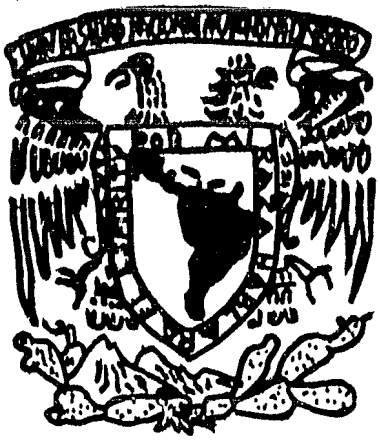


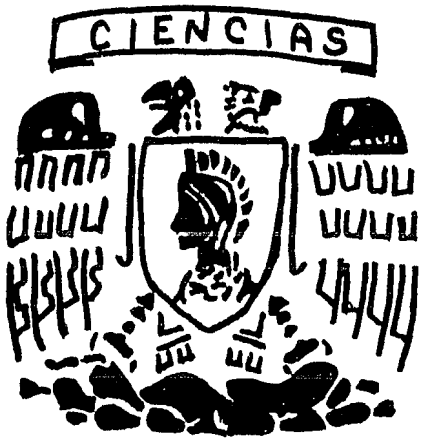
UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

ASPECTOS DE CALCULO(

DEL

ALGEBRA LINEAL



T E S I S

QUE PARA OBTENER EL TITULO DE :

A C T U A R I A

P R E S E N T A

BEATRIZ VALADEZ BAUTISTA

MEXICO, D.F. 1988

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

I N D I C E

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

Este estudio de aspectos del Cálculo en Algebra Lineal está dirigido a personas que sin ser especialistas en Análisis Numérico, con frecuencia se enfrentan a este tipo de cálculo en la práctica. El objetivo de este trabajo será conducir a este tipo de personas a aspectos generales del Algebra Lineal referentes a los cálculos matriciales, incluyendo normas para conocer el pocedimiento de la Eliminación Gausiana, así mismo se darán exposiciones intrínsicas de los métodos de cálculo del Algebra Lineal.

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

Este estudio de aspectos del Cálculo en Algebra Lineal está dirigido a personas que sin ser especialistas en Análisis Numérico, con frecuencia se enfrentan a este tipo de cálculo en la práctica. El objetivo de este trabajo será conducir a este tipo de personas a aspectos generales del Algebra Líneal referentes a los cálculos matriciales, incluyendo normas para conocer el pocedimiento de la Eliminación Gausiana, así mismo se darán exposiciones intrínsicas de los métodos de cálculo del Algebra Líneal.

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

C A P I T U L O

I.

M A T E R I A L E L E M E N T A L

D E

A L G E B R A L I N E A L

I) MATERIAL ELEMENTAL DE ALGEBRA LINEAL

1.1.- MATRICES.

Un conjunto de números arreglados en forma rectangular, es llamada una matriz rectangular. Este arreglo puede tener m renglones y n columnas, y se puede representar en la forma.

$$1) \qquad A = \begin{vmatrix} a_{11} & a_{12} & a_{1n} \\ a_{21} & a_{22} & a_{2n} \\ a_{m1} & a_{m2} & a_{mn} \end{vmatrix},$$

donde el subindice m describe los renglones y el subindice n describe las columnas; asi un punto ó elemento de la matriz es localizado por renglón y columna.

Esto puede ser abreviado de la forma :

$$A = (a_{ij}) \quad (i = 1,2,...m \; ; \; j = 1,2,...n)$$

Dos matrices son iguales si sus correspondientes elementos son iguales.

Las matrices compuestas solo de un renglón son llamadas Vector Renglón.

Las matrices compuestas solo de una columna son llamadas Vector Columna. Si el número de renglones de una matriz es igual al número de columnas entonces se dice que es una matriz cuadrada.

Entre las matrices cuadradas una regla importante son las Matrices Diagonales, es decir, matrices en las cuales todos los elementos son cero a excepción de los elementos principales de la diagonal, los cuales son diferentes de cero.

$$2)\quad \begin{vmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & a_n \end{vmatrix} = (a_1 \; a_2 \; \cdots \; a_n)$$

Si todos los números componentes de la diagonal son iguales, entonces se dice que la Matriz es Escalar.

$$3)\quad \begin{vmatrix} a & 0 & \cdots & 0 \\ 0 & a & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & a \end{vmatrix} = a$$

Y si a es igual a 1, se dice que es una Matriz Unitaria.

$$4)\quad \begin{vmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & 1 \end{vmatrix} = I.$$

Por último una matriz en la cual todos sus elementos son cero se le llama Matriz Nula, y la designaremos por el número cero.

El determinante cuyos elementos son los elementos de una Matriz Cuadrada, se dice que es el Determinante de esta Matriz, y escribimos el determinante de la Matriz, A como, $|A|$, o bien $d(A)$.

MULTIPLICACION DE UNA MATRIZ POR UN NUMERO; LA ADICION DE MATRICES.

Una matriz cuyos elementos son obtenidos por multiplicar todos los elementos de una Matriz A por una escalar a es llamado producto del número a y la Matriz A.

$$5) \qquad ca = \begin{vmatrix} ca_{11} & ca_{12} & \cdots & ca_{1n} \\ ca_{21} & ca_{22} & \cdots & ca_{2n} \\ \vdots & & & \\ ca_{m1} & ca_{m2} & \cdots & ca_{mn} \end{vmatrix}$$

Una matriz C, cuyos elementos son las sumas de los elementos correspondientes de A y B, matrices que tienen igual número de columnas y renglones se dice que es la suma de A y B :

$$6) \quad A + B = \begin{vmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & & & \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{vmatrix}$$

Las operaciones que se han introducido arriba tiene las siguientes propiedades:

1.- $A + (B + C) = (A + B) + C$

2.- $A + B = B + A$

3.- $A + 0 = A$

4.- $(r + s)A = rA + sA$ donde r, s son números

5.- $r(A + B) = rA + rB$

Aquí, A, B, C, son matrices, r, s son números, generalmente.

1.3.- LA MULTIPLICACION DE MATRICES.

La multiplicación de dos matrices A y B, es definida siempre que el número de columnas de A sea igual al número de renglones de B. Así los elementos del producto C = AB, están definidos de la siguiente manera; el elemento en el i-ésimo renglón y la j-ésima columna de la matriz C es igual a la suma de los productos de los elementos del i-ésimo renglón de la matriz A por los elementos correspondientes de la j-ésima columna de la matriz B, así.

7) AB

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{vmatrix} \quad \begin{vmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & & b_{2n} \\ \vdots & & & \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{vmatrix}$$

$$c_{ij} = \begin{vmatrix} c_{11} & c_{12} & . & . & . & c_{1p} \\ c_{21} & c_{22} & & & & c_{2p} \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ c_{m1} & c_{m2} & & & & c_{mp} \end{vmatrix} = C, \text{ donde}$$

$$(8)\quad c_{ij} = a_{i1}\,b_{1j} + a_{i2}\,b_{2j} + \ldots + a_{in}\,b_{nj} = \Sigma\, a_{ik}\,b_{kj}$$

$(i = 1,2, \ldots m; \quad j = 1,2, \ldots p)$.

Notese que el producto de dos matrices rectangulares es una Matriz Rectangular, cuando no es una matriz rectangular, entonces nos da una matriz en la cual el número de renglones, es igual al número de renglones de la primera matriz (que se multiplica) y el número de columnas es igual al número de columnas de la segunda matriz es decir :

$$A_{m*n} \cdot B_{n*p} = C_{m*p}.$$

Asi tenemos que el producto de una matriz cuadrada por una matriz compuesta por una sola columna, nos da una matriz de una columna.

El producto de las matrices AB y BA se puede hacer solo si el número de renglones de la primera matriz es igual al número de columnas de la segunda, y el número de columnas de la primera es igual al número de renglones de la segunda, dada ésta última condición, se sigue que las matrices deben de ser cuadradas, asi el producto AB y BA, es posible siempre que A y B sean matrices cuadradas, en caso contrario y en general tenemos que AB =/= BA.

En casos particulares la multiplicación puede ser conmutativa, y por ejemplo, las matrices conmutan con cualquier matriz cuadrada del mismo orden, así.

$$\begin{vmatrix} a & 0 & \cdots & 0 \\ 0 & a & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & a \end{vmatrix} \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

$$= \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \begin{vmatrix} a & 0 & \cdots & 0 \\ 0 & a & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a \end{vmatrix}$$

$$= \begin{vmatrix} aa_{11} & aa_{12} & \cdots & aa_{1n} \\ aa_{21} & aa_{22} & \cdots & aa_{2n} \\ \vdots & & & \\ aa_{n1} & aa_{n2} & \cdots & aa_{nn} \end{vmatrix}$$

Asi se sigue que para una Matriz unitaria la multiplicación por otra Matriz (cuadrada) es conmutativa, esto es,

$$AI = IA = A .$$

ASOCIATIVIDAD DE MATRICES.- La multiplicación de Matrices es asociativa, si AB y (AB)C, es decir, si multiplicamos primero AB y luego efectuamos la multiplicación por C, también se puede hacer primero la operación BC y después hacer la multiplicación A (BC).

entonces,

$$i) \qquad A(BC) = (AB)C$$

El producto de matrices tiene la siguientes propiedades :

$$ii) \qquad a(AB) = (aA)B = A(aB)$$

$$iii) \qquad (A + B)C = AC + BC$$

$$iv) \qquad C(A + B) = CA + CB$$

Donde A, B, C son matrices y a un número.

TRANSPUESTA.- La transpuesta de una Matriz se obtiene intercambiando los renglones y columnas de una matriz A.

$$A = \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \\ a_{m1} & \cdots & a_{mn} \end{vmatrix} = (a_{ij})$$

entonces obtenemos la transpuesta de la matriz como :

$$A' = A^{t} = (a_{ij})' = \begin{vmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & & a_{m2} \\ \vdots & & & \\ a_{1n} & a_{2n} & & a_{mn} \end{vmatrix} = (a_{ji})$$

Asi, tenemos la siguiente regla para las matrices transpuestas,

$$(AB)'' = B' A'$$

En conclución podemos remarcar que el determinante de un producto de matrices cuadradas es igual al producto de los determinantes de las matrices multiplicadas :

$$|AB| = |A||B|,$$

y este resultado es tomado de la teoria de determinantes.

PARTICION DE MATRICES.- Al tomar matrices de orden grande, requiere como regla un gran número de operaciones, y es conveniente reducir los cálculos de matrices de orden grande a cálculos con matrices de menor orden. Tal reducción puede ser efectuada por particionar las matrices dadas; cada matriz puede ser concebida como compuesta de varias matrices de menor orden, y esta subdivisión puede ser llevada a cabo de varias maneras, por ejemplo.

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{vmatrix} = \left| \begin{array}{c:ccc} a_{11} & a_{12} & a_{13} & a_{14} \\ \hdashline a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{array} \right| = \left| \begin{array}{cc:cc} a_{11} & a_{12} & a_{13} & a_{14} \\ \hdashline a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{array} \right|$$

Las matrices es las cuales se subdivide la matriz dada son llamadas submatrices.

Las operaciones básicas con matrices particionadas en las cuales las matrices diagonales son del mismo orden, se relacionan de manera natural con las matrices, entonces si tenemos.

$$A = \begin{vmatrix} A_{11} & A_{12} & \cdots & A_{1K} \\ A_{21} & A_{22} & \cdots & A_{2K} \\ \vdots & & & \\ A_{K1} & A_{K2} & \cdots & A_{KK} \end{vmatrix}$$

y

$$B = \begin{vmatrix} B_{11} & B_{12} & \cdots & B_{1K} \\ B_{21} & B_{22} & & B_{2K} \\ \vdots & & & \\ B_{K1} & B_{K2} & & B_{KK} \end{vmatrix}$$

Donde A_{ii} y B_{ii} son matrices cuadradas del mismo orden, entonces

$$A + B = \begin{vmatrix} A_{11} + B_{11} & A_{12} + B_{12} & \cdots & A_{1K} + B_{1K} \\ A_{21} + B_{21} & A_{22} + B_{22} & \cdots & A_{2K} + B_{2K} \\ \cdots & \cdots & \cdots & \cdots \\ A_{K1} + B_{K1} & A_{K2} + B_{K2} & \cdots & A_{KK} + B_{KK} \end{vmatrix}$$

Un caso especial muy importante de una matriz particionada, es la matriz bordeada. Teniendo una matriz cuadrada de orden n-1 :

$$A_{n-1} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1,n-1} \\ a_{21} & a_{22} & \cdots & a_{2,n-1} \\ \vdots & & & \\ a_{n-1,1} & a_{n-1,2} & \cdots & a_{n-1,n-1} \end{vmatrix}$$

formamos una matriz cuadrada de orden n, A_n, al agregar a la matriz A_{n-1} un renglón : $v = (a_{n-1} , \ldots , a_{n,n-1})$, una columna : $u = (a_{1n} , \ldots , a_{n-1,n})$, y un numero a_{nn} :

$$A_n = \begin{bmatrix} & & a_{1n} \\ A_{n-1} & & a_{2n} \\ & & \vdots \\ & & a_{n-1,n} \\ a_{n1} \ \cdots \ a_{n,n-1} & & a_{nn} \end{bmatrix} = \begin{bmatrix} A_{n-1} & u_{n-1} \\ v_{n-1} & a_{nn} \end{bmatrix}$$

Diremos que la matriz A ha sido obtenida por bordear la matriz A_{n-1} y la matriz A es particionable naturalmente.

Las operaciones bajo matrices bordeadas son realizadas de acuerdo con las reglas generales de operaciones con matrices particionadas.

sean $A = \begin{vmatrix} M & u \\ v & a \end{vmatrix}$ y $B = \begin{vmatrix} P & y \\ x & b \end{vmatrix}$

dos matrices bordeadas de orden n, entonces los siguientes exposiciones son válidas:

i) $aA = \begin{bmatrix} aM & au \\ av & aa \end{bmatrix}$

ii) $A + B = \begin{vmatrix} M + P & u + y \\ v + x & a + b \end{vmatrix}$

iii) $AB = \begin{vmatrix} MP + ux & My + ub \\ vP + ax & vy + ab \end{vmatrix}$

aqui MP y ux son matrices de orden (n-1); My, ub, son columnas compuestas de n-1 elementos; vP y ax son renglones análogos y, por último vy + ab, es un número.

Matriz Quasi-diagonal, son de la forma :

$$\begin{bmatrix} A & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & B & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & C \end{bmatrix}$$

Donde las celdas de la matriz son obviamente.

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} ; \quad B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} ; \quad C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

y las otras seis celdas son nulas.

El determinante de una matriz Quasi diagonal es el producto de los determinantes de las celdas diagonales.

El producto de dos matrices Quasi-diagonales es una matriz quasi-diagonal

MATRICES INVERSAS Y ADJUNTA.- Si una matriz cuadrada $A = (a_{ij})$ se dice que es no singular si su determinante es distinto de cero, en caso contrario la matriz A es singular.

Un concepto muy importante de una matriz inversa es ahora introducido.

Una matriz B es llamada la inversa ó reciproca de una matriz A si

$$AB = I.$$

Una condición necesaria y suficiente para la existencia de la matriz inversa es la no singularidad de la matriz A.

entonces, $AB = I$, $|A||B| = I$, y consecuentemente $A \neq 0$

Supongamos ahora que $A \neq 0$, para construir la inversa de la matriz primero definamos la adjunta de una matriz. i.e., la matriz.

$$C = \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & & & \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix}$$

Aqui A_{ij} es el complemento algebraico (cofactor de los elementos a_{ij}) en el determinante de la matriz A, i.e., es desginado por el menor determinante de los elementos a_{ij} notese que esta puesta en posición transpuesta.

y tiene la siguiente propiedad (la matriz adjunta)

$$AC = |A| I$$

De esta igualdad se sigue, $$B = \frac{1}{|A|} C$$

como A es no singular entonces, se tiene ,

$$AB = A = \frac{1}{|A|} C = \frac{1}{|A|} AC = I$$

La matriz asi construida tiene tambien la propiedad.

$$BA = I$$

La matriz inversa es unica, si esta existe.

Polinomios en una matriz.- Definiremos la potencia entera positiva de una matriz cuadrada, haciendo.

$$\underbrace{A.A.A.A.A.A. \ldots . A}_{n \text{ veces}} = A^n$$

en vista de la ley asociativa, es evidente de la definicion que,

$$A^{n}\,A^{m} = A^{n+m}$$

$$(A^{n})^{m} = A^{nm}$$

de aqui se sigue que las potencias de una misma matriz son conmutativas.

Y por definición tenemos que

$$A^{0} = I$$

Una expresion de la forma

$$\alpha_0 A^{n} + \alpha_1 A^{n-1} + \ldots + \alpha_n I$$

donde α_0 , α_1 , ... α_n , son números complejos, es llamada un Polinomio en una matriz ó bien Matriz Polinomial. Esta matriz polinomial puede ser tomado, como el resultado de retomar la variable γ en un polinomio algebraico.

$$\mu(\gamma) = \alpha_0 \gamma^{n} + \alpha_1 \gamma^{n-1} + \ldots + \gamma_n$$

Por la matriz A.
Es importante notar que las reglas de operación para Matrices Polinomiales no difieren de las reglas de operación para polinomios algebraicos, :

$$\text{Dado } \mu(\gamma) = \Psi(\gamma) \pm \chi(\gamma)$$
$$w(\gamma) = \Psi(\gamma)\chi(\gamma)$$

$$\text{entonces } \mu(A) = \Psi(A) \pm \chi(A)$$
$$w(A) = \Psi(A)\chi(A)$$

Esto se sigue de la conmutatividad de las potencias de una matriz.

1.7.- EL POLINOMIO CARACTERISTICO.

El teorema de Cay Ley - HAMILTON. El polinomio mínimo
La ecuación.

$$\begin{vmatrix} a_{11}-\gamma & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-\gamma & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn}-\gamma \end{vmatrix} = 0$$

Es llamada la ecuación caracteristica, ó singular de la matríz $A = (a_{ij})$
El miembreo izquierdo de esta ecuación, el cual puede ser escrito en la forma abreviada (A - γI), recibe el nombre de polinomio característico la función característica de la matriz. Las ecuaciónes características son frecuentemente encontradas en matematicas aplicadas.

El cáculo directo de la función característica presenta dificultades técnicas considerables.

$$\text{Si } \Phi(\gamma) = a - \gamma I = (-1)^n (\gamma^n - p_1 \gamma^{n-1} - p_2 \gamma^{n-2} - \cdots - p_n)$$

Entonces,

$$p_1 = a_{11} + a_{22} + \cdots + a_{nn}$$

$$p_n = (-1)^{n-1} A ,$$

Y los coeficiéntes restantes p son las sumas tomadas con el signo $(-1)^{k-1}$ de todos los menores principales del

determinante de la matriz A de orden K es decir, de los menores involucrados en la diagonal principal. El número de tales menores es igual al número de combinaciónes de n cosas tomadas de K en K.

Las raices de la ecuación característica, son llamados los números propios (de raices) características valores propios,ó eigen valores de la matríz A, del bíen conocido teorema de Vieta dando la relación entre las raices de una ecuación y sus coeficientes tenemos.

$$\gamma_1 + \gamma_2 + \ldots + \gamma_n = p_1 = a_{11} + a_{22} + \ldots + a_{nn}$$

$$\gamma_1 \gamma_2 \ldots \gamma_n = (-1)^{n-1} p_n = A.$$

La cantidad $p_1 = a_{11} + a_{22} + \ldots + a_{nn}$ es llamada la traza de la matríz A, y se denota com trA.

Para cualquier matríz cuadrada la siguiente relación conocida como el teorema CAYLEY - HAMILTON, obtenemos que sí $\mu(\gamma)$ es el polinomio característico de la matríz A, entonces $\mu(A) = 0$, que es en un sentido que la matríz es una raiz de su propia ecuación característica.

