A \right\|_2} \\ \left\| A \right\|_2 \left\| x \right\|_2 \dfrac{\left\| A \right\|_E}{\left\| A \right\|_2} + \left\| b \right\|_2 \end{bmatrix} \cdot f(n) 2^{-t} \qquad (8.19)$$

donde :

$$f(n) = (1 + r)\, 1.9\, n^{\frac{1}{2}} (n + 1)$$

Observamos que :

$$\frac{\|\bar{r} - r\|_2}{\|A\|_2 \|\bar{x} - x\|_2} \leq \begin{bmatrix} \dfrac{0.79 - \varkappa(A)}{\sqrt{1 - \beta}} & 0.81 \\ \dfrac{0.24\ \varkappa^2(A)}{1 - \beta} & \dfrac{\varkappa(A)}{\sqrt{1 - \beta}} \end{bmatrix} *$$

$$* \begin{bmatrix} \|r\|_2 \dfrac{\|A\|_E}{\|A\|_2} \\ \|A\|_2 \|x\|_2 \dfrac{\|A\|_E}{\|A\|_2} + \|b\|_2 \end{bmatrix} \cdot f(n) 2^{-t}$$

da un límite más pequeño para el error que aquel

obtenido en :

$$\left\| \begin{array}{cc} \left\| \delta r \right\|_2 & \\ \left\| A \right\|_2 \left\| \delta x \right\|_2 & \end{array} \right\| \leq \left[\begin{array}{cc} \dfrac{\kappa(A)}{\sqrt{1-\alpha}} & 1 \\ \dfrac{\kappa^2(A)}{1-\alpha} & \dfrac{\kappa(A)}{\sqrt{1-\alpha}} \end{array} \right] *$$

$$* \left[\begin{array}{cc} \left\| r \right\|_2 & \dfrac{\left\| \delta A \right\|_2}{\left\| A \right\|_2} \\ \left\| A \right\|_2 \left\| x \right\|_2 & \dfrac{\left\| \delta A \right\|_2}{\left\| A \right\|_2} + \left\| \delta b \right\|_2 \end{array} \right]$$

cuando :

$$\left\| \delta A \right\|_2 = f(n) \cdot 2^{-t} \left\| A \right\|_E , \qquad \left\| \delta b \right\|_2 = f(n)\, 2^{-t} \left\| b \right\|_2$$

y esto de acuerdo a :

$$\beta = 3.42\, c = 3.42\, n^{\frac{1}{2}} (n + 1)\, 2^{-t} \left\| A \right\|_2 \left\| R^{-1} \right\|_2 < 1$$

y :

$$\alpha = (\sqrt{2}+1) \left\| R^{-t} \right\|_2 \left\| \delta A \right\|_2$$

implica : (nota 161)

$$\alpha = (1 + \delta)(\sqrt{2} + 1) \cdot 1.3\, n^{1/2} (n + 1) \times$$

$$\times\, 2^{-t} \left\| R^{-1} \right\|_2 \left\| A \right\|_E > \beta$$

Como el factor (1 + δ) esencialmente es de importancia, podemos decir que la desviación de la solución, de la verdadera solución es aproximadamente menor que la desviación resultante, de alguna perturbación δA, δb tal que :

$$\left\| \delta A \right\|_E \div \left\| A \right\|_E \approx \left\| \delta b \right\|_2 \div \left\| b \right\|_2 \approx 2 \cdot n^{1/2}\, 2^{-t}$$

Además cuando A está mal condicionada podemos, como remarcado en el final de la solución 5, esperan un resultado mucho mayor que este.
Finalmente observamos que cuando el procedimiento modificado es usado para computar R, entonces el procedimiento clásico, no será contado para ser usado para calcular y si esto fue hecho tendremos :

$$Y \approx \bar{Q}^{-t}\, b = (\bar{Q} - \tilde{Q})^{t}\, b + \tilde{Q}^{t}\, b$$

y consecuentemente :

$$\left\| y - \tilde{Q}^{t} \right\|_2 \approx \left\| \bar{Q} - \tilde{Q} \right\|_2 \left\| b \right\|_2 =$$

$$= \left\| U \right\|_2 \left\| b \right\|_2 = \text{Const.} \times (A) 2^{-t} \left\| b \right\|_2$$

puede ser verificado que si $\| r \|_2 < \| b \|_2$, esto significa que los errores en la solución calculada generalmente serán multiplicados por $\kappa(A)$.