Para probarlo concideremos la matríz B, la adjunta de la matríz $A - \gamma I$, dado que cada cofactor en el determinante $| A - \gamma I |$ es un polinómio en γ de grado que no exceda n-1, la matríz adjunta puede ser representada como un polinómio algebraico, con coeficiéntes matriciales, es decir, en la forma.

$$B = B_{n-1} + B_{n-2} \gamma + \ldots B_0 \gamma^{n-1}$$

donde $B_{n-1}, \ldots, B_0$, son ciertas matrices no dependientes de γ.

En lo fuerte de la propiedad fundamental de la matríz adjunta tenemos

$$(B_{n-1} + B_{n-2}\gamma + \ldots + B_0\gamma^{n-1})(A - \gamma I) = |A - \gamma I| I$$

$$= (-1)^n(\gamma^n - p_1\gamma^{n-1} - \ldots - p_n)I$$

esta ecuación es equivalente al sistema de ecuaciónes

$$\begin{cases} B_{n-1}A = (-1)^{n+1} p_n I \\ B_{n-2} A - B_{n-1} = (-1)^{n+1} p_{n-1} I \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ B_0 A - B_1 = (-1)^{n+1} p_1 I \\ -B_0 = (-1)^n I \end{cases}$$

multiplicando estas ecuaciónes por la derecha por $I, A, A^2, \ldots, A^{n-1}, A^n$, respectivamente, y sumando obtenemos una matríz nula en el lado izquierdo y en la derecha.

$$(-1)^n(-p_n I - p_{n-1}A - p_{n-2}A^2 - \ldots - p_1 A^n) = \mu(A).$$

Así $\mu(A) = 0$ la cúal es lo que se quería probar.

La relación CAYLEY-HAMILTON muetra que para una matríz cuadrada dada un polinómio existe, para la cúal es una

raíz. Evidentemente tal polinómio no es único sí $\Psi(\gamma)$, tiene tal propiedad,tiene cualquier polinómio divisible por $\Psi(\gamma)$ el polinómio de menor grado teniéndo la propiedad de que la matríz A es un cero del polinómio, y es llamdo el polinómio minimo de A.

En seguida se probará que el polinómio caracteristíco es divisible por el polinómio mínimo.

Sean $q(\gamma)$ $r(\gamma)$, el coeciénte y recidúo obtenidos al dividir el polinómio caracteristico $\mu(\gamma)$ por el polinómio minimo $\Psi(\gamma)$:

$$\mu(\gamma) = \Psi(\gamma)\, q(\gamma) + r(\gamma),$$

El grado $r(\gamma)$ siendo por supuesto menor que el grado de $\Psi(\gamma)$, sustituyendo A por γ en esta ecuación tenemos

$$r(A) = \mu(A) - \Psi(A)\, q(A) = 0$$

Así la matríz A prueba que es un cero del polinómio $r(\gamma)$ de aquí se sigue que $r(\gamma) = 0$ dado que de otro modo $\Psi(\gamma)$ no seria el polinómio minimo, consecuente mente $\Psi(\gamma)$ divide a $\mu(\gamma)$.

1.8 Matrices Semejantes. Decimos que la matríz B, es semejante a la matríz A si existe una matríz C no singular tal

que $B = C^{-1} AC$

La matríz B se puede obtener de la matríz A por una transformacion semejante.
La transformacíon semejanza tiene las siguientes propiedades.

$$1.-\ C^{-1}A_1C + C^{-1}A_2C + \ldots + C^{-1}A_nC = C^{-1}(A_1 + A_2 + \ldots + A_n)C.$$

$$2.-\ C^{-1}A_1C.C^{-1}A_2C \ldots C^{-1}A_nC = C^{-1}(A_1A_2\ldots A_n)C.$$

En particular, $(C^{-1} AC)^n = C^{-1} A^n C$.

De aquí,

3.- $f(C^{-1} AC) = C^{-1} f(A) C$, para cualquier polinómio $f(\gamma)$

de la última propiedad se sigue directamente que las matrices semejantes tienen la misma función minima.

Mostraremos que las matrices semejantes tiene también la misma función caracteristica.
tenemos

$$| B - \gamma I | = | C^{-1} AC - \gamma I | = | C^{-1} AC - \gamma C^{-1} IC |$$

$$= \left[C \right]^{-1} \left[A - \gamma I \right] \left[C \right] = \left[A - \gamma I \right].$$

1.9.- Transformaciónes Elementales.
Frecuentemente es necesario efectuar las siguientes operaciónes con matrices.

a).- Multiplicación de los elementos de algunos renglones por un número.,

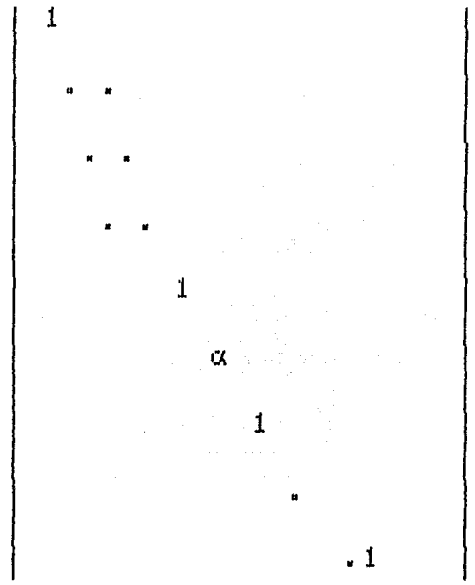
b).- Sumando a los elementos de algún renglon, números proporciónales a los elementos de algún renglón siguiente.

b").- Sumando a los elemntos de algún renglón números proporciónales a los elementos de algún renglón precedente.

Algunas veces estas transformaciónes pueden ser hechas en las columnas. Las transformaciónes del tipo indicado son llamadas tranformaciónes elementales de la matríz.

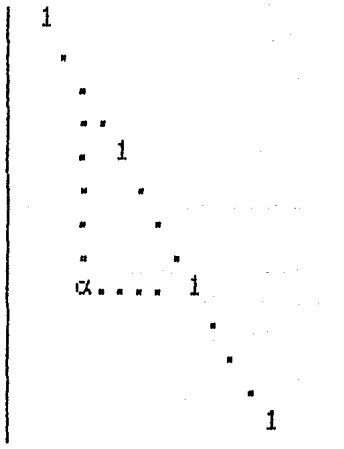
Cualquier transformación elemental de los renglones es equivalente a una Premultiplicación de la matríz (Multiplicación por la izquierda) por una matríz no singular de una forma especial.

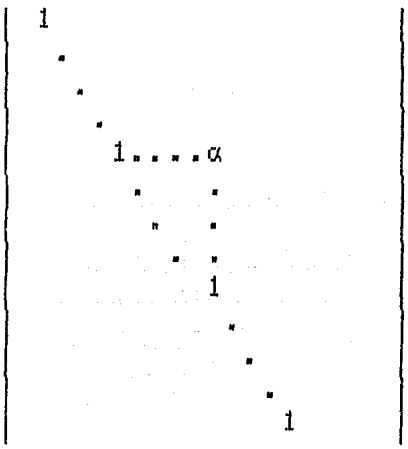
La operación a es equivalente a una premultiplicación de la matríz por la matríz.



La operación b' es equivalente a una premultiplicación por la matríz.

La operación b" es equivalente a una premultiplicación por la matríz.





Las operaciónes a, b', b". son ejecutadas en las columnas al usar tales matrices elementales en las post multiplicaciónes de la matríz bajo las transformaciónes

(33) ahora efectúando las operaciónes B en las columnas y
(34) efectuando B.

Ejemplos de operaciónes por renglones.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & b & c \\ x & y & z \\ u & v & w \end{bmatrix} = \begin{bmatrix} a & b & c \\ ax & \alpha y & \alpha z \\ u & v & w \end{bmatrix};$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \alpha & 0 \end{bmatrix} \begin{bmatrix} a & b & c \\ x & y & z \\ u & v & w \end{bmatrix} = \begin{bmatrix} a & b & c \\ x & y & z \\ u+\alpha x & v+\alpha y & w+\alpha z \end{bmatrix};$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \alpha \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a & b & c \\ x & y & z \\ u & v & w \end{bmatrix} = \begin{bmatrix} a & b & c \\ x+\alpha u & y+\alpha v & z+\alpha w \\ u & v & w \end{bmatrix}$$

En el trabajo futuro tendremos que ejecutar transformaciones de los tipos a y b' sobre matrices.

El resultado de varias transformaciones de ese tipo es equivalente a la premultiplicación de la matriz por algúna matríz triangular, es decir, una de la forma

$$\begin{bmatrix} \neq_{11} & 0 & \cdots & 0 \\ \neq_{21} & \neq_{22} & \cdots & 0 \\ \neq_{n1} & \neq_{n2} & \cdots & \neq_{nn} \end{bmatrix}$$

Con elementos deiagonales $\neq_{ii} =/= 0$, es una matríz triangular inferiór

1.10.- Descomposición de matrices en el producto de dos matrices triangulares.

Las matrices triangulares, esto es, matrices de la forma.

$$\begin{bmatrix} c_{11} & 0 & \cdots & 0 \\ c_{21} & c_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \quad y \quad \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ 0 & b_{22} & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & b_{nn} \end{bmatrix}$$

$$\text{con } c_{ii} \neq 0,\ b_{ii} \neq 0$$

El producto de dos matrices triangulares de la misma estructura nos da otra vez una matríz triangular de la misma forma, una matríz no singular triangular es invertible facilmente y su inversa es de la misma forma.

Los siguientes teoremas son por consiguiente de interes.

TEOREMA. Una condición para que las submatrices resultantes de la matríz A sean no singulares es decir que, A pueda ser representada como el producto de una matríz triangular superior y una matríz triangular inferior,

Demostración :

Tenemos por inducción que para n = 1 es obvio que
$(a_{11}) = (b_{11})(c_{11})$, y uno de los factores pueda ser tomado arbritariamente.

Luego para (n - 1) el teorema es verdadero, para una matríz de orden (n - 1).

Ahora se mostrará que para una matríz de orden n la partición de la matríz A es una matríz bordeada.

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{n2} \end{bmatrix} = \begin{bmatrix} & & a_{1n} \\ & A_{n-1} & \vdots \\ & & \vdots \\ a_{n1} & a_{n2} \cdots a_{n,n-1} & a_{nn} \end{bmatrix} = \begin{bmatrix} A_{n-1} & u \\ v & a_{nn} \end{bmatrix}$$

Buscamos una descomposición A + CB de la matriz A en el producto de dos matrices B y C de las formas requeridas, primero tomando estás matrices particiónadas en la forma bordeada como la de A :

$$C = \begin{vmatrix} C_{n-1} & 0 \\ x & c_{nn} \end{vmatrix} \qquad B = \begin{vmatrix} B_{n-1} & y \\ 0 & b_{nn} \end{vmatrix}$$

Por regla de multiplicación para matrices particiónadas.

$$CB = \begin{vmatrix} C_{n-1} & 0 \\ x & c_{nn} \end{vmatrix} \begin{vmatrix} B_{n-1} & y \\ 0 & b_{nn} \end{vmatrix} = \begin{vmatrix} C_{n-1}B_{n-1} & C_{n-1}Y \\ xB_{n-1} & xy+c_{nn}b_{nn} \end{vmatrix} = A$$

De aquí tenemos :

$$C_{n-1}B_{n-1} = A_{n-1}$$

Ahora bien, tales matrices triangulares, C_{n-1} y B_{n-1}, Existen por hipótesis de inducción.

Además la suposición de que $|A_{n-1}| \neq 0$ se sigue que

$$|C_{n-1}| \neq 0 \quad y \quad |B_{n-1}| \neq 0$$

Ahora bien x, y son encontradas por medio de las formulas.

$$y = C^{-1}_{n-1} u \quad , \qquad x = v B^{-1}_{n-1}$$

Y con estas son determinados de manera unica en terminos de u y v así solo nos resta determinar loe elementos diagonales c y b de la ecuación.

$$c_{nn} \quad b_{nn} = a_{nn} - xy$$

Esta última ecuación muestra que uno de los elementos diagonales puede ser tomado arbitrariamente.
Así la descompisición de la Matríz en el producto de dos matrices triangulares, superior e inferior, sera unica si damos valores a los elementos diagonales de una de las matrices triangulares.
Es conveniente considerar, por ejemplo que sí,

$$b_{ii} = 1 \text{ , con } i = 1, \ldots, n$$

entonces ,

$$c_{nn} = a_{nn} - xy$$

De acuerdo con la matríz C será unicamente determinado.

1.11.- Notación Matricial para un Sistema de Ecuaciónes Líneales, concideremos el sistema de ecuaciónes líneales con n incognitas.

$$(1) \quad \begin{cases} a_{11} x_1 + a_{12} x_2 + \ldots + a_{1n} x_n = b_1 \\ a_{21} x_1 + a_{22} x_2 + \ldots + a_{2n} x_n = b_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1} x_1 + a_{n2} x_2 + \ldots + a_{nn} x_n = b_3 \end{cases}$$

Utilizando la definición de multiplicación de matrices, el sistema puede ser escrito como una simple ecuación en notación matrícial.

(2)
$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{bmatrix}$$

ó simplemente como $Ax = b$

Donde A es la matríz de coeficientes del sistema, b la columna de términos constantes, y x la columna cuyos elementos son las incógnitas.
Sí la matríz del sistema es no singular, obtenemos la solución del sistema (1) por premultiplicar (2) por A^{-1}

$$x = A^{-1} b = \frac{1}{|A|} B\, b$$

Donde B es la matríz adjunta de A y esta última expresión es la notación matrícial para la Regla de Cramer.

$$x_i = \frac{|A_i|}{|A|}$$

Donde A_i es la matríz que se obtiene de A por sustituir los elementos a_{ki} de la i - ésima columna por las componentes b_i de b.

Realmente la ecuación matricial.

$$x = A^{-1} b = \frac{1}{|A|} Bb$$

es equivalente a las n-ecuaciónes.

$$x_i = \frac{A_{1i} b_1 + A_{2i} b_2 + \ldots + A_{ni} b_n}{A} \qquad \text{con } i=1, \ldots n$$

Dado que los A_{ki} son los cofactores de los elementos a_{ki}, en el determinante de la matríz A.
Tenemos obviamente .

$$A_{1i} b_1 + A_{2i} b_2 + \ldots + A_{ni} b_n = |A_i|$$

y con esto queda probada nuestra exposición.

1.12.- Espacio Vectorial n-dimensiónal.
Un punto X de éste espacio es un arreglo de n números reales en un orden definido.

$$X = (x_1 , x_2 , \ldots , x_n)$$

X es también llamado vector n-dimensional y los números $x_1 , x_2 , \ldots x_n$ son las componentes del vector y n es la dimensión del espacio.

Dos vectores son iguales solo si sus componentes correspondientes son iguales.
Las operaciónes fundamentales con vectores están definidas a continuación.

si $X = (x_1 , x_2 , \ldots x_n)$ y $Y = (y_1 , y_2 , \ldots y_n)$

Son dos vectores n-dimensiónales y a es un número complejo arbritario entonces tenemos por definición.

$$X + Y = (x_1 + y_1 , x_2 + y_2 , \ldots , x_n + y_n)$$

$$aX = (ax_1 , ax_2 , \ldots , ax_n)$$

La suma de vectores sastiface las leyes asociativa y conmutativa .

$$X + Y = Y + X$$

$$(X + Y) + Z = X + (Y + Z)$$

En el espacio n-dimensiónal, el producto escalar está definido de acuerdo con la fórmula.

$$(X , Y) = \sum_{k=1}^{n} X_k \bar{Y}_k$$

Donde $\bar{Y}_k$ designa el conjugado complejo de Y_k

y el producto escalar tiene las siguientes propiedades.

1).- $(X X) \neq 0$ si, $X \neq 0$; $(X X) = 0$, si $X = 0$

2).- $(X Y) = \overline{(Y X)}$

3).- $(X_1 + X_2 , Y) = (X_1 Y) + (X_2 Y)$

4).- $(aX, Y) = a (X Y)$

5).- $(X , Y_1 + Y_2) = (X Y_1) + (X Y_2)$

6).- $(X a Y) = \bar{a} (X Y)$

En resumen $\sqrt{(x,x)}$ es llamada la longitud del vector, en lo que sigue lo designaremos como $|| x ||$.
También el espacio n-dimensiónal complejo, introducido arriba, es usado para considerar el espacio real n-dimensiónal, es decir arreglos de vectores con componentes reales.

En el espacio real, el producto escalar es igual a la suma de los productos de las componentes correspondientes de los vectores.
La longitud de un vector es igual a la raíz cuadrada de la suma de los cuadrados de sus componentes. A menudo tendremos que trabajar con el espacio real n-dimensiónal, mientras que el espacio complejo solo en algunas ocasiones se requiere.

1.13.- Dependencia Líneal.

Un vector $Y = c_1 X_1 + c_2 X_2 + \ldots + c_m X_m$, decimos que es una combinación líneal de los vectores $X_1, X_2, \ldots Xm$.
Es fácil ver que si los vectores $Y_1, \ldots Y_k$, son combinaciónes líneales de los vectores $X_1, \ldots X_m$, cualquier combinación líneal $\mu_1 Y_1 + \ldots \mu_k Y_k$ será tambien una combinación de los vectores $X_1, X_2, \ldots X_m$.

Los vectores $X_1, X_2, \ldots X_m$, se dice que son linealmente dependientes, si existen constantes $c_1, c_2 .. c_m$ no todas cero, tales que la ecuación

$$c_1 X_1 + c_2 X_2 + \ldots + c_m X_m = 0 \qquad (5)$$

se cumple. Sin embargo, si esta ecuación se cumple solo cuando todas las constantes c_i son iguales a cero los vectores $X_1, X_2, \ldots X_m$ se dice que son linealmente independientes.

Si los vectores $X_1, \ldots X_m$ son linealmente dependientes, entonces al menos uno de ellos será una combinación líneal del resto.

Por ejemplo, $c_m \neq 0$, encontramos de (5)

$$X_m = \frac{c_1}{c_m} X_1 - \ldots - \frac{c_{m-1}}{c_m} X_{m-1}$$

1.14.- Bases.

Un sistema de vectores linealmente independiente constituye una base para un espacio si cualquier vector del espacio es una combinación líneal de los vectores del sistema.

Un ejemplo de una base, es el conjunto de vectores.

$$\left\{ \begin{array}{l} e_1 = (1, 0, \ldots 0) \\ e_2 = (0, 1, \ldots 0) \\ \ldots\ldots\ldots\ldots \\ e_n = (0, 0, \ldots 1) \end{array} \right.$$

Y es obvio que para cualquier vector $X = (x_1, x_2, \ldots x_n)$ tenemos que

$$X = x_1 e_1 + x_2 e_2 + \ldots + x_n e_n .$$

Y a esta expresión la llamaremos, la base inicial del espacio ó base canónica. Y dicha base no es la única posible, sino todo lo contrario. El número de vectores que forma una base no depende de la selección de la misma. La prueba de esto es. Sean $Y_1, \ldots Y_k$ y $Z_1, \ldots Z_m$, dos bases, donde $k > m$.
Los vectores $Y_1, \ldots Y_k$, son combinación de los líneal de los vectores $Z_1, \ldots Z_m$, entonces $Y_1, \ldots Y_k$, son

linealmente dependientes, lo cual contradice la definición de base, dado que k = m.
Adicionalmente, dado que la base inicial esta constituida por n vectores, cualquier otra base consistirá también de n vectores. El número de vectores que forma una base, coincide con la dimensión del espacio.

Sea $U_1, \ldots, U_n$, los vectores que forman una base para el espacio. Cualquier vector X será entonces una conbinación lineal de $U_1, \ldots, U_n$:

$$X = \epsilon_1 U_1 + \epsilon_2 U_2 + \ldots + \epsilon_n U_n .$$

Los coeficientes en esta expresión unicamente define al vector X, para

$$X = \epsilon_1 U_1 + \ldots + \epsilon_n U_n = \epsilon'_1 U_1 + \ldots \epsilon'_n U_n ,$$

entonces $(\epsilon_1 - \epsilon'_1)U_1 + \ldots + (\epsilon_n - \epsilon'_n)U_n = 0$

y por lo tanto

$$\epsilon_1 - \epsilon'_1 = 0 , \ldots 0, \epsilon_n - \epsilon'_n = 0$$

En vista de la independencia líneal de los vectores $U_1 ,..U_n$ los coeficientes $\epsilon_1, \ldots \epsilon_n$, son llamados las coordenadas del vector X con respecto a la base $U_1, \ldots, U_n$. Nótese que las componentes de un vector $x_1, \ldots x_n$ son las coordenadas del vector X con respecto a la base inicial.

1.15.- Sistemas Ortogonales de Vectores.

Los vectores distintos de cero de un espacio se dice que son ortogonales si su producto escalar es igual a cero. Un sistema de vectores se dice que es ortogonal si cualquiera dos vectores son ortogonales mutuamente. Al referirnos a un sistema ortogonal, debemos asumir que todos los vectores de este sistema son diferentes de cero.

Teorema.- Los vectores que forman un sistema ortogonal son linealmente independientes.

Teorema.- Sean $X_1, \ldots X_n$, linealmente independientes.
Un sistema ortogonal de vectores $X_1, \ldots X_n$ puede ser construido si este es relacionado con el conjunto original por las relaciones :

$$X'_1 = X_1$$

$$X'_2 = X_2 + \alpha_{21} X_1$$

. . . . . .

$$X'_k = X_k + \alpha_{k1} + \ldots + \alpha_{k,k-1} X_{k-1}$$

La prueba se realiza por inducción

De manera que se puede pasar de cualquier sistema ortogonal de vectores al sistema ortonormal correspondiente, dividiendo cada vector por su longitud (norma)
El proceso descrito permite una gran elasticidad en la elección de una base ortonormal, de una se puede pasar de cualquier base a una ortonormal por otorgonalización y normalización.
El producto escalar de dos vectores es muy fácil de expresar en terminos de las coordenadas de esos vectores con respecto a cualquier base ortonormal, si para $U_1 \ldots U_n$ una base ortonormal, y

$$X = \epsilon_1 U_1 + \ldots + \epsilon_n U_n, \quad Y = \eta_1 U_1 + \ldots + \eta_n U_n$$

entonces

$$(X,Y) = (\epsilon_1 U_1 + \ldots + \epsilon_n U_n, \eta_1 U_1 + \ldots + \eta_n U_n)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} (\epsilon_i U_i, \eta_j U_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_i \eta_j (U_i U_j) = \sum_{i=1}^{n} \epsilon_i \eta_i$$

Así la expresión del producto escalar en terminos de vectores (de sus coordenadas) con respecto a cualquier base ortonormal coincide con su expresión en términos de las componentes de los vectores, es decir, en términos de las coordenadas con respecto a la base inicial.

1.16.- Transformación de Coordenadas.

Veamos ahora el cambio en las coordenadas de un vector que acompaña un cambio de base.
Sean $e_1, e_2, \ldots e_n$, y $e'_1, e'_2 \ldots e'_n$ dos bases, y sean

$$e'_1 = a_{11} e_1 + a_{21} e_2 + \ldots + a_{n1} e_n$$

$$e'_2 = a_{12} e_1 + a_{22} e_2 + \ldots + a_{n2} e_n$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$e'_n = a_{1n} e_1 + a_{2n} e_2 + \ldots + a_{nn} e_n$$

La relación con la transformación de coordenadas en la matríz A, donde las columnas son las coordenadas de los vectores $e'_1, e'_2, \ldots e'_n$, con respecto a la base $e_1, e_2, \ldots e_n$, es decir la matríz.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

Donde la matriz A es no singular, para que tenga inversa, por medio de la cual los vectores $e_1, e_2, \ldots e_n$, son expresados en términos de los vectores $e'_1, e'_2, \ldots e'_n$,

Ahora designamos por $x_1, x_2, \ldots x_n$, las coordenadas del vector X con respecto a la base $e_1, e_2, \ldots e_n$ y por $x'_1, x'_2, \ldots x'_n$, sus coordenadas con respecto a la base $e'_1, e'_2 e'_n$. Determinemos la relación de dependencia entre las coordenadas anteriores y las nuevas. Donde por la independencia lineal de los vectores $e_1, e_2, \ldots e_n$:

$$x_1 = a_{11} x'_1 + a_{12} x'_2 + \ldots + a_{1n} x'_n$$

$$x_2 = a_{21} x'_1 + a_{22} x'_2 + \ldots + a_{2n} x'_n$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$x_n = a_{n1} x'_1 + a_{n2} x'_2 + \ldots + a_{nn} x'_n$$

Y esta ecuación puede ser escrita en la forma,

$$x = Ax'$$

donde

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad y \quad x' = \begin{bmatrix} x'_1 \\ x'_2 \\ \cdot \\ \cdot \\ \cdot \\ x'_n \end{bmatrix}$$

Son las coordenadas de las columnas del, vector X con respecto a las bases $e_1, e_2, \ldots e_n$ y $e'_1, e'_2, \ldots e'_n$ respectivamente.

1.17.- Subespacios.

Un subespacio de un espacio vectorial es un subconjunto que satisface dos condiciones.

i).- Si sumamos dos vectores cualesquiera x e y del subespacio su suma x+y permanece en el subespacio.

ii).- Si multiplicamos cualquier vector x del subespacio por cualquier escalar c, el multiplo cx permanece en el subespacio,

En otras palabras, un subespacio es un conjunto " cerrado " bajo la suma y la multiplicación por un escalar; estas operaciones nos mantienen dentro del espacio, sin sacarnos. No es necesario verificar las ocho propiedades del Espacio Vectorial ya que estas se satisfacen en el espacio mayor y por ende se satisfacen en cualquier subespacio. Nótese que en particular el vector cero pertenece a cada subespacio debido a la propiedad ii) : basta elegir el escalar c = 0

1.18.- Rango de una Matríz.

Dado que una base de un espacio vectorial es un conjunto de vectores con dos propiedades simultaneas :

1).- Es linealmente independiente

2).- Genera espacio

Entonces dos bases cualesquiera de un espacio vectorial V contienen el mismo número de vectores. Este número, compartido por todas las bases, y que expresa el número de " grados de libertad " del espacio, se llama dimensión de V.

Ahora bien, supongamos que reducimos la matríz A de m x n, mediante operaciónes elementales e intercambio de filas a una matríz triangular escalonada. Habrá entonces r elementos distintos de cero; las últimas m-r filas de la matríz escalonada son cero. Entonces habrá r variables básicas y n-r variables libres, correspondientes a las columnas de la matríz escalonda. El espacio nulo formado por las soluciónes de Ax = 0 tiene las n-r variables libres como parámetros independientes.
Si n = r, no hay variables libres y el espacio nulo sólo contiene ax = 0.

Existen soluciónes para cada lado derecho b si y solo si r = m con este número de pivotes la matríz escalonada no tiene filas iguales a cero y puede resolverse mediante sustitución regresiva. En el caso de que r > m, la matríz escalonada tendrá m-r filas iguales a cero y habrá m-r restricciones para b de modo que Ax = b se puede resolver. Si existe una solución particular, entonces todas las demás soluciones difieren de la particular en un vector del espacio nulo de A.

En otras palabras el rango de una matríz es un número r tal que entre los menores de la matríz existe un menor distinto de cero de orden r, pero todos los menores de orden r +1 y más grande son iguales a cero.
Y es válido el siguiente teorema importante.

Teorema ;

El número máximo de renglones linealmente independientes de una matríz, como también el número máximo de columnas linealmente independientes, coinciden con el rango de una matríz.

1.19.- Tansformaciones Líneales.

Asociaremos a cada vector X de un espacio un cierto vector Y del mismo espacio. Tal asociación la llamaremos una transformación del espacio. Designaremos el resultado de la aplicación de transformación A al vector X por AX.
Y llamaremos la trasnformación A líneal si

1).- $A(\gamma X) = \gamma AX$ para cualquier número complejo γ ;

2).- $A(X_1 + X_2) = AX_1 + AX_2$.

Ahora definiremos las operaciónes sobre las transformaciónes líneales.

El producto de las transformaciones líneales A y B, AB = C, será una transformación constituida por las transformaciones B y A por turnos, primero se completará la transformación B y entonces se llevará a cabo A.

El producto de transformaciones líneales es una transformación líneal.

La suma de transformaciones líneales A y B será una transformación C la cual asocia el vector X con el vector AX + BX, y la suma de estas es asimismo, una transformación líneal.

Representación de una Transformación Líneal por una Matríz.
Sea una transformación líneal.

$$L : K^n \longrightarrow K$$

entonces existe un vector A en K^n , tal que $L = L_A$, es decir tal que para todo X tenemos.

$$L(X) = AX$$

Sean $E_1, \ldots, E_n$, los vectores unitarios en K^n .

Si $X = x_1 E_1 + \ldots + x_n E_n$ es cualquier vector, entonces

$$L(X) = L(x_1 E_1 + \ldots + x_n E_n)$$

$$= x_1 L(E_1) + \ldots + x_n L(E_n)$$

Si hacemos $a_i = L(E_i)$

vemos que $L(X) = x_1 a_1 + \ldots + x_n a_n = XA$

De modo que las componentes de A son, presisamente los valores : $L(E_1), \ldots, L(E_n)$, donde E_i, donde i= 1,..n
son los vectore unitarios de K.
En general, tenemos en términos de los vectores columna.

$$L(E_1) = \begin{bmatrix} a_{11} \\ . \\ . \\ . \\ a_{m1} \end{bmatrix}, \ldots, L(E_n) = \begin{bmatrix} a_{1n} \\ . \\ . \\ . \\ a_{mn} \end{bmatrix}$$

Por tanto.

$$L(X) = x_1 (a_{11} e_1 + \ldots + a_{m1} e_m) + \ldots + x_n (a_{1n} e_1 + \ldots + a_{mb} e_m)$$

$$= (a_{11} x_1 + \ldots + a_{1n} x_n) e_1 + \ldots + (a_{m1} x_1 + \ldots + a_{mn} x_n) e_m$$

En consecuencia si hacemos $A = (a_{ij})$, entonces se ve que

$$L(X) = AX$$

Desarrollando completamente, ésta expresión tiene la siguiente forma.

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{bmatrix}$$

Así L es la aplicación lineal asociada con la matríz A.

También designamos a A como la matríz asociada a la transformación líneal L, y se sabe que esta matríz está determinada de manera única.

1.20.- El Rango de Una Tranformación Líneal.

Sea A una cierta transformación líneal, el conjunto de vectores AX constituirá obviamente un subespacio, el cual denotaremos por AR^n

La dimensión de esta espacio se dice que es el rango de la transformación A.

Mostraremos ahora que el rango de una transformación es igual al rango de la matríz correspondiente a esta transformación en cualquier base, $e_1, e_2, \ldots, e_n$.

Obviamente el subespacio AR^n está constituido por los vectores

$Ae_1, Ae_2, \ldots, Ae_n$, es decir, la dimensión de AR^n es igual al rango de la matríz cuyas columnas están compuestas de las coordenadas de los vectores $Ae_1, Ae_2, \ldots, Ae_n$, esto es, el rango de la matríz correspondiente a la transformación.

Dado que la dimensión de un subespacio no depende de la elección de la base, se sigue que los rangos de matrices semejantes son iguales.

1.21.- Los Vectores Propios de una Transformación Lineal.

Por un vector propio (vector característico, ó eigenvector) de una transformación lineal A por medio de un vector X diferente de cero tal que.

$$AX = \gamma X$$

Donde γ es un número.
El número γ es llamado el número propio (valor propio, ó eigenvalor) de la transformación.

Los valores propios y vectores propios de la transformación pueden ser determinados de la siguiente manera. Sea A la transformación relacionada con la matríz $A = (a_{ik})$ con respecto a alguna base; sea X el vector propio de coordenadas, con respecto a esta base $x_1, \ldots, x_n$.

Este sistema de ecuaciónes homogéneo en $x_1, \ldots, x_n$, tendrá una solución diferente de cero en el caso solo.

$$\begin{bmatrix} a_{11} - \gamma & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \gamma & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \& \end{bmatrix} = 0$$

es decir, si γ es un cero del polinomio característico de la matríz.
Así los siguiente es válido.

TEOREMA.

Los valores propios de una transformación coinciden con los ceros del polinomio característico de una matríz que está relacionado con ésta transformación con respecto a una base arbitraria.

Ejemplo.
Consideremos la transformación líneal con la matriz.

$$A = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$$

Entonces $| A - \gamma I | = (\gamma - 3)^2$, y así $\gamma = 3$ es una raíz doble del polinomio característico.
El sistema de ecuaciones para determinar las coordenadas del vector propio de la transformación A, será :

$$3x_1 + x_2 = 3x_1$$

$$3x_2 = 3x_2$$

Donde $x_2 = 0$, y así todos los vectores propios de la transformación en cuestion serán $(x_1 , 0) = x_1 (1 , 0)$.

Propiedades de los valores propios y los vectores propios de una matríz.

1).- Una matríz y su transpuesta tienen polinomios característicos identicos.

2).- Para una matríz real simetrica, las propiedades de ortogonalidad son simples gracias al hecho de que todos sus valores propios son reales.

3).- Los vectores propios que pertenecen a valores propios distintos son ortogonales, y viceversa.

1.22.- Valores Propios de una Forma Cuadrática Positiva Definida.

Un polinomio homogéneo de segundo grado en varias variables $x_1, \ldots, x_n$ es llamado una forma cuadrática. Consideremos solamete aquellos con coeficientes reales. Cualquier forma cuadrática puede ser escita como.

$$\emptyset(x_1, x_2, \ldots, x_n) = \sum_{i,k=1}^{n} a_{ik} x_i x_k ,$$

Donde $a_{ik} = a_{ki}$

Una forma cuadrática se dice que es positiva definida si sus valores son positivos para cualquier valor $x_1, \ldots x_n$, no todos son cero simultaneamente.

Es evidente que los coeficientes diagonales de una forma positiva definida son positivos, para

$$a_{11} = \emptyset(1, 0, \ldots 0) , \quad a_{22} = \emptyset(0, 1, \ldots 0), \ldots$$

$$a_{nn} = \emptyset(0, 0, \ldots, 1)$$

Denotado por X el vector con componentes $(x_1, \ldots x_n)$, podemos escribir una forma cuadrática como (AX , X), A siendo la matríz de los coeficientes de la forma. Esta matríz es simétrica por la definición. Los valores propios de la matríz son llamados los valores propios de la forma cuadrática. En vista de los resultados anteriores, todos los valores propios de una forma cuadrática son reales. Y, si una forma cuadrática es positiva definida, sus valores propios son positivos.

1.23.- Los Valores Propios y Vectores Propios de Matrices Semejantes.

Hemos establecido previamente, que las matrices semejantes tienen polinomios característicos identicos, y

consecuentemente identico espectro de valores propios. Por consiguiente los vectores propios de matrices semejantes son las columnas de las coordenadas de los vectores propios de la transformación bajo consideración, con respecto a bases diferentes, y así estan relacionados por la relación

$x' = C^{-1} x$, donde C es la matríz de transformación de coordenadas. Esta circunstancia puede ser verificada formalmente: Si $Ax = \gamma x$,

$$(C^{-1} A C) (C^{-1} x) = \gamma (C^{-1} x)$$

1.24.- Los Valores Propios de un Polinomio en una Matríz.

Sea A una matríz con valores propios $\gamma_1, \ldots, \gamma_n$ y sea

$\mu(x) = a_0 + a_1 x + \ldots + a_m x^m$ el polinomio dado.

Entonces, los valores propios de la matríz $\mu(A)$ serán : $\mu(\gamma_1), \mu(\gamma_2), \ldots, \mu(\gamma_n)$.

Esto está establecido para una matríz cuyos valores propios son distintos (todos). Realmente, tal matríz puede ser reducida a la forma diagonal por una transformación semejante.

$$A = C^{-1} \lceil \gamma_1, \gamma_2, \ldots, \gamma_n \rfloor C$$

De acuerdo a : $\mu(A) = C^{-1} \mu \lceil \gamma_1, \gamma_2, \ldots, \gamma_n \rfloor C$

Pero $\mu(\lceil \gamma_1, \ldots, \gamma_n \rfloor) = \lceil \mu(\gamma_1), \mu(\gamma_2), \ldots \mu(\gamma_n) \rfloor$,

lo cual se sigue del hecho de que,

$$\Gamma_{\gamma_1, \ldots, \gamma_n}^{k} = \Gamma_{\gamma_1^k, \ldots, \gamma_n^k}$$

Consecuentemente

$$\mu\left(\Gamma_{\gamma_1, \ldots \gamma_n}\right) = \sum_{k=0}^{m} a_k \Gamma_{\gamma_1^k, \ldots \gamma_n^k}$$

$$= \Gamma_{\sum_{k=0}^{m} a_k \gamma_1^k, \ldots \sum_{k=0}^{m} a_k \gamma_n^k}$$

$$= \Gamma_{\mu(\gamma_1), \mu(\gamma_2), \ldots \mu(\gamma_n)}.$$

Así la matriz $\mu(A)$ es similar a la matriz $\Gamma_{\mu(\gamma_1), \ldots, \mu(\gamma_n)}$,

y de acuerdo a sus valores propios son $\mu(\gamma_1), \mu(\gamma_2), \ldots, \mu(\gamma_n)$

q.e.d

Este resultado es verdadero para cualquier matriz.

Notese en particular que los valores de la matriz A^k son γ^k

1.25.- La Normalización de los Vectores Propios de una Matriz. Segundo Grupo de Relaciónes de Ortogonalidad.

Sea A una matriz real cuyos valores propios $\gamma_1, \ldots \gamma_n$ son distintos, y sean $X_1, \ldots, X_n$ los vectores propios correspondientes a ellos.

Como sabemos, la matríz transpuesta A', tiene los mismos vectores propios. Sean $X'_1, \ldots, X'_n$, los vectores propios de la matríz A', y su enumeración elegida tal que X_i y X'_i corresponden al conjugado complejo de valores propios, y por consiguiente, se sigue la relación de ortogonalidad :

$$(X'_i , K_j) = 0 \quad \text{para } i \neq j$$

Mostraremos ahora que habiendo elegido $X'_1, \ldots, X'_n$ de cualquier manera (están determinados pero, para un multiplicador numérico), podemos normalizar los vectores $X_1, \ldots, X_n$, tal que $(X'_i , X_i) = 1$

Demostración.
Sean $X_1, \ldots, X_n$ linealmente independientes.
resolviendo X'_i , en terminos de está base :

$$X'_i = \mu_1 X_1 + \mu_2 X_2 + \ldots + \mu_n X_n .$$

Formando el producto escalar (X'_i , X'_i), obtenemos

$$(X'_i, X'_i) = \mu_1 (X'_i, X_1) + \ldots + \mu_i (X'_i, X_i) + \ldots + \mu_n (X'_i , X_n)$$

$$= \mu_i (X'_i , X_i)$$

De aqui concluimos que :

$$(X'_i , X_i) = \alpha \neq 0, \text{ para } (X'_i , X'_i) > 0$$

Tomando en lugar de los vectores $X_1, \ldots, X_n$, los vectores

$1/\bar{\alpha}_1 X_1$, . . ., $1/\bar{\alpha}_n X_n$, llegamos a la normalización requerida

$$(X'_i , 1/\bar{\alpha}_i X_i) = 1/\alpha_i (X'_i , X_i) = 1$$

1.26.- La forma Canónica de Jordan.

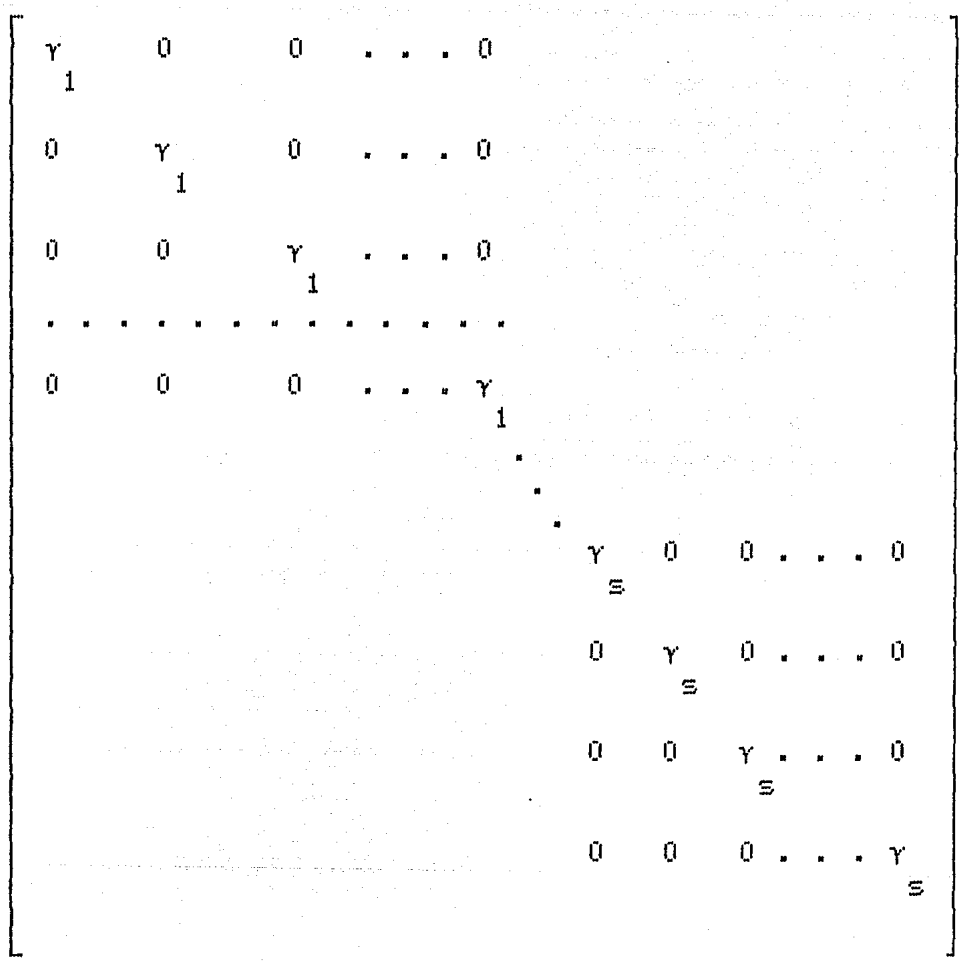
Hemos visto que si una matríz tiene n valores propios, puede trasladarse a su forma diagonal por transformaciones semejantes.

Dada la presencia de raices multiples, sin embargo, no siempre la transformación puede ser posible. Asi el problema de reducir una matríz a una forma tan simple como sea posible por medio de una transformación semejante hecha.

El problema es equivalente al encontrar una base con respecto a la cual la transformación lineal relacionada con la matríz dada tuviera una matríz de la forma más simple y esto último es la forma canónica de Jordan.

Omitiremos la demostración y nos concretaremos a la descripción de ésta forma canónica.

La forma (clásica) canónica es una matríz cuasi-diagonal compuesta de bloques canónicos.



Y es posible que el mismo valor γ_i aparezca en varios bloques canónicos. Todos los valores γ_i figuran en los diferentes del bloques, son los valores propios de la matríz canónica, y la multiplicidad, de un valor propio es igual a la suma de los ordenes de los bloques en el cual figura como elemento diagonal. Así la característica polinomial de la matríz canónica es igual al producto de los polinomios característicos de los bloques por separado, cada uno de ellos es igual a $(\gamma_i - \gamma)^m$ donde γ_i es el valor propio

es el orden de la i-ésima caja ó bloque.
   i

La determinación de los bloques de Jordan para una matríz dada A presenta ciertas dificultades. La característica
                                  mi
polinomial $\pi (\gamma - \gamma)$ coincide con la característica
                   i
polinomial de la matríz original, y consecuentemente es posible encontrarla sin conocer la matríz canónica en sí misma. Pero el conocimiento de la caraterística polinomial no hace posible la determinación completa de la forma canónica para un valor propio γ de multiplicidad
                                                  i
k estos pueden corresponder a varios bloques de Jordan
 i
conteniendo este valor como un elemento diagonal, y considerandolos solamente la suma de sus órdenes, no el orden de cada bloque en particualr. Sí la forma canónica debe ser determinada, el conocimiento de los " divisores elementales " de la matríz debe tenerse.
Designamos por $D_i(\gamma)$ el común divisor más grande de todos los menores del i-ésimo orden del determinante $| A - \gamma I |$.
En particular, $D_n(\gamma)$ coincide con la característica polinomial.
Puede ser probado que todo $D_i(\gamma)$, como $D_n(\gamma)$, son comunes a la clase de matrices similares, y puede demostrarse que $D_{i-1}(\gamma)$ divide a $D_i(\gamma)$.

Pero

$$\frac{D_i(\gamma)}{D_{i-1}(\gamma)} = E_i(\gamma);$$

Obviamente

$$D_n(\gamma) = \prod_{i=1}^{n} E_i(\gamma)$$

Se sigue que

$$E_n(\gamma) = \frac{D_n(\gamma)}{D_{n-1}(\gamma)}$$

es el polinomio mínimo de la matríz

Resolviendo $E_i(\gamma)$ en factores lineales, entonces

$$E_i(\gamma) = \prod_{j=1}^{s} (\gamma_j - \gamma)^{m_{ij}}$$

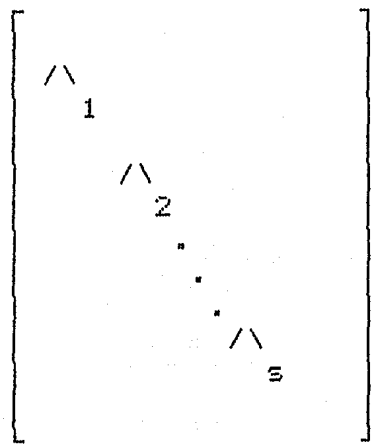
Aquí s denota el valor de valores propios distintos ,

$$\sum_{i=1}^{n} m_{ij} = k_j$$

Es obvio que entre los exponentes m_{ij} solamente algunos serán diferentes de cero.

Concideremos la matríz que transforma una matríz A dada en la forma canónica. Entonces las columnas de la matríz transformada serán las componentes de los vectores de esas bases en términos de lo cual la transformación en cuestión se describe como una matríz canónica.

Esta matriz canónica tiene la forma .



Donde

$$/\!\!\backslash_r = \begin{bmatrix} \gamma_r & 0 & \cdots & 0 \\ 0 & \gamma_r & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \gamma_r \end{bmatrix}$$

Y el orden de $/\!\!\backslash_r$ es igual a m ; sean

$$U_1^{(1)}, \ldots U_{m1}^{(1)}, \ldots U_1^{(s)}, \ldots U_{ms}^{(s)}$$ la base correspondiente.

Entonces las siguientes formulas para la transformación son válidas.

$$AU_1^{(r)} = \gamma_r U_1^{(r)} + U_2^{(r)}$$

$$AU_2^{(r)} = \gamma_r U_2^{(r)} + U_3^{(r)}$$

. . . . . .

$$AU_{mr-1}^{(r)} = \gamma_r U_{mr-1}^{(r)} + U_{mr}^{(r)}$$

$$AU_{mr}^{(r)} = \gamma_r U_{mr}^{(r)}$$

Para todo r = 1, . . . s .

Veamos que entre los vectores de la base canónica son los vectores propios de la matríz dada, uno por block. Puede ser probado que con todos estos vectores propios linealmente independientes de la matríz dada, son iguales al número de blocks canónicos en su forma canónica. En particular, el número de vectores propios linealmenete independientes corresponden a un valor propio que es igual al número de blocks canónicos que contienen este número. No es de mayor multiplicidad que el valor propio, y es igual a esta multiplicidad en caso, y solo en caso de que, todos los bloques contengan el valor propio dado y sean de orden 1, es decir, cuando los divisores elementales correspondientes sean lineales.

C A P I T U L O

II.

P R O B L E M A S D E C A L C U L O

D E L

A L G E B R A L I N E A L

PROBLEMAS DE CALCULO DE ALGEBRA LINEAL

Los problemas de cálculo más comunes del Algebra Lineal son los concernientes a las matrices de números reales.
Este capitulo nos llevará unos pasos más en la resolución de estos problemas y se tratan los principales problemas y los más antiguos del tema $Ax = b$ y $Ax = \gamma x$, sin embargo podemos decir que cada uno de ellos ha sido considerado por la nueva generación de matemáticos, y los dividimos en cinco grupos.

1).- Sea A una matríz de números reales con n renglones y n columnas, es decir una matríz cuadrada de n x n, y b un vector columna de n renglónes. El problema que frecuentemente se presenta en las ecuaciónes líneales es encontrar un vector columna " x " de n renglónes, tal que.

$$Ax = b$$

Por regla general se supone que A es una matríz no singular puesto que, solo si esto cumple entonces existe una solución unica para toda b. En caso contrario, es decir si A es una matríz singular, entonces se tendra una infinidad de soluciónes para b.

2).- Otro problema tradicional, si tenemos una matríz de números reales, con n renglones y n columnas es encontrar la matríz inversa, esto es A^{-1}, y la condición para que está exista es que su determinante sea distinto de cero.

3).- El tercer problema que se encuentra con frecuencia, es que si A es una matríz de n renglones y n columnas de números reales,la cual es simétrica, esto es que cumpla que sea cuadrada e igual a su transpuesta, es decir, si $A^T = A$, es simétrica.

Ejemplo:

$$\text{Sea } A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad \text{entonces, } A^T = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$T$$
y , $A = A$, por lo tanto A es una matríz simétrica.

$$T \qquad T$$
Dado que $A = A = (a) = (a)$.
$$ij \qquad ji$$

Aquí el problema consiste en encontrar alguno de todos (necesariamente real) los eigen valores de A. Recordando que un eigen valor de A es un número γ para el cual existe un vector columna v, tal que ;

$$Av = \gamma v$$

Así un vector (columna) v es llamado eigen vector de A correspondiente a γ, y con frecuencia el problema computacional encierra el encontrar un v correspondiente a cada eigen valor computado. Existen n eigen vectores ortonormales de A donde uno corresponde a cada eigen valor de A.

4).- Sea A una matríz no simétrica de n renglones y n columnas de números reales. Otro problema tradicional del Algebra Lineal, es de encontrar alguno de todos sus eigen valores y algunas veces tambien su eigen vector columna correspondiente. Recordando que un eigen vector renglón corresponde a γ y un vector renglón de n columnas v, es tal que ;

$$Av = \gamma v$$

Cuando A es no-simétrica el problema se complica de muchas maneras:

PRIMERO.- Algunos de los eigen valores de A son con frecuencia números complejos.

SEGUNDO.- Puede ser que no sea un eigen vector de n columnas linealmente independientes, y estos,no son por lo general ortogonales. Realmente es probable que se acerquen a la dependencia lineal, y lo mismo se tiene para los eigen vectores renglón.

TERCERO.- Si un eigen valor γ es una raíz de multiplicidad $K > 1$ de la ecuación caracteristica $\det (A - \gamma I) = 0$,

entonces puede existir en cualquier parte de 1 a k, de la columna linealmente independiente de los ingen -vectores correspondientes a γ. (Si A fuera simétrica, habria siempre un k) si el número es menor que k, corresponde a uno ó más blocks no diagonales en la forma canónica de Jordan de A, es decir.

$$\begin{bmatrix} \gamma_i & 0 & 0 & \cdots & 0 & 0 \\ 1 & \gamma_i & 0 & \cdots & 0 & 0 \\ 0 & 1 & \gamma_i & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & \gamma_i \end{bmatrix}$$

Donde en la diagonal principal se encuentra γ_i ; y directamente bajo la diagonal (en la subdiagonal) están dispuestos elementos los cuales son todos la unidad; todo el resto de los elementos son cero.

O equivalentemente los tan llamados divisores elementales no lineales de A.

CUARTO.- Múltiples o cercanamente múltiples eigen-valores de A son probables de ser funciones de cambio muy rapido de los elementos $a_{i,j}$ de A, de modo que los cálculos son de lo más complicado.

5).- Para cualquier vector columna y definimos la norma potencia p - ésima de y, como;

(1)
$$|| y ||_p = \left(\sum_{i=1}^{n} | y_i |^p \right)^{1/p}$$

Donde p es un número real con $1 \leq p, < \infty$, $y_1, y_2, \ldots y_n$

son las componentes de y, en un sistema coordenado dado. Definiremos la norma maxima como en el caso del limite de p cuando p ——> ∞ de (1).

$$\left|\left| y \right|\right|_{\infty} = \max_{1 \leq i \leq n} \left| y_i \right|$$

Las normas más usadas en análisis numérico son
p = 1, 2, . . . ∞, pero las estadísticas le están prestando atención a valores de p, entre 1 y 2.

Sea una matríz de n renglónes y k columnas de números reales y sea b el vector columna de n renglónes. Dado algún p, un problema computaciónal más reciente es el de encontrar un vector columna x de k renglónes tal que,

$$\left|\left| Ax - b \right|\right|_{p}$$

Sea minimizado cuando p =2, el caso común, este es el caso del problema de minimos cuadrados líneales, dado que tenemos una matríz de n renglones y k columnas, entonces podemos imaginar que el número de observaciones es mayor que el número de incógnitas, así que se debe expresar que el sistema Ax = b sea inconsistente. Probablemente no exista una selección de x que coincida perfectamente con los datos b; o, en otras palabras probablemente el vector b no sea una conbinación de las columnas de A.
Y el problema es selecciónar $\bar{x}$ que minimice el error, y de nuevo está minimización se hará en el sentido de los minimos cuadrados. El error es E = | Ax - b |, que es exactamente la distacia de b al punto Ax en el espacio columna de A, por lo tanto lo que hacemos al usar minimos cuadrados es localizar el punto $p = A\bar{x}$ que está más cerca de b que cualquier otro punto en el espacio columna.

Así la solución en términos de minimos cuadrados para un sistema inconsistente Ax = b de n ecuaciónes con k incógnitas satisface.

$$A^T A\bar{x} = A^T b .$$

Ahora bien para p = 2 la esfera unitaria es muy pareja y los metodos de trabajo de análisis buenos.

Sin embargo para p = 1 ó ∞ , la esfera unitaria tiene muchos ángulos y métodos para minimizar $||\ Ax - b\ ||_p$, haciéndose conbinatorio ó discreto.

6).- Para dos vectores columna de n renglónes x y y definimos x ≥ y, para hacer notar que $x_i \geq y_i$, para todas las componentes de x y y

Sea A y b, una matríz de n renglónes y q columnas de números reales y un vector columna de n renglónes respectivamente, entonces un problema de cálculo importante es el describir el conjunto " S " de vectores columna x de n renglónes tal que ;

$$Ax \geq b\ .$$

Algunas veces, como en los problemas de programación líneal uno parecido para vectores x en S tal que, $C^t x$ es un mínimo, donde C es un vector columna dado, con k renglónes.

Hasta este momento hemos hablado extensamente solo de matrices de números reales, problemas similares se presentan ocasionalmente para matrices de números complejos de la forma.

$$A = \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix}$$

Mucho de los problemas pueden ser formulados también para matrices en las cuales sus elementos son expresiónes incógnitas ó literales. Como método para simbolizar la manipulación en computadoras digitales, para hacerlo más accesible a los usuarios de las computadoras. Los problemas del Algebra Lineal con matrices literales serán posteriormente más estudiadas. La manipulación de simbolos prácticos probablemente será de más interes para los matemáticos en computación que para cualquier otro.

La presente discusión está limitada a las matrices de números ó mejor dicho a los problemas :

1).- Donde A es una matríz de n x n de números reales y b un vector columna de n x 1 y se desea encontrar x tal que.

$$Ax = b$$

2).- Donde A es una matríz de n x n de números reales, y se requiere.

$$A^{-1}$$

3).- Donde A es una matríz simétrica de n x n de números reales en la cual el problema es encontrar alguno de todos los eigen-valores de A, es decir.

$$Av = \gamma v$$

4).- Donde A es una matríz no simétrica de n x n de números reales, en la cual el problema es análogo al anterior.

$$Av = \gamma v$$

Porque los problemas líneales anteriores, surgen con frecuencia ?. Por que son importantes ?. La respuesta es que los operadores líneales son los más simples en matemáticas y solo los operadores líneales son plenamente estudiados, por consiguiente son un modelo natural para una matamática aplicada, para ser usada al atacar un problema.

Aun cuando los operadores líneales en espacios de dimensión infinita ocurrieran en análisis ó ecuaciónes diferenciales. (por ejemplo), la realidad del significado de la computación en el cual solo espacios de dimensión finita pueden ser manejados con computadoras digitales.

Modelos más realistas de matamáticas aplicadas son generalmente no líneales. Pero cuando algunos operadores no líneales son usados, la solución actual de ecuaciónes, funcionales siempre involucra la aproximación de operadores no líneales por algunos líneales. Un ejemplo típico de esto es el uso del método de Newton para resolver un sistema de ecuaciónes no líneales en el cual a todo paso un sistema de ecuaciónes localmente más apropiado debe ser resuelto. Los problemas no líneales generalmente son muy dificiles. Al atacarlos por métodos líneales, es esencial que nuestros conceptos líneales sean muy claros, de modo que ellos

puedan formalizarse para trabajar sin falla. Sólo de esta manera puede el análisis concentrarse en las dificultades reales del aspecto no lineal.

Este punto de vista no solo enfatiza la importancia de poder resolver problemas líneales, sino también la necesidad de resolver sistemas líneales con métodos extremadamente formales.

Del sistema de ecuaciónes líneales 1) es decir, Ax = b surgen dos opciones.

- Una de aproximación a ecuaciónes líneales funcionales generalmente ecuaciónes diferenciales parciales u ordinarias.

- La otra opción es un problema en el que dado un conjunto de datos, implica que el método de interpolación ó aproximación por familias de funciones líneales, tendran que ser aplicados.

El problema de los eigen-valores generalmente surge de estudios de vibración ó estabilidad ó resonancia de sistemas físicos líneales (aeronaves, reactore, etc) ó problemas de análisis de factor, valiendose del texto de Noble tomamos los siguientes ejemplos físicos de problemas del cálculo de matrices.

EJEMPLO No 1

Primero, retomando la ecuación diferencial para el movimiento armónico simple.
Si una partícula de masa m se mueve en línea recta y es atraida a un punto en la línea por una fuerza la cual es proporcional a la distancia X de la partícula desde el punto, entonces la ecuación de movimiento de la partícula está dada.

$$m \frac{d^2 x}{dt^2} + PX = 0 \qquad (a)$$

donde P es una constante positiva de proporcionalidad, para resolver está ecuación agrupamos, en la forma acostumbrada,

$$x'' = xe^{iwt}$$

Donde w es la frecuencia angular de vibracion y x es una constante que representa la amplitud de vibracion, la ecuación (a) da.

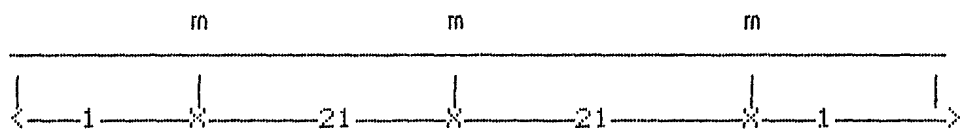
$$(- mw^2 + p) xe^{iwt} = 0$$

Dado que la partícula está en movimiento, x =/= 0. También e^{iwt} es distinto de cero para toda t. De aquí la ecuación anterior implica que.

$$mw^2 = p \quad ó \quad w = \sqrt{\frac{p}{m}}$$

Esto significa físicamente que hay solo una frecuencia de vibración libre.

Ahora veamos un ejemplo más complicado. Concideremos tres particulas, cada una de masa m, tomadas en las posiciónes 1, 31, 51, respectivamente, a lo largo de un alambre elástico de longitud 61, como lo muestra la figura a continuación.



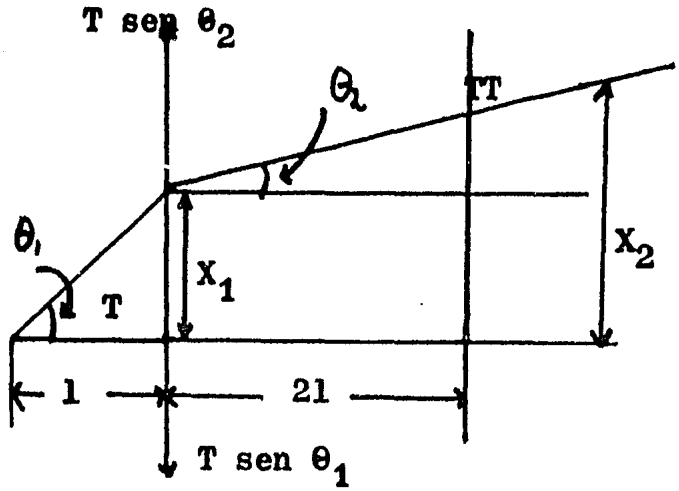
Tres camas en un alambre.

El alambre es fijo por ambos lados y está bajo una tensión T.

Supongamos que las partículas ejecutan pequeñas vibraciones trasnversales, bajo fuerzas no externas, los desplazamientos de estas tres partículas en una dirección perpendicular a la línea de equilibrio del alambre siendo.

X_1 , X_2 , X_3 , respectivamente.

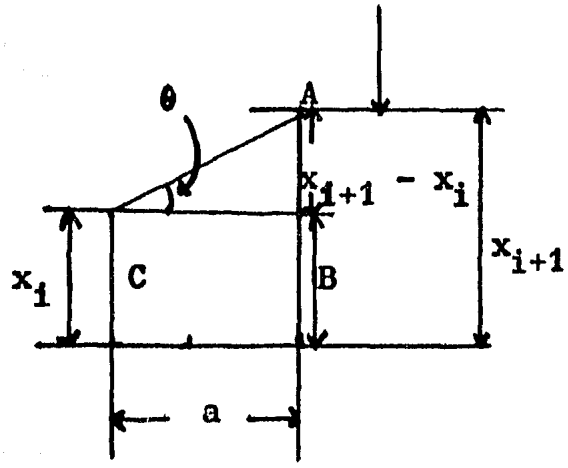
Las fuerzas en la cama 1 se muestra en la figura siguiente.



La fuerza resultante perpendicular a la línea de equilibrio es.

$$T (\text{sen } \theta_2 - \text{sen } \theta_1)$$

Tomando en cuenta que los desplazamientos son pequeños, tal que la tensión T puede llegar a ser constante, y el $\text{sen}\theta_1$ - $\text{sen}\theta_2$ pueden ser aproximados por $\tan\theta_1$, $\tan\theta_2$ como se muestra en la figura a continuación.



La aproximacion para $\text{sen}\theta$

$$\text{sen}\theta = \frac{AB}{AC} \approx \frac{AB}{BC} = \frac{X_{i+1} - X_1}{a}$$

Vibración de camas perpendicualar a un alambre.

Por la segunda ley de movimiento de Newton (los tiempos de aceleración de la masa es igual a la fuerza en la partícula) da para la cama 1, al usar los resultados anteriores.

$$m \frac{d^2 X_1}{dt^2} = - \frac{T X_1}{1} + \frac{T(X_2 - X_1)}{21}$$

Similarmente, para el mivimiento en las camas 2,3, obtenemos

$$m \frac{d^2 X_2}{dt^2} = - \frac{T(X_2 - X_1)}{21} + \frac{T(X_3 - X_2)}{21}$$

$$m \frac{d^2 X_3}{dt^2} = - \frac{T(X_3 - X_2)}{21} - \frac{TX_3}{1}$$

Ahora asumimos que todas las cantidades varian sinusoidalmente con el tiempo y con la misma frecuencia, y agrupando.

$$X_r = x_r e^{iwt} \qquad \text{donde } r = 1,2,3,$$

Entonces las ecuaciónes anteriores se convierten en :

$$(3-\gamma)x_1 - x_2$$

$$-x_1 + (2-\gamma)x_2 - x_3 \qquad (b)$$

$$-x_2 + (3-\gamma)x_3 = 0\ ,$$

Donde $\gamma = 2w^2 ml/T$; esto establece que γ debe ser eigen valor y $x = (x_1, x_2, x_3)^T$ su eigen vector asociado para la matriz

$$\begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{bmatrix}$$

Como vimos antes, las soluciónes distintas de cero x existiran precisamente cuando el coeficiente del determinante en la ecuación (b) sea cero, que es cuando,

$$\det \begin{bmatrix} 3-\gamma & -1 & 0 \\ -1 & 2-\gamma & -1 \\ 0 & -1 & 3-\gamma \end{bmatrix} = 0\ ,$$

Este determinante es fácil de evaluar, tal que de la expresión inmediatamente anterior se tiene que

$$-\gamma^3 + 8\gamma^2 - 19\gamma + 12 = 0,$$

cuyas raices son

$$\gamma = 1, \quad \gamma = 3, \quad \gamma = 4\ .$$

Para $\gamma = 1$ (b) se convierte en

$$2x_1 - x_2 = 0$$

$$-x_1 + x_2 - x_3 = 0$$

$$-x_2 + 2x_3 = 0$$

la cual se resuelve para obtener ,

$$\gamma = 1,\ x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad \text{para una } \alpha \text{ arbitraria } \neq 0$$

Para $\gamma = 3$ (b) es ,

$$\gamma = 3,\ x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \beta \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \quad \text{para una } \beta \text{ arbitraria } \neq 0$$

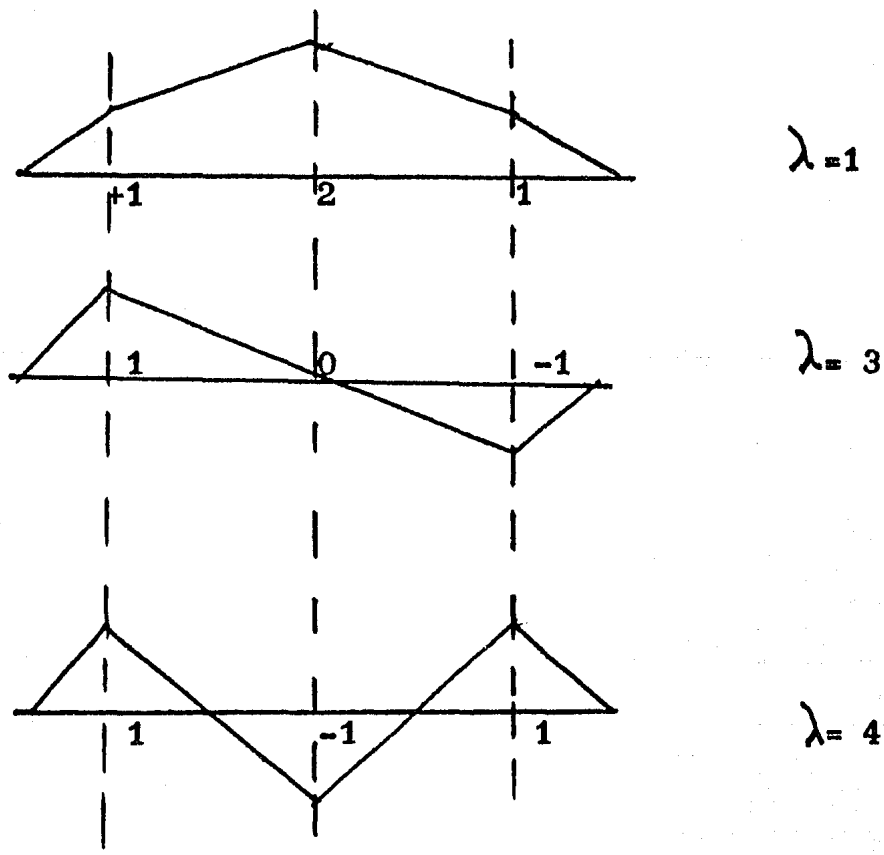
Para $\gamma = 4$ (b) es ,

$$\gamma = 4,\ x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \Gamma \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \quad \text{para una } \Gamma \text{ arbitraria } \neq 0$$

El significado físico de los resultados matemáticos anteriores en eigen sistemas es el siguiente. Correspondiendo a $\lambda = 1$, por ejemplo, hay una vibración libre con frecuencia angular w dado por $w = \sqrt{T/ml}$, y correspondiendo a esta frecuencia de vibración hay un modo de oscilación dado por $X_r = x_r e^{iwt}$ y (b), tal que las proporciones $x_1 : x_2 : x_3$, son 1 : 2 : 1.

Similarmente para $\lambda = 3$ y $\lambda = 4$.

Los modos de vibración estan ilustrados gráficamente en la figura siguiente. Estas tres frecuencias y modos de vibración son solo algunos que pueden existir.



Modos de vibración.

C A P I T U L O

III.

U N E N F O Q U E M A S S U T I L

D E L O S

P R O B L E M A S

UN ENFOQUE MAS SUTIL DE LOS PROBLEMAS

Dado que las computadoras actuales tienen capacidad de almacenaje finito y desde que tienen precisión finita, necesitamos estudiar más de cerca la naturaleza de las matrices A y sus problemas de cálculo.

Es A una matríz densa (tiene más elementos $a_{ij} \neq 0$, ó es una matríz hueca (tiene más elementos $a_{ij} = 0$) ?

$$A = \begin{bmatrix} 1 & 0 & -2 \\ 0 & 1 & 3 \\ 4 & 6 & 4 \end{bmatrix}$$

Matríz densa

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 6 & 0 & 0 \end{bmatrix}$$

Matríz hueca.

Si A es hueca, los elementos distintos de cero forman un paso significativo ? Por ejemplo es A triangular (con $a_{ij} = 0$ para $i > j$ ó para $i < j$) ?

$$A = \begin{bmatrix} c_{11} & 0 & \cdots & 0 \\ c_{21} & c_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}$$

Matríz triangular inferior

$$A = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ 0 & b_{22} & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & b_{nn} \end{bmatrix}$$

Matríz triangular superior

O es de la forma de Hessenberg. ($a_{ij} = 0$ para $i > j + 1$ ó $j = i + 1$) ?

Es una matríz banda ($a_{ij} = 0$ para $|i - j| > m$, donde $m << n$)?

$$A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 & 0 \\ a_{21} & a_{22} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & a_{n,n-1} & a_{nn} \end{bmatrix}$$

Es una matríz tridiagonal (es decir una matríz banda con $m = 1$) ?

$$A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & a_{nn} \end{bmatrix}$$

Todas estas formas ocurren con frecuencia y pueden ser dadas con consideración especial. Si A es simétrica (es decir $A = A^T$). Positivamente definida. Esto es que cada uno de los siguientes criterios es una condicion necesaria y suficiente para que una matríz simétrica real A sea positivamente definida :

I .- $x^T Ax = 0$ para todos los vectores x distintos de cero.

II.- Todos los valores propios de A satisfacen $\gamma_i > 0$

III.- Todas las submatrices A_k tienen determinantes positivos

IV.- Todos los pivotes (sin intercambio de filas) satisfacen $d_i > 0$.

Ahora bien si es hueca, es el paso asociado con la matríz adyacente de alguna gráfica. Frecuentemente las matrices asociadas con estructuras ó con ecuaciónes diferenciales parciales son mejor entendidas en términos de una gráfica asociada. Por ejemplo ;

En los elipsoides de n dimensiones, para entender la geometría de una matríz positivamente definida es $x^T Ax = 1$ Los vectores x que satisfacen esta ecuación se localizan en una superficie en el espacio n - dimensional. Y se quiere mostrar que es una elipsoide n - dimensional con centro en el origen.

$$\text{Sea } A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \quad \text{y} \quad x^T Ax = 2u^2 - 2uv + 2v^2 = 1$$

Reescribiendo esta ecuación como una suma de cuadrados con $\gamma_1 = 1$ y $\gamma_2 = 3$ como coeficientes. Se tiene ,

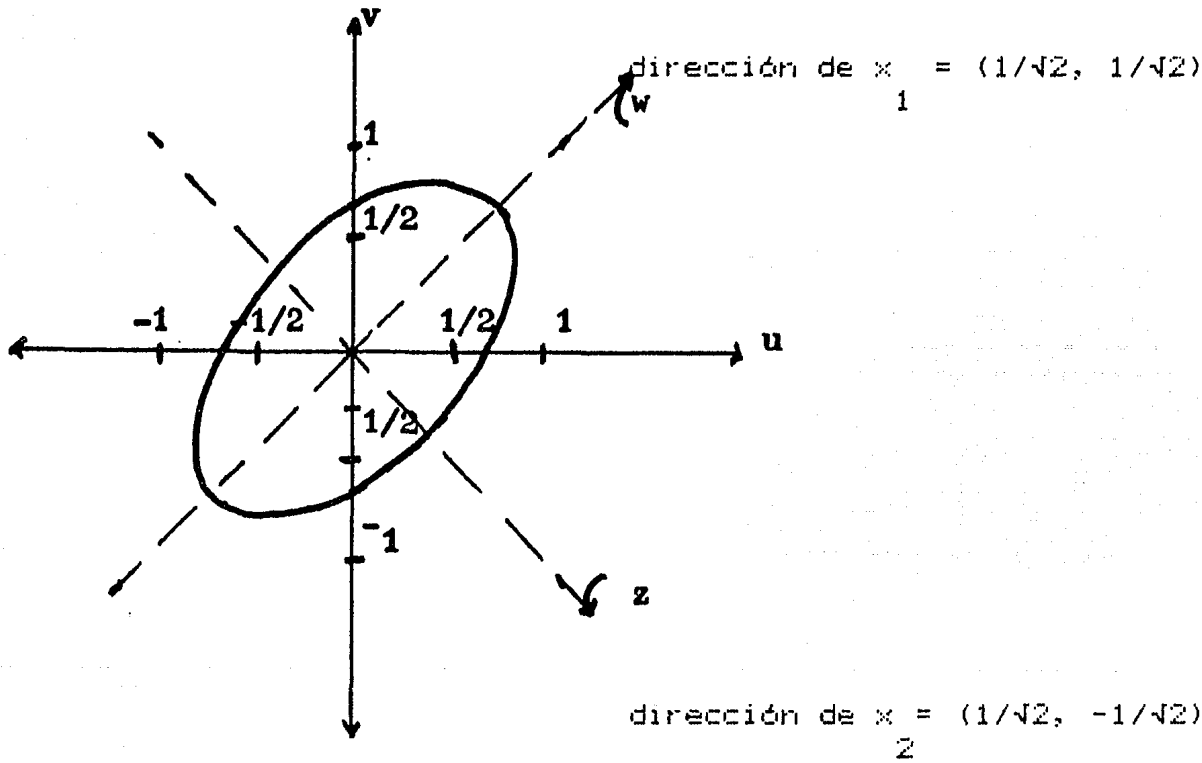
$$2u^2 - 2uv + 2v^2 = 1 \left(\frac{u}{\sqrt{2}} + \frac{v}{\sqrt{2}} \right)^2 + 3\left(\frac{u}{\sqrt{2}} - \frac{v}{\sqrt{2}} \right)^2 = 1.$$

$$\text{Haciendo} \quad \left(\frac{u}{\sqrt{2}} + \frac{6v}{\sqrt{2}} \right) = w \quad , \text{ y , } \quad \left(\frac{u}{\sqrt{2}} - \frac{v}{\sqrt{2}} \right) = z$$

Resulta la ecuación $w^2 + 3z^2 = 1$, la cual corresponde a una elipse. Cuyo eje más largo termina en el punto $w = 1$, $z = 0$ y el eje más corto termina en $w = 0$, $z = 1/\sqrt{3}$. Así la longitud del eje mayor es $2 \frac{1}{\sqrt{\gamma_1}}$, y la longitud del eje menor es $2 \frac{1}{\sqrt{\gamma_2}}$. Además, estos ejes están dirigidos hacia

los vectores propios $x_1 = (1/\sqrt{2},\ 1/\sqrt{2})$ y $x_2 = (1/2,\ -1/\sqrt{2})$, por lo tanto, la geometría esta completamente vinculada con los valores y vectores propios.

Gráfica de la elipse $2u^2 - 2uv + 2v^2 = 1$



Otro ejemplo ilustrativo es el de un Problema de Programación Lineal. En el problema del mínimo, se tiene que deseamos minimizar cx sujeto a, $Ax = b$ con $x = 0$ es decir.

En forma matricial podemos escribir.

$$\text{Min } Cx$$

s.a.

$$Ax \geq b$$

$$x \geq 0$$

En un ejemplo numérico sea

$$\text{Min } x + y$$

s.a.

$$x + 2y \geq 6$$

$$2x + y \geq 6$$

$$x \geq 0$$

$$y \geq 0$$

Matricialmente se puede escribir

$$\text{Min } C = (1, 1) \begin{bmatrix} x \\ y \end{bmatrix}$$

$$A = \begin{vmatrix} 1 & 2 \\ 2 & 1 \end{vmatrix} \quad \begin{vmatrix} x \\ y \end{vmatrix} = \begin{vmatrix} 6 \\ 6 \end{vmatrix} \quad \Big| \; > b$$

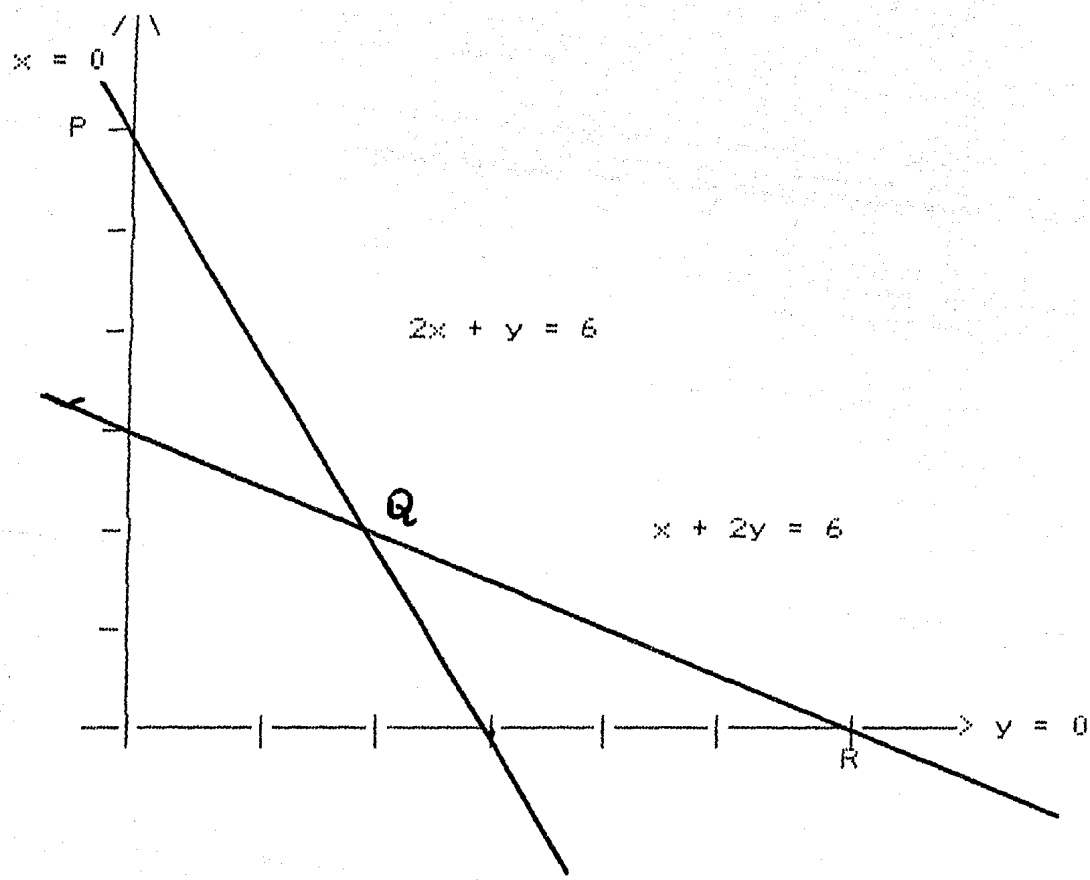
$$x \geq 0 \qquad y \geq 0$$

Interpretación geométrica, la primera condición $x \geq 0$ restringe al vector al cuadrante positivo, que en dos dimensiones es un cuarto del plano. Las otras m restricciones producen m semiespacios adicionales y los vectores factibles son todos aquellos que cumplen las m + n condiciones; están en el cuadrante $x \geq 0$ y al mismo tiempo satisfacen $Ax \geq b$. Dicho de otra manera, el conjunto factible es la intersección de m + n semiespacios; puede ser acotado, puede no estar acotado o puede estar vacío. Ahora bien, la función de costo Cx, introduce al problema una familia de planos paralelos, que dan los costos posibles, cuando varian, estos planos barren todo el espacio y el vector optimo x^* está en el punto donde tocan primero al conjunto factible.

En nuestro ejemplo numérico se tiene n=2 variables y m = 2 restricciónes; entonces hay seis intersecciones posibles, como se muestra en la gráfica, donde tres de ellas son realmente esquinas del conjunto factible y son los puntos

marcados con las letras A, B, C. respectivamente (0,6), (2,2) y (6,0) y uno de ellos debe ser el vector óptimo.

Gráfica.



Ahora, volviendo a lo anterior, están los elementos a_{ij}, almacenados en la memoria de la computadora, ó son generados de algún algoritmo conforme se necesite ? ó podrían definir el contenido informacional de una matríz como el número de espacios necesitados (en una cierta computadora) para almacenar datos y programar para obtener todos los a_{ij} ?

Cuál es el tamaño de la matríz A, relativo al tamaño de la memoria y rapidez de una computadora dada ? . Si nosotros estamos resolviendo un sistema de ecuaciónes lineales Ax=b tenemos muchas soluciones para b ó solo una ? Tenemos muchas matrices A diferentes que están juntas o sólo una A? Son los elementos de A precisamente números matemáticos (por ejemplo enteros) ó son números físicos (aproximaciones) ?

Los mejores propósitos de un sistema de ecuaciónes lineales no han considerado esta cuestión y miran al análisis numérico para explicar las posibilidades y seleccionar opciones.

Si para un propósito se necesita la matríz inversa A^{-1} es generalmente el descubrir por qué. Frecuentemente se desea un camino lo más práctico para resolver Ax = b, para un vector arbitrario b. Teniendo A^{-1} almacenada, el propósito que se espera es obtener la solución " x " en la forma $A^{-1}b$, para algún b nuevo que venga. Y será mostrado que hay otros métodos para obtener $A^{-1}b$, para nuevos vectores b, los métodos que no requieren de más almacenaje y no toman el camino más largo para la misma precisión, que la multiplicación de A^{-1} por b. Porque de estos factores, la computación de A^{-1} puede frecuentemente ser omitida. Sin embargo ciertas aplicaciónes estadísticas realmente requieren de conocimientos de un número de elementos diagonales de A^{-1}.

El problema de eigen valor, es decir al tener una matríz de n x n, A la cual es simétrica, el encontrar alguno de todos los eigen valores de A, $Av = \gamma v$, puede requerir de encontrar todos los eigen valores ó solo algunos, dependiendo de los requerimientos que se tengan, ya sea que los correspondientes eigen vectores sean necesitados ó no?

Si un conjunto completo de eigen-vectores es necesitado, es importante que sean mutuamente ortogonales?. Consiguiendo eigen-vectores ortogonales correspondientes a eigen-valores múltiples, es más largo y más dificil que encontrar eigen-valores?.

En el problema del valor característico, si las respuestas a las preguntas anteriores son afirmativas, entonces, para matrices no-simetricas A, uno tiene similares selecciones? Queremos todos los valores característicos ó solo algunos? queremos eigen-vectores columna ? Queremos eigen-vectores renglon ó ambos ?. Pero entonces viene una nueva elección. Si algunos eigen valores son multiplicados y corresponden a divisores elementales no lineales, que vectores es nuestro propósito encontrar ? En algunos textos de álgebra se aprende sobre cadenas de vectores principales que con el vector característico forman una base para el espacio nulo

$$k$$

N (Kernel) de $(A - \gamma I)$, donde γ es un eigen-vector de multiplicidad k, estos vectores principales están asociados con la forma canónica de Jordan de A. Un propósito que tenga una buena base en álgebra, querrá ver un conjunto de vectores principales (no son únicos). Pero estos vectores principales son extremadamente difíciles de evaluar, en parte por que son funciones discontinuas de los datos. Es como una base ortogonal para el espacio nulo N que sería un conjunto más usual de vectores.
El asunto parece ser poco entendido, para propósitos de problemas y análisis numérico.

Las matrices con eigen-valores multiples actualmente son muy raras, y una pequeña perturbación computacional de estos destruye la igualdad de eigen-valores.

Por consiguiente se da por hecho que no interese en la práctica que hacer con ellos. Pero en efecto el comportamiento de los divisores no lineales arrastrados en la práctica, para un largo y sorprendente conjunto de

matrices cercanas. Estas matrices cercanas tienen distintos eigen-valores, pero la columna k de eigen-vectores son tan cercanos a la dependencia lineal que no pueden ser separados en una computación normal.

Un aspecto ahora en el problema de mínimos cuadrados, se dice que es un análisis para x, para lograr minimizar.

$$f(x) = || \quad Ax - b \; ||_2$$

Y el propósito realmente es que se desea un mínimo para f(x) ó realmente es que se desea una x que de un valor de f(x) lo más aproximado al mínimo ?. En un problema de curva propia por ejemplo, uno puede muchas veces conseguir una cantidad sorprendentemente buena para un polinomio con coeficientes muy diferentes de los polinomios minimizados.

En todos los problemas de los cálculos anteriores es importante verificar cual de los siguientes tipos de respuesta al problema propuesto se está buscando :

a).- Una supuesta respuesta sin estimadores de su exactitud.

b).- Alguna respuesta, junto con algunas afirmaciones probabilísticas acerca de su exactitud.

c).- Alguna respuesta, junto con límites matemáticos probables para su error.

Normalmente es más largo para obtener, alguna respuesta junto con algunas afirmaciones probabilísticas acerca de su exactitud y, una supuesta respuesta sin estimadores de su exactitud, y hasta más largo de obtener alguna respuesta, junto con límites matemáticos probables para su error.

No es muy fácil ver cual de los tipos de respuesta anteriores al problema propuesto querríamos. Frecuentemente el desear una supuesta respuesta sin extimadores de su exactitud es completamente sastifactoria, los Científicos, Físicos e Ingenieros, frecuentemente hacen sus propios chequeos en cuanto a la validez de una respuesta y pueden no necesitar ninguna ni desear los métodos rigurosos de los

Matemáticos. Ellos pueden reconocer por ejemplo, que el modelo matemático es como una ligera aproximación a la realidad y que los métodos matemáticos serian solo utópicos. Cuando los matemáticos entran al mundo práctico de ingenieria, las reglas con las cuales los matemáticos están jugando frecuentemente tienen poca relevancia. El análista numérico frecuentemente tiene el problema de decidir cuando debe de jugar el juego de acuerdo a la reglas matemáticas y cuando jugar el juego de acuerdo a los ingenieros.

Es por supuesto, extremadamnete favorable para encontrar esos ejemplos ocasionales, donde los métodos matemáticos probables pueden ser encontrados, estos son justamente tan exactos y fáciles de encontrar como las supuestas respuestas.

Y nunca cesará uno de buscar tales espejismos, porque estos espejismos ocurren .

C A P I T U L O

I V.

E L E S T A D O D E L A C O M P U T A C I O N

D E S D E 1 9 4 6 H A S T A A H O R A

Y

L A N A T U R A L E Z A D E L H A R D W A R E

Y S O F T W A R E D E L A C O M P U T A D O R A

En las tres últimas decadas, se han logrado avances asombrosos en la construcción y el uso de los sistemas de computo. Cada avance hacía obsoletos tecnológicamente los modelos anteriores.

En el año de 1946 vió el nacimiento de la primera generación de computadoras con la computadora ENIAC (Electonic Numerical Integrador and Calculator) una de sus primeras utilizaciones fué para la construcción de tablas para el cálculo de trayectoria de proyectiles, es decir, este avance vino, como en muchas otras ciencias, por las necesidades militares que surgieron con la segunda guerra mundial.

Despues en 1951 surge la computadora UNIVERSAL I, y esto también marcó el inicio del desarrollo de los programas y de los lenguajes de programación. Muchas de las computadoras de la Primera Generación tenian que programarse en un lenguaje de máquina, que consistía en una serie de ceros y unos, lenguaje que era muy dificil de aprender y de utilizar. Como resultado de lo anterior, fueron desarrollados los lenguajes simbólicos y ensamblador. Estos nuevos lenguajes utilizaban símbolos como D para dividir y M para multiplicar, pero tenía que existir un programa que tradujera el lenguaje simbólico ó ensamblador al lenguaje máquina. En 1952 se desarrollo uno de los primeros traductores de programas y este hecho determinó el inicio del desarrollo en los lenguajes de programación.

Durante esta primera generación, los sistemas de computación se utilizarón principalmente con fines científicos.

Mientras que la primera generación de computadoras definitivamente significó una majoria sobre las máquinas antecesoras, aún dejaba mucho que desear la utilización de bulbos al vacío. Estos bulbos producían altas temperaturas y, en consecuencia, se quemaban con mucha facilidad y obligaba a utilizar costosos sistemas de refrigeración. Otra característica de las computadoras de esta generación era la escasa fiabilidad; por ejemplo, el tiempo medio entre averias de una unidad central era inferior a la hora, esto implicaba que para garantizar el buen funcionamiento de un equipo se necesitaba la total dedicación de un grupo

de personas encargadas del mantenimiento. La forma de ejecutar los trabajos era estrictamente secuencial: el programa que previamente se había perforado en tarjetas, se cargaba en la memoria de la computadora y, a continuación se ejecutaba, procesando las instrucciónes de cálculo y las de salidas de información. Estos problemas y el deseo de producir mejores sistemas de cómputo alentaron el desarrollo de transistores, lo cual marcó el inicio de la segunda generación de computadoras.

LA SEGUNDA GENERACION de computadoras utilizó transistores en lugar de bulbos. La razón de ello es que los transistores ofrecían mayores ventajas, tales como: Eran más pequeños, más rapidos, más confiables y producían menos calor durante su operación. Esta segunda generación de computadoras se ubica aproximadamente de 1959 a 1965.

Algunas majoras en los equipos de computación acompañaron a esta segunda generación de computadoras. Disminuyó el empleo de tarjetas como instrumento de almacenamiento y se introdujo el uso de dispositivos de cinta magnética y discos. Por primera vez, un sistema de cómputo tenía acceso directo a los datos almacenados de manera permanente por medio del disco.

Los dispositivos de entrada y salida fueron perfeccionados y también se desarrollaron mejores impresoras, terminales, lectoras de tarjetas, y además, se inició la utilización de núcleos de memoria (almacenamiento temporal). Los dispositivos de equipo también fueron modulares, una inovación que facilitaba la localización de problemas de equipo y mantenimiento.

También hubo algunas mejoras en los sistemas de programación. Los sistemas operativos ya estaban reemplazando muchas de las funciones que estaban siendo desempeñadas por la persona uqe operaba el sistema de cómputo. Se desarrollaron los compiladores para traducir lenguajes de programación de alto nivel a lenguaje de maquina y los programas de utilería se desarrollaron para intercalar, ordenar, transferir archivos de datos de una localidad de almacenamiento a otra.

Esta segunda generación fué testigo del cambio en el énfasis de equipo a programas. En muchas instalaciones, el costo de los programas ya se aproximaba al costo del equipo.

Más aún, algunas empresas empezaron a útilizar la computadora para suministrarle a sus ejecutivos la información que podía ayudarles a incrementar las utilidades y a mejor administrar las empresas. Esto fué el inicio de los sistemas computarizados de información administrativa.

Al mismo tiempo empezaron a surgir emopresas de procesamiento de datos. Agunas de éstas se fueron especializando en diversas funciones específicas dentro de la industria del procesamiento de datos, en términos de organizaciones avocadas a la programación ó bien a un servicio de procesamiento de datos. Al principio de la decada de los sesentas, muchas de las primeras organizaciones de procesaminto de datos se empezaron a ocupar en la capacidad del personal encargado del procesamiento de datos.

Aún cuando el uso de transistores mejoró considerablemente las capacidades de las computadoras de la segunda generación, científicos e ingenieros seguían trabajando con ahínco en areas de una nueva y más depurada tecnología. Estos esfuerzos llevaron al desarrollo de los circuitos miniaturizados y a la tercera generación de computadoras.

LA TERCERA GENERACION de computadoras se inició con la introducción de la IBM 360, lo cual aconteció en 1965. Los transistores dieron paso a los circuitos integrados y miniaturizados. De 1965 a la fecha, el desarrollo de las computadoras ha sido muy rápido y se han dado muchísimas mejoras en todos los aspectos del sistema de cómputo. Estas mejoras incluyen :

1).- Mayor velocidades de precesamiento.

2).- Mayor exactitud.

3).- La integración del equipo y los programas.

4).- La capacidad de desempeñar simultáneamente varias operaciónes.

5).- Avances en materia de comunicacion de datos.

6).- Una mejoría en la proporción entre desempeño y precio.

Los programas han mejorado significativamente, al grado de que, hoy en día, el mayor costo de un sistema de cómputo le corresponde a los programas. En un futuro cercano, los programas pudieran significar tanto como un 75% del costo total de un sistema de cómputo. Los lenguajes de programación de alto nivel se usan casi exclusivamente para el desarrollo de programas de aplicación, y ya existe un buen número de nuevos lenguajes de programación de alto nivel, aclarando que en este sentido, la tendencia es hacer más fácil la utilización de estos lenguajes. También existen nuevas formas de diseñar y desarrollar programas de aplicación, y un ejemplo de ello lo constituye el diseño estructurado. Así mismo, han existido unas mejorías realmente dramáticas en lo tocante a programas del sistema y lo mismo puede decirse de los sistemas operativos, los compiladores, ensambladores y los programas de utilería.

La tercera generación también a permitido hacercar la informática a los usuarios finales, tanto a los profesionales infomáticos como a los de otras especialidades, a travéz del teleprocesamiento, de los sistemas conversacionales y, sobre todo a travéz de las computadoras personales. El teleprocesamiento permite realizar la entrada de datos desde terminales remotas, y recibir los resultados en el mismo lugar. Los sistemas conversacionales permiten a los usuarios no sólo enviar y recibir datos desde sus terminales, sino también seguir e intervenir en el desarrollo de sus programas ha travéz de una " conversación " con el sistema. Por último, los computadores personales han conseguido popularizar la informática y es sorprendente comprobar cómo las prestaciones de una computadora personal, micro-computadora ó minicomputadora, son cada vez más parecidas a las ofrecidas por los grandes equipos tradicionales.

LA CUARTA GENERACION de computadoras ha surgido con los circuitos integrados de alta escala de integración. Algunos autores cuestionan su existencia por no haber sufrido

modificación el métpdo de explotación de los grandes equipos. Sin embargo, el importante auge de las mini-computadoras y microcomputadoras, ha supuesto importantes modificaciónes en el método de explotación, por lo que parece incuestionable que nos encontremos en la cuarta generación de computadoras. Algunos discrepan de esta opinión, dado que los procedimientos de explotación no han variado sustancialmente. En cualquier caso, lo que si se puede afirmar es que, en la decada de los noventas superaremos e inauguraremos la quinta generación de sistemas informáticos.

Por otra parte han continuado su crecimiento las organizaciónes de procesamiento de datos y estas mismas organizaciónes están ahora publicando revistas especializadas sobre computadoras y procesamiento de datos. A medida que se constituyen nuevas organizaciónes, son reconocidos los logros de sus predecesores.
Además es conveniente decir que la computación de matrices ha pasado por diversas etapas, así como la computación en general. Ciertamente no se puede aprender álgebra lineal teórica un lenguaje de programación algebraico ó algún otro, y enpezar a escribir programas los cuales se ejecutarán aceptablemente por modelos actuales. Ya se ha ganado mucha experiencia en los mejores algoritmos, y esto es mencionado en los textos matemáticos de álgebra lineal.

El monto de literatura en computación de matrices está en etapas. En 1958 Fadeev y Fadeeva en su libro registraron una reseña muy completa de métodos computaciónales. Wilkinson da más acerca del conocimiento sobre la computación de eigen-valores y de matrices densas almacenadas y escasas (ambas,simétricas y no simétricas), con límites de error para muchos algoritmos. Hay una diferencia muy pequeña entre estos libros, porque Wilkinson y unos pocos contemporaneos crearon más del material de su libro en los años posteriores a 1958. Posiblemente nadie podría enpezar la búsqueda en las matemáticas numéricas del álgebra lineal sin un completo conocimiento del material relevante en estos libros.

Lo más conocido para los principiantes, para matrices densas ó huecas, la eliminación gaussiana, de supuesto. Sabemos que algunas veces produce pobres resultados. No siempre estamos seguros por que, debatimos eternamente

sobre como opera el pivote para la eliminación, sin arreglarlo. El debate aún continúa, pero ahora principalmente entre personas quienes no entienden que las líneas principales de la respuesta han sido arregladas. Porque de no haber comprendido las dificultades de la eliminación gaussiana, buscamos otros métodos los cuales podrían hacerlo mejor. El método de gradiente conjugado ha sido ideado para matrices densas ó huecas por Lanczos.

Von Neumann y Goldstine, evitaron el problema del pivoteo para reducir cualquier sistema de ecuaciónes líneales Ax=b

$$T \qquad T$$

a un sistema positivo definido A Ax = A b. Sabemos que está normalización del problema fué costoso en tiempo y de emmpeorarse la condición del problema Von Neumann y Goldstine presentaron límites de error para la solución; actualmente los errores observados fuerón encontrados para ser acaso 100 veces más pequeños en casos razonables.

La forma de error de análisis fué una comparación directa de la máquina aritmética con operaciones exactas. La no-asociatividad y no distributividad de la máquina aritmética hizo el análisis extremadamente difícil. En algún caso, podría solo manejarse encerrado en la aritmética de punto fijo. Porque del tamaño de los límites de los errores, Von Neumann y Goldstine eran inecesariamente pesimistas sobre la posibilidad de invertir matrices en general de ordenes sobre 15. en máquinas con 27 bits de precisión.

Para los problemas de eigen valores, el estado de cosas estaba mucho peor, se tenía el método potencia con deflación de matríz. Mientras era razonablemente satisfactorio para unos cuantos renglones dominantes, su aplicación general requiere de intuición y suerte, y supone una algoritmización completa. Para matrices densas y matrices simétricas almacenadas teníamos el método de Jacobi y era muy satisfactorio.

Para matrices no simétricas, las cosas eran horribles. Si el método potencia no trabajara no tendríamos practicamente alternativas. Podríamos examinar para ceros del det(A - zI) de alguna manera u otra. Se han tratado métodos para determinar la característica polinomial, tan descrito en Fadeeva y encontrandolas irremediables. Fué también increible tanto como malos los métodos estandar

para n = 4, y ejecutarlos para n = 10. No obstante el método original de Lanczos necesitó cuidadoso manejo porque el nuevo resultó que con frecuencia era pobre.

El caracter de soluciones realizables a los problemas computacionales de álgebra líneal, está gratamente influenciada por la naturaleza de los sistemas computacionales aprovechables para nosotros. Es acostumbrado separar los sistemas computacionales en las siguientes líneas.

a).- HARDWARE.- Es la naturaleza de los circuitos electronicos de una computadora, con la palabra Hardware se hace referencia en informatica a las diferentes máquinas llamadas computadoras.

Analicemos las características y pecularíedades de las microcomputadoras (computadoras cuya unidad central de proceso - el cerebro ejecutor - es un microprocesador) de mayor vigencia y difusión. Para poder evaluar las diferencias y similitudes que existen entre los diversos sistemas, es preciso aplicar a todos ellos un cuestionario análogo que evidencie claramente la potencia y posibilidades de cada computadora. El avance de la informática y su definitiva penetración en la sociedad actual, se apoya sobre la base de los microprocesadores, la amplia variedad y diversificación de este tipo de sistemas hacen que su campo de aplicación sea casi ilimitado.

La entrada difinitiva de IBM en el mundo de la microinformática se ha realizado con la llegada del IBM-PC y el microprocesador 8088, cuya arquitectura interna es de 16 bits, un procesador aritmético destinado al cálculo rápido de operaciónes matemáticas en punto flotante, con este se consigue mayor versatilidad y rapidez.
El rango de memoria disponible en la IBM-PC ocila entre 64 y 544 kbytes.

b).- LENGUAJES DE COMPUTACION.- Los lenguajes en el cual son descritos los algoritmos para la solución de un problema dado en una computadora.

Con el fin de facilitar el trabajo del programador, surge la necesidad de que la computadora entienda un lenguaje diferente al suyo propio. Entramos en la etapa de la simbolización. Ya el programador no necesita conocer realmente donde ubica sus datos, le basta con referirse a direcciónes simbolicas. Así nacen los lenguajes de programación del tipo ensamblador y, consecuentemente, nace el soft-ware traductor ó conjunto de programas que permiten convertir los programas escritos en lenguaje del programador al lenguje que entiende la máquina.
Por esta vía se avanza más y se llega a un nuevo paso que permite al hombre dar al ordenador las fórmulas o notaciones que normalmente usa en su trabajo. Aparecen los lenguajes de programación de alto nivel y los compiladores programas cuya misión es traducirlos al lenguaje de la computadora.

c).- SOFT-WARE.- Son los programas los cuales hacen posible para una computadora actualmente ejecutar los algoritmos descritos en el lenguaje de la computadora.

La característica fundamental de las computadoras es la posibilidad de almacenar en su memoria programas que pueden modificarse y ejecutarce automáticamente. Pero esa ejecución automática implica que existan otros programas que permiten a la computadora coordinar el que varios programas puedan ejecutar se, bien secuencialmente. También es evidente que existe un gran número de operaciónes repetitivas, tales como clasificar un grupo de datos en un orden determinado, que podrían prepararse de forma que el usuario del ordenador no tenga que preocuparse de ello.
Esto es el soft-ware, también denominado " el componente lógico del sistema informático ".

En un sentido más estricto el soft-ware es el conjunto de programas que se utilizan en una computadora.
Es el conjunto de programas y otras ayudas generalmente proporcionados por el fabricante de la computadora que facilitan al usuario una operación más eficiente del equipo.

Tipos de soft-ware. Podemos clasificar los programas en cuatro grandes grupos.

Soft-ware especifico, ó programas de aplicación que pueden ser escritos tanto por el constructor, como por el usuario, todos los problemas técnicos y científicos pertenecen a este grupo.
Soft-ware traductor ó programas de ayuda para escribir nuevos porogramas. Constituido por los programas que permiten que los programas escritos por los usuarios en un lenguaje distinto al de la máquina se conviertan en programas con instrucciónes en código de lenguaje de máquina. Son escritos y proporcionados por los fabricantes.
Soft-ware funciónal ó programas de ayuda para ejecutar otros programas. Más comunmente conocido como sistema operativo. Es un conjunto de programas que facilitan una explotación más racional de las computadoras, guiando todas las tareas y ayudando a los programas en ciertas funciones. Son desarroyados por los constructores.
Soft-ware ó rutinas de utilidad. Programas que permiten la realización de funciones de uso frecuente y que generalmente son escritos por los fabricantes; aunque también pueden desarrollarlos los propios usuarios.
Al buscar en el Hardware de la computadora, para cálculos de álgebra líneal, se quiere saber que precisión es aprovechable para computación, cuantos dígitos están en el significante de los operadores de punto flotante, y en que base.
También se está interesado en el costo y rapidez de las operaciónes de doble precisión. En el trabajo de álgebra de matrices la operación crítica es frecuentemente la computación de la aproximación del redondeo de precisión simple al de doble precisión, el producto interno de dos vectores en los cuales sus componentes tienen números de punto flotante con precisión simple.

C A P I T U L O

V

P R O B L E M A S D E L A S
E C U A C I O N E S L I N E A L E S

Para resolver el problema de las ecuaciónes líneales de la forma Ax = b. La eliminación gaussiana es un algoritmo perfecto, excepto si un problema en particular tiene algunas propiedades especiales, como las tienen casi todos los problemas.

Para matrices grandes y huecas como aquellas que aparecen en aproximaciónes de diferencias finitas en ecuaciónes diferenciales parciales. Los métodos vistos dependen para su ocurrencia de sus exitos en la naturaleza de los problemas continuos siendo aproximados. Porque si las matrices son huecas, los métodos que prevalecén son los iterativos. Omitiremos la discución adicional de estas y daremos atención a las matrices densas almacenadas.

El análisis simple de multiplicaciónes anidadas (Capitulo XII) produjo grandes resultados; muestra que un detalle de análisi de error de redondeo de muchos posibles algoritmos es completamente inecesario. Son excluidos a-priori por simples concideraciónes de las limitaciones inherentes.

Para la solución de sistemas lineales la situación no es tan extrema; sin embargo un análisis de las limitaciónes inherentes es otra vez muy ilustrativa. Y ha precedido al intento de llevar a cabo un detallado análisis de error de la eliminación gaussiana muchas de las confusiónes que han sido evitadas.

Antes de esto examinemos las propiedades de la solución exacta del sistema :

$$Ax = b \qquad (1)$$

Para cualquier matríz consistente y un vector normado tenemos de.

$$Ax = b$$

que $$|| A || \; || x || \leq || b || \qquad (2)$$

La solución de

$$Ax = b$$

Está dada por : $x = A^{-1} b \qquad (3)$

Y dado que

$$|| x || \leq || A^{-1} || \, || b || \qquad (4)$$

muestra que

$$|| b || / || A || \leq x \leq || A^{-1} || / || b || \qquad (5)$$

Si A está definida por

$$|| A || = \text{mín. lím. sup. } || Ax || / || x || \qquad (6)$$

Entonces ambos límites son alcanzables en

$$|| b || / || A || \leq x \leq || A^{-1} || \quad || b ||$$

de aquí que el radio del valor más grande de x, al más pequeño correspondiente a un valor dado de || b ||, es $|| A || \, || A^{-1} ||$.

y es escribimos :

$$k(A) = || A || \, || A^{-1} || \qquad (7)$$

donde k (A) es llamada la condición de A con respecto a la inversion en la norma dada; la A en k (A) será de nuestro interés principalmente con las normas L1, L2 y L∞ y la norma Frobenius. De la relación $AA^{-1} = I$, tenemos,

$$k(A) = || A || \, || A^{-1} || \geq 1 \qquad (8)$$

Cuando k (A) es grande se dice que la matríz está mal condicionada con respecto a la inversión, por algunas razones, las cuales serán discutidas despues.

En el caso de la norma L2, entra en la variación de x para una b fija, está estipulada por el valor singular de descomposición.

$$A = U \, \text{diag}(\sigma_i) \, V^T \qquad (9)$$

De aquí U y V son ortogonales y la σ_i, la cual puede ser ordenada ta que.

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0 \qquad (10)$$

Son los valores singulares de A. La condición $\sigma_n \geq 0$ está basada en la suposición de que A es no singular. De las relaciónes.

$$|| A ||_2 = \sigma_1 \quad , \quad || A^{-1} ||_2 = 1 / \sigma_n \qquad (11)$$

Tenemos

$$k(A) = \sigma_1 / \sigma_n \qquad (12)$$

Y las matrices mal condicionadas se caracterizan por la propiedad $\sigma_1 >> \sigma_n$. Supongamos ahora que b es un rango por la derecha. Podemos escribir, entonces

$$b = || b ||_2 \sum \alpha_i u_i \text{ , donde } \sum \alpha_i^2 = 1 \qquad (13)$$

Ahora de la relación.

$$A = U \text{ diag } (\sigma_i) V^T$$

implica que :

$$AV = U \text{ diag } (\sigma_i), \text{ es decir, que } Av_i = \sigma_i u_i \qquad (14)$$

De aquí la solución de Ax = b, está dadas por :

$$x = || b ||_2 \sum (\alpha_i / \sigma_i) v_i \qquad (15)$$

Dando

$$|| x ||_2 / || b ||_2 = \left[\sum (\alpha_i / \sigma_i)^2 \right]^{1/2}$$

donde

$$\alpha_n / \sigma_n = \alpha_n || A^{-1} ||_2 \qquad (16)$$

A no ser que b pase a ser patologicamente deficiente en su componente α_n de u_n , la relación

$$|| x ||_2 / || b ||_2 = \left[\sum \alpha_i / \sigma_i)^2 \right]^{1/2} , \quad \alpha_n / \sigma_n = \alpha_n || A^{-1} ||_2$$

muestra que $|| x ||_2 / || b ||_2$ es del orden de magnitud de $|| A^{-1} ||_2$.

Dado que los u_i son ortogonales, cuando A está mal condicionada, la probabilidad de que un rango b de una x, la cual es del orden de magnitud del límite superior en la siguiente relación es grande, más bien que del límite inferior.

$$|| b || / || A || \leq x \leq || A^{-1} || || b ||$$

Si varias de las σ_i son comparables con σ_n , entonces

$|| x ||_2 / || b ||_2$ serán del orden de magnitud de $|| A^{-1} ||$, a menos que todos los α_i correspondientes sean pequeños y la probabilidad de que ésto suceda es totalmente mínima.

Si en el otro lado tomamos un rango x, y escribimos :

$$x = || x || \sum \beta_i v_i \quad \text{donde} \quad \sum \beta_i^2 = 1 \qquad (17)$$

entonces

$$b = Ax = \| x \| \sum \beta_i \sigma_i u_i$$

Dado que $\sigma_n << \sigma_1$, para una A mal condicionada la probabilidad es grande de que una b construida de está manera, sea particularmente débil en su componente u_n.

Realmente esto será verdad, a menos que una x pase a ser excepcionalmente deficiente en todos estos u_i, para los cuales $\sigma_i >> \sigma_n$. Este resultado es importante, porque se ha experimentado con frecuencia, y construye del lado derecho vectores de rango x, para tener sistemas de los cuales se saben las soluciónes. Tales lados derechos constituyen una inclinación muy simple.

Teoría de Pertubación para Sistemas Líneales.
El objetivo principal es que es casi inconcebible que pueda haber un algoritmo general propuesto para resolver Ax = b, el cual podría garantizar hacer mejor el trabajo que para dar la solución exacta de algún sistema perturbado.

$$(A + \delta A) x = b + \delta b$$

con

$$|\delta a_{ij}| \leq k^{1-t} \beta |a_{ij}|, \quad |\delta b_i| \leq k^{1-t} \beta |b_i|$$

Y en cuanto es probable para involucrar $0(n^3)$ operaciónes, los límites grandes para δA y δb son virtualmente inevitables.

Por lo tanto discutiremos aquí el efecto de las pertubaciones que satiface.

$$|\delta a_{ij}| \leq e|a_{ij}|, \quad |\delta b_i| \leq e|b_i|$$

Dado que tenemos límites para las perturbaciones en elementos individuales, no usaremos ninguno de estos; en efecto nuestro análisis involucra solo los requerimientos.

$|| \delta A || \leq \epsilon || A ||$, $|| \delta b || \leq \epsilon || b ||$, los cuales son consecuencia de .

$$| \delta a_{ij} | \leq \epsilon | a_{ij} | \quad , \quad | \delta b_{i} | \leq \epsilon | b_{i} |$$

Para las normas L1, L ∞ . Para la norma L2

$$\delta a_{ij} \leq k \beta^{1-t} | a_{ij} | , \quad | \delta b_{i} | \leq k \beta^{1-t} | b_{i} |$$

implica solo que :

$$|| \delta A ||_2 \leq n^{\frac{1}{2}} \epsilon || A ||,$$ pero ignoraremos esto por el momento. Observamos primero que podemos dejar la perturbación b. En efecto

$$(A + \delta A) x = b + \delta b$$

puede ser expresado de la forma.

$$(A + \delta A - \delta b x^T / x^T x) x = b$$

y tenemos.

$$|| \delta b x^T / x^T x ||_2 = || \delta b ||_2 / || x ||_2 \leq || b ||_2 / || x ||_2$$

y de $(A + \delta A) x = b + \delta b$

resulta

$$(|| A ||_2 + || \delta A ||_2) || x ||_2 \geq || b ||_2 - || \delta b ||_2$$

dando

$$|| A ||_2 (1 + \epsilon) || x ||_2 \geq || b ||_2 (1 - \epsilon)$$

de aquí

$$|| \delta b x^T / x^T X ||_2 \leq \epsilon || A ||_2 (1 + \epsilon)/(1 - \epsilon) = \epsilon || A ||_2$$

Si $\epsilon << 1$, y si estamos preparados para doblar el límite en δA, podemos distribuir con δb. Nótese que cuando A está mal condicionada.

$$|| A ||_2 (1 + \epsilon) || x ||_2 \geq || b ||_2 (1 - \epsilon)$$

será muy débil y para un rango b la probabilidad es grande de que.

$$|| b ||_2 / || x ||_2 << \epsilon || A ||_2 (1 + \epsilon)/ (1 - \epsilon)$$

y de aquí δb puede ser retomada por una perturbación adicional, y A es mucho más pequeña que $\epsilon || A ||$.

En cambio no es posible retomar la δA con el modesto argumento b, tenemos en efecto.

$$Ax = b + \delta b - \delta A x$$

Para un rango b vimos que $|| x ||$ es probable de ser del orden de magnitud de $|| A^{-1} || || b ||$ y dado que

$$|| \delta A x || \leq || \delta A || || x || = 0$$

$$(\epsilon || A || || A^{-1} || || b ||) = 0 \epsilon k || b ||$$

si ahora escribimos

$$(A + \delta A) y = b \quad , \quad Ax = b \quad , \quad || \delta A || \leq \epsilon || A ||$$

Tenemos

$$(I + A^{-1} \delta A) y = x$$

Observamos que, a menos que

$$|| A^{-1} \delta A || < 1$$

No podemos garantizar que A + δA no sea singular, tenemos

$$|| A^{-1} \delta A || \leq || A^{-1} || \, || \delta A || \leq \epsilon \, || A || \, || A^{-1} || = \epsilon k$$

En general no podemos esperar que y sea en algún sentido una aproximación de x a no ser que εk sea significativamente más pequeña que la unidad y esto se puedde asumir desde ahora.

De

$$(I + A^{-1} \delta A) y = x$$

tenemos

$$y - x = - A^{-1} \delta A y$$

$$|| y - x || \leq \epsilon k \, || y ||$$

y

$$|| x || \geq || y || \, (1 - \epsilon k)$$

Dado

$$|| y - x || / || x || \leq \epsilon k / (1 - \epsilon k)$$

La expresión de la izquierda es el error relativo en y y vemos que está efectivamente limitado por εk. Para casi todos los x y δA ésta es una estimación realista para el error relativo. La presencia del factor k en el límite de error para cualquier algoritmo, para resolver sistemas lineales es por consiguiente natural y realmente inevitable.

Cuando A es positiva definida se cumple que

$$\sigma_1 = \gamma_1$$

$$\sigma_n = \gamma_n$$

y

$$k = \gamma_1 / \gamma_n$$

Supongamos ahora que se intenta asignar un aproximado vía el tamaño del correspondiente residual. tenemos

$$r = b - A y = \delta A y$$

$$|| r || \leq || \delta A || \, || y || \leq \epsilon || A || \, || x || / (1 - \epsilon k)$$

este resultado es trivial y mucho más remarcable que el visto.
Consideremos una solución aproximada z teniendo un error del mismo orden de magnitud, como aquel en y pero arbitrario.

De $|| y - x || / || x || \leq \epsilon k / (1 - \epsilon k)$

podemos escribir

$z = x + w$ donde $|| w || \leq \epsilon k \, || x || / (1 - \epsilon k)$

y de aquí

$$b - Az = b - Ax - Aw = - Aw$$

Dando

$$|| b - Ax || \leq || A || \, || w || \leq \epsilon k \, ||A|| \, ||x|| /(1 - \epsilon k)$$

Este límite es más grande por el factor k que el que se tiene en y. De este la totalidad de las soluciónes, el tener errores del orden de magnitud $\epsilon k \, || x || / (1 - \epsilon k)$ casi todos ellos dan muy grandes residuales que los otros, las cuales son soluciónes exactas correspondientes a una matríz A + δA con una perturbación relativa de orden ε. Hay mucho más aspectos remarcables de esta propiedad de y. Consideremos una solución aproximada x + w con

$$|| w || \leq \epsilon || x ||$$

Es decir, las cuales realmente tienen un error relativo bajo ahora tenemos

$$|| b - A (x + w) || =$$

$$= || - Aw || \leq || A || \, || w || \leq \epsilon || A || \, || x ||$$

ignorando el factor 1/ (1 - ek), el límite en el residual para una solución muy precisa no es más pequeña que para y. Un ejemplo particular de como una x + w es estipúlada por la correcta versión de redondeo de x, en el caso cuando los elementos de x no son digitales. Para esto tenemos.

$$|| w || \leq k \beta^{1-t} x$$

Veamos que en toda esta cantidad de virtudes entre soluciones aproximadas, no cabe esperar que de un residual más pequeño que la solución exacta de y.

$$(A + \delta A) y = b \quad \text{con} \quad || \delta A || \leq k \beta^{1-t} || A ||$$

Nos lleva al hecho de que

$$y = x - A^{-1} \delta Ax + A^{-1} \delta AA^{-1} \delta Ax$$

y de aquí

$$A y = A x - AA^{-1} \delta Ax + \text{terminos más chicos}$$

$$= b - \delta Ax + ...$$

El principal error en y es $A^{-1} \delta Ax$, y cuando esto se multiplica por A^{-1}, la A y A^{-1} se cancelan si : A es sustituida por cualquier otra matríz B con la misma norma, entonces podemos decir solo que.

$$|| A B || \leq || A || || B || = || A || || A^{-1} || = k (A)$$

De aquí el factor extra k.

Notese que a no ser que la solución exacta tenga elementos digitales entonces, no habrá en general alguna solución digital aproximada, dando un residual más pequeño que $k \beta^{1-t} || A || || x ||$ el cual corresponda a un residual relativo.

$$k\,\beta^{1-t}\,||\,A\,||\,||\,x\,||\,/\,||\,b\,||\,\leq\,k\,\beta^{-t}\,||\,A\,||\,||\,A^{-1}\,||$$

y para una b aleatoria hemos visto que la x correspondiente es tal que este es un límite realista.

Hemos propuesto consideraciónes detalladas de una perturbación en b porque en general esto es menos importante.

Supongamos ahora que

$$A\,y = b + \delta b \qquad ||\,\delta b\,|| \leq \epsilon||\,b\,||$$

entonces

$$y = A^{-1}\,b + A^{-1}\,\delta b = x + A^{-1}\,\delta b$$

Dando

$$||\,y - x\,|| \leq \epsilon||\,A^{-1}\,||\,||\,b\,|| \leq \epsilon||\,A^{-1}\,||\,||\,A\,||\,||\,x\,||$$

ó

$$||\,y - x\,|| / ||\,x\,|| \leq \epsilon k$$

Aparte del factor 1 / (1 - ϵk), este es el mismo límite para el error relativo en y, como está dado en

$$||\,y - x\,|| / ||\,x\,|| \leq \epsilon\,k / (\,1 - \epsilon k\,)$$

Para la perturbación δ A en A. No obstante éste está perdido en

$$||y - x||\,\leq \epsilon\,||\,A^{-1}\,||\,||\,b\,||\,\leq \epsilon\,||\,A^{-1}\,||\,||\,A\,||\,||\,x\,||$$

b es sustituido por || A || || x || en la última desigualdad y para la mayoría de b ésta en una desigualdad extremadamente débil.
Para la mayoría de b hemos visto que ,

$$||\,x\,|| = 0\,||\,A^{-1}\,||\,||\,b\,||$$

y cuando esto es verdad tenemos

$$|| y - x || \leq \epsilon || A^{-1} || \, || b || = 0 \epsilon || x ||$$

Tal que para la mayoría de b, cualquier perturbación δb tal que $|| \delta b || \leq \epsilon || b ||$ da solo un error relativo pequeño en la y correspondiente.
Quizá en este punto debemos enfatizar dos factores.

1).- Hay siempre una aproximación digital a x teniendo un error relativo bajo. En efecto si x_R es el vector derivado de x por redondear cada elemento a la aproximción dígital correspondiente será.

$$|| x_R - x || / || x || \leq \frac{1}{2} \beta^{1-t}$$

2).- Cuando A está mal condicionada entonces, para la mayoría de b no habrá aproximación digital para x dando un residual relativo bajo. Hemos visto que toda x_R (la cual puede ser considerada como la mejor aproximación digital) dará un residual relativo del orden de magnitud de $k \beta^{1-t}$ para la mayoría de b. En general es solo cuando b es tal que $x = 0$ ($|| b || / || A ||$) y esto hará aproximaciones digitales para x dando un residual relativo bajo.

Solución Práctica de Sistemas de Ecuaciónes Lineales

Aquí discutiremos brevemente algunas concideraciónes computacionales prácticas, concernientes a la eliminación gaussiana.
Muchos problemas en las ciencias aplicadas involucran la inversión de matrices de m x n ó la solución de n ecuaciónes con n incógnitas. Si n es grande (más grande que 3 o 4) ó las entradas en la matríz no son enteros con solo pocos digitos la solución de tales problemas a mano son tediosos e ingratos, en la práctica estos problemas son resueltos por lo general por programas de cómputo, en computadoras digitales electrónicas, pero con la guía

humana para simplificar el problema del comienzo. El procedimiento sistemático da una excelente base para determinar la naturaleza del conjunto de soluciones y para calcular las soluciones, podemos hacer de este procedimiento las bases para el método computacional elegido, si se toman las precauciones propias. Antes debemos considerar como hacerlo, haciendo algunos comentarios acerca de la naturaleza del sistema de ecuaciónes, el cual surge en la práctica, tal que el lector pueda ver el criterio usado para medir las eventualidades de cualquier procedimiento propuesto.

Consideremos el sistema de ecuaciónes líneales.

1).- El problema real que queremos resolver ha sido modelado de alguna manera matemática.

Supongamos que un paso en la solución del modelo matemático requiere la solución del conjunto de ecuaciónes

$$A^* x^* = b^*$$

Donde A^* y b^*, son matrices, cuyos elementos estan precisamente definidos, pero generalmente no necesariamente conocidos.

2).- Dado el sistema ideal $A^* x^* = b^*$, tenemos actualmente a la mano un conjunto de ecuaciónes

$$A x = b$$

Donde tenemos que $A \neq A^*$ y $b \neq b^*$. Los errores en los datos son a menudo tambien al menos en parte debido a errores de redondeo y para una conversión binaria cuando los números son almacenados en la computadora; comunmente los datos tambien llevan un error experimental causado por nuestra poca habilidad para medir A^* , b^* precisamente.

El sistema

$$A x = b$$

por supuesto tiene una solución exacta, que es, una para la cual $A x - b$ es exactamente igual a cero, con tal de que las computadoras se involucren en evaluar $A x - b$ exactamente, pero lo mejor que podemos esperar hacer en una computadora digital ó en una calculadora de escritorio ó

probablemente a mano, es obtener una solución aproximada x', no igual a la solución exacta x generalmente, para

$$A\,x = b$$

Nuestros métodos prácticos para resolver las ecuaciónes podria ser, por supuesto, solamente con el dato aproximado A,b en la ecuación Ax = b. Cuando tratamos de evaluar lo que hacen nuestros métodos, esto es, como se produce la precisión y aproximación x', podemos solamente medir la precisión con respecto a la solución exacta x " para los datos que nos dan en Ax = b ", trataremos de garantizar que nuestros procedimientos hacen el error x - x' " pequeño ".

Los científicos aplicados, por supuesto que solo quieren la respuesta x', y dado que los errores x - x' no nos conciernen de momento, entramos al procedimiento de eliminación gaussiana para obtener alguna idea sobre el monto del trabajo involucrado en la eliminación gaussiana, ambos para resolver ecuaciones y para invertir matrices, y enseguida lo concerniente a ejemplos concretos con un tratamiento más teórico de la solución de sistemas líneales de la forma A x = b.

Por simplicidad consideremos las adiciones y las sustracciones juntas como " adiciones "
multiplicaciones y divisiones son concideradas juntas como multiplicaciones.
Concideremos la solución de k conjuntos de ecuaciones como en

$$A\,x = b$$

entonces

$$A\,x_j = b_j \qquad \text{con } j = 1,2\ldots, k$$

El cálculo de los elementos de

$$A^{(r+1)}, \quad b^{(r+1)}, \quad \text{de } A^{r}, \quad b^{(r)}$$

involucran lo siguiente,

1).- Cálculo de m_{ir}, es decir.

$$m_{ir} = \frac{a_{ir}^{(r)}}{a_{rr}^{(r)}}, \qquad \text{con } i = r+1,\ldots, n$$

involucrando n-r dividiones.

2) El cálculo de $a_{ij}^{(r+1)}$, de

$$a_{ij}^{(r+1)} = a_{ij}^{(r)} - m_{ir}\, a_{rj}^{(r)} \qquad \text{con } i,j = r+1,\ldots, n$$

involucrando $(n-r)^2$ multiplicaciones y sumas.

3) El cálculo de k conjuntos $b_i^{(r+1)}$ involucrando k (n-r) multiplicaciones y adiciones.

$$b_i^{(r+1)} - b_i^{(r)} - m_{ir}\, b_r^{(r)} \qquad \text{con } i=r+1,\ldots, n$$

donde para r = 1, debemos tomar

$$a_{ij}^{(1)} = a_{ij}$$

$$b_i^{(1)} = b_i$$

La eliminación progresiva para obtener.

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1^{(1)}$$

$$a_{22}x_2 + \ldots + a_{2n}x_n = b_n^{(2)} \qquad (1)$$

$$+ a_{nn}x_n = b_n^{(n)}$$

Por consiguiente se requiere de

$$\sum_{r=1}^{n-1} \{(n-1)^2 + (k+1)(n-r)\} = n\{(1/3\, n^2 - 1/3) + 1/2\, k(n-1)\}$$

multiplicaciónes

$$\sum_{r=1}^{n-1} \{(n-1) + k(n-r)\} = n\{(1/3\, n^2 - 1/2\, n + 1/6) + 1/2 k(n-1)\}$$

sumas ó adiciones

donde hemos usado los resultados

$$\sum_{s=1}^{m} s = 1/2\, m(m+1)$$

$$\sum_{s=1}^{m} s = 1/6\, m(m+1)(2m+1)$$

En la solución de (1) por sustitución hacia atrás de x, involucra n-r multiplicaciones y adiciones y una división. Dejando la división como equivalente a una multiplicación, las k sustituciones hacia atrás involucran.

$k(1 + 2 + \ldots + n) = 1/2\, k\, n\, (n+1)$ multiplicaciones

y

$k(1 + 2 + \ldots + n-1) = 1/2\, k\, n\, (n-1)$ adiciones.

Los totales para la solución completa de k conjuntos de ecuaciones son :

$$n\left(1/3\, n^{2} - 1/3 + k\, n \right) \text{ multiplicaciones}$$

y

$$n\left((1/3\, n^{2} - 1/2\, n + 1/6 + k\, (n - 1)\right) \text{ adiciones}$$

Cuando invertimos A, empezamos con n vectores unitarios e_j en vez de b_j a la derecha de

$$A\, x_j = b_j$$

Al recordar con cuidado donde hay ceros y unos a la derecha y entonces, evitando multiplicaciones por unos y ceros y sumas de ceros, podemos facilmente reducir el trabajo total para la eliminación progresiva a,

$$1/2\, n^{2}\, (\, n - 1\,) \text{ multiplicacion}$$

y

$$1/2\, n\, (\, n - 1\,)^{2} \text{ adiciones}$$

Y también siendo cautelosos en la sustitución hacia atrás, finalmente llegamos a los requerimientos totales para la inversión entera.

$$n^{3} \text{ multiplicaciones}$$

y

$$n^{3} - 2n^{2} + n \text{ adiciones}$$

Las conclusiones importantes de estos resultados para la eliminación gaussiana son .

1).- El número de multiplicaciones y adiciones requeridas para resolver un conjunto simple de ecuaciones son del orden de $1/3\, n^{3}$

2).- El número de multiplicaciones y adiciones requeridas
3
para invertir una matríz son del orden de n

Esto no dice que :

3
1).- El monto de trabajo es proporcional a n tal que si duplicamos el número de ecuaciones tenemos que hacer 8 veces el monto del trabajo involucrado, y rapidamente se convierte en sustancial, a medida que el sistema se incrementa.

2).- Alrededor de 3 veces del monto de trabajo es involucrado cuando invertimos una matríz, comparado con resolver un conjunto de ecuaciones, no n veces, como podriamos esperar. El radio es esencialmente independiente de la medida del sistema.
Para una n muy grande estas operaciones contadas, son
3
consideradas un tanto impropias, dado que n será extremadamente grande. Sin embargo, se tiende a olvidar que
3
multiplicar dos matrices de orden n juntas requieren de n multiplicaciones y adiciones. De aquí el monto de trabajo
2 T
requerido para formar A ó A A. Además se sabe que en ningún método las operaciones son enteramente por renglón y columna,y puede requerir menos que este monto de trabajo.

En suma, justamente olvidamos que las computadoras pueden operar rapidamente, para implementar la eliminacion gauissiana. Pues a mano es laborioso resolver 5 ecuaciones con 5 incognitas. En el presente las computadoras es lo más práctico para resolver sistemas de 100 ecuaciones con 100
4
incógnitas, con todos los 10 coeficientes distintos de cero. Las operaciones aritméticas toman más o menos 10 segundos en hacerlo y el costo es menor que un dolar. Este es el porque las computadoras han revolucionado el camino en el cual formulamos los problemas en las ciencias aplicadas. Sin embargo todavía titubearíamos antes de tomar un sistema con n = 100, donde la mayoria de los coeficientes son cero.

El Método de Gauss

En las páginas anteriores se ilustraron algunas maneras en las cuales surgen las matrices en las ciencias aplicadas. En lo que sigue veremos las matrices desde un punto de vista más teórico, para obtener un entendimiento más amplio de las potencialidades y las limitaciones de los métodos matriciales.

Como se ha ilustrado en los ejemplos considerados, la utilidad de las matrices en las ciencias aplicadas, está relacionada a menudo con el hecho de que se lleve a cabo un método conveniente para formular problemas físicos en términos de un conjunto de ecuaciones lineales algebraicas. Sin embargo es importante entender teóricamente las diversas situaciónes que pueden surgir cuando se resuelven conjuntos de ecuaciónes lineales. Este es el objetivo de lo siguiente.

Pasando a la eliminación gaussiana. El punto inicial es el sistema.

$$A\,x = b$$

comenzaremos con un ejemplo en tres dimensiones, con el sistema.

$$2x + y + z = 1$$

$$4x + y = -2$$

$$-2x + 2y + z = 7$$

Nuestro problema es encontrar el valor de las incógnitas x,y,z, y aplicando la eliminación gaussiana tenemos.

a).- Sustrayendo de la segunda ecuación, la primera multiplicada por 2;

b).- Sustrayendo de la tercera ecuación, la primera multiplicada por -1.

el resultado es un sistema equivalente de ecuaciones.

$$2x + y + z = 1$$

$$-y - 2z = -4$$

$$3y + 2z = 8$$

c).- Al sustraer de la tercera ecuación la segunda multiplicada por -3 nos queda.

De tal forma que hemos obtenido de la ecuación original.

$$Ax = \begin{bmatrix} 2 & 1 & 1 \\ 4 & 1 & 0 \\ -2 & 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \\ 7 \end{bmatrix}$$

Un sistema equivalente pero más simple, con una matríz de coeficientes nueva que denotaremos por U :

$$Ux = \begin{bmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ -4 \\ -4 \end{bmatrix}$$

Ahora la matríz de coeficientes es triangular superior : son cero todas las entradas debajo de la diagonal principal.
El lado derecho, que es un nuevo vector c, se derivó del vector original b mediante los mismos pasos que cambiaron a A en U. de modo que la eliminación gaussiana consiste en;
Comenzar con A y b;
Aplicar los pasos a),b) y c), en ese orden;
Terminar con U y c.
La última etapa de la solución Ux = c, mediante la sustitución hacia atrás.
La matríz A puede escribirse, mientras que ninguno de los

pivotes sea cero, como un producto LU de una matríz triangular inferior L y una matríz triangular superior U, las entradas de la diagonal principal de L son unos; debajo de la diagonal están los multiplicadores l_{ij} de la fila j que fueron sustraidos de la fila i en el proceso de eliminación. U es la matríz de coeficientes que obtenemos después de la eliminación y antes de la sustitución

regresiva; las entradas de su diagonal son los pivotes.

La Existencia de Soluciónes para un conjunto de ecuaciones. Algunos ejemplos.

Hemos visto hasta aquí los métodos para encontrar la solución numerica de ecuaciones líneales simultaneas, pero no hemos puesto mucha atención a la cuestión de que tales soluciones existan, ó que la solución sea única. Este es el principal problema que ahora trataremos.

Consideremos ahora la ecuación simple ax = b, donde a,x,b, son escalares. Y decimos inmediatamente que la solución de esta ecuacion es x = b / a, pero en esto hay tres posibilidades.

1).- Si a =/= 0, entonces x = b/a, y ésta es la única solución de la ecuación, para cualquier valor de b. (El cual puede ser cero, en tal caso la solución es x = 0)

2).- Si a = 0, hay dos posibilidades, dependiendo del valor de b.

a).- si b =/= 0, entonces la ecuación es 0 x = b =/= 0 y no existe una solución finita. La solución " x = ∞ ", no puede ser considerada como una solución posible. Decimos que " la solución no existe ", ó, alternativamente, que " la ecuación es inconsistente ", implica que 0 = b =/= 0, lo cual es una contradicción.

b).- si b = 0, entonces cualquier número es una solución de la ecuación, para 0 x = 0, siempre que el valor sea dado a x. Se excluye otra solución infinita, dado que 0 ∞ no está definido. Es un notable resultado que precisamente las mismas posibilidades existen en el caso de dos ecuaciones con dos incónitas. Así,

$$x_1 + x_2 = 2$$

$$x_1 - x_2 = 0$$

tienen una solución única.

$$x_1 + x_2 = 2$$

$$x_1 + x_2 = 1$$

son inconsistentes, y

$$x_1 + x_2 = 2$$

$$2x_1 + 2x_2 = 4$$

tienen una infinidad de soluciónes,

sean $x_1 = k$, $x_2 = 2 - k$ para todo k .

Y un resultado más notable es que precisamente las mismas posibilidades existen en el caso general de n ecuaciónes con n incógnitas, como veremos. Para algún ejemplo específico es posible descubrir que la situación es por solución directa de la ecuaciónes ó, equivalentemente, por usar la reducción triangular de la matríz aumentada, así consideremos las ecuaciones.

$$x_1 + 2x_2 - 5x_3 = 2$$

$$2x_1 - 3x_2 + 4x_3 = 4 \qquad (1)$$

$$4x_1 + x_2 - 6x_3 = 8$$

La matríz aumentada es,

$$\begin{bmatrix} 1 & 2 & -5 & 2 \\ 2 & -3 & 4 & 4 \\ 4 & 1 & -6 & 8 \end{bmatrix}$$

Las operaciónes por renglón reducen esto a,

$$\begin{bmatrix} 1 & 2 & -5 & 2 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Nótese que todos los elementos en el tercer renglón son cero

Esto significa que la tercera ecuación se ha reducido a $0x_3 = 0$ tal que una solución de la ecuación original existe para la cual $x_3 = k$, donde k es cualquier número

Por la sustitución hacia atrás, nos da.

$$x_2 = 2k$$

$$x_1 = 2 + k$$

Y ésta solución puede ser expresada de la forma

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} + k \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

Si de otra manera, la tercera ecuación en (1) hubiera sido

$$4x_1 + x_2 - 6x_3 = 0$$

encontrariamos en lugar de

$$\begin{bmatrix} 1 & 2 & -5 & 2 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

lo siguiente

$$\begin{bmatrix} 1 & 2 & -5 & 2 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

La tercera ecuación es ahora $0.x_3 = 1$, tal que este conjunto de ecuaciones es inconsistenete.

En los ejemplos consideramos, que el número de ecuaciones ha sido exactamente igual al número de incógnitas, pero los mismos métodos (llamados reducción de una matriz aumentada

a la forma triangular ó, equivalentemente, la solución directa de las ecuaciones), pueden ser utilizados para checar un conjunto dado de m ecuaciones con n incógnitas, para algún m,n, que posea ó no solución, una solución única, ó una infinidad de soluciones. Así podemos verificar los siguientes ejemplos siguiendo el método descrito anteriormente.

Por ejemplo, si la ecuación

$$x_1 + x_2 + x_3 = 6$$

es sumada al conjunto de ecuaciones

$$x_1 + 2x_2 - 5x_3 = 2$$

$$2x_1 - 3x_2 + 4x_3 = 4$$

$$4x_1 + x_2 - 6x_3 = 8$$

como una cuarta ecuación; el sistema resultante de 4 ecuaciones con 3 incógnitas, tiene solución única.

$$x_1 = 3$$

$$x_2 = 2$$

$$x_3 = 1$$

Si la ecuación

$$3x_1 - x_2 - x_3 = 6$$

es sumada al conjunto de ecuaciones

$$x_1 + 2x_2 - 5x_3 = 2$$

$$2x_1 - 3x_2 + 4x_3 = 4$$

$$4x_1 + x_2 - 6x_3 = 8$$

el sistema resulta de 4 ecuaciones con 3 incógnitas, tiene el mismo conjunto de soluciónes, del sistema original. El cual es

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} + k \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

Y el sistema de dos ecuaciones con 3 incógnitas, dado por las primeras dos ecuaciones en el sistema anterior, es decir

$$x_1 + 2x_2 - 5x_3 = 2$$

$$2x_1 - 3x_2 + 4x_3 = 4$$

tiene el mismo conjunto de soluciones como, el sistema original anteriormente descrito en (1)

La moral de estos ejemplo es que en general, no es posible decir si un conjunto de ecuaciones, no tiene solución, una solución única, ó una infinidad de soluciones, meramente del conocimiento del número de ecuaciones y del número de incógnitas.- es decir, m y n. Por tanto 10 ecuaciones con 2 incógnitas pueden tener una solución única, y 2 ecuaciones con 10 incógnitas pueden ser inconsistentes.

Consideremos el siguiente ejemplo.

$$x_1 + 2x_2 - x_3 = 2$$

$$2x_1 + 4x_2 + x_3 = 7$$

$$3x_1 + 6x_2 - 2x_3 = 7$$

Usando la primera ecuación para eliminar x_1 de las 2 restantes obtenemos.

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 2 \\ 3x_3 &= 3 \\ x_3 &= 1 \end{aligned} \qquad \begin{bmatrix} 1 & 2 & -1 & 2 \\ 0 & 0 & 3 & 3 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

La incógnita x_2 no aparece en la segunda y tercera ecuaciones.
Hemos reducido el coeficiente de x_3 en la segunda ecuación a la unidad