C O N C L U C I O N E S

El procedimiento modificado de GRAM-SCHMIDT ha sido usado sucesivamente, desde 1960 como parte del sistema de datos procesados para espectros opticos.

Tiene más ó menos la misma gran estabilidad numérica como otros algoritmos comparables basados en transformaciones ortogonales y generalmente será preferible el procedimiento clasico de GRAM-SCHMIDT. La reortogonalización es entonces no siempre necesaria, y la solución es obtenida en aproximadamente $m n^2$ multiplicaciones. El pivoteo puede ser hecho facilmente, pero no es requisito para la estabilidad,el requerimiento de almacenage es aproximadamente :

$$m n + n^2 \div 2$$

Esto seria comparado a $2 m n + 4 n^2 \div 3$ multiplicaciones y n^2 lugares de almacenage necesitadas para formar y resolver las ecuaciones normales en doble precisión.

Soluciones más precisas para problemas de Mínimos Cuadrados pueden ser obtenidas por usar el refinamiento iterativo, este fue el primero propuesto por GOLUB y usado tambien en BAHUER. Esquemas diferentes para el refinamiento son discutidos en las notas de refinamiento iterativo de minimos cuadrados. Los resultados obtenidos son sin embargo, solo satisfactorios cuando las ecuaciones son tambien compatibles es decir cuando :

$$\varepsilon \, \kappa_2(A) \left\| r \right\|_2 < \left\| A \right\|_2 \left\| x \right\|_2$$

En notas posteriores de BAHUER donde tambien se resuelven ejemplos numéricos se pplica el refinmamiento iterativo a probemas de minimos cuadrados para calcular los residuales del sistema :

$$\begin{Vmatrix} r + Ax \\ A^t r \end{Vmatrix} = \begin{Vmatrix} I & A \\ A^t & 0 \end{Vmatrix} \begin{Vmatrix} r \\ x \end{Vmatrix} = \begin{Vmatrix} b \\ 0 \end{Vmatrix}$$

en precisión doble, y usando la inversa :

$$\begin{Vmatrix} I & A \\ A^t & 0 \end{Vmatrix}^{-1} \approx \begin{Vmatrix} I - \bar{Q}\,\bar{Q}^t & \bar{Q}\,\bar{R}^{-t} \\ \bar{R}^{-1}\,\bar{Q}^t & \bar{R}^{-1}\,(I + U)^{-t}\,\bar{R}^{-t} \end{Vmatrix}$$

Para los incrementos en r y x. Esta inversa resulta como una pequeña generalización del procedimiento modificado de GRAM-SCHMIDT y como se mostró que la iteración generalmente convergerá a la solución exacta, correctamente redondeada cuando :

$$\chi^2(A) \left\| r \right\|_2 \text{ es del mismo orden de magnitud que :}$$

$$2^t \left\| A \right\|_2 \left\| x \right\|_2$$

B I B L I O G R A F I A

1.- Bauer, F. L. Optemally Scaled Matrices
Numerische Mathematik. Vol .5 1963
p. 73 - 87

2.- Golub. G. H. & J. H. Wilkinson.- Note on the Iterative Refenement of Least Squares Solution
Numerische Mathemalik 1966 p. 139 - 148

3._ Strang, Gilbert.- Linear Algebra.- Academic Press,
Nuw - York. 1980

4.- Valdez Bautista Betatriz.- Tesis Profesional
" Aspectos de Cálculo del Algebra Lineal "
Cd. Universitaria de México 1988, p.100 - 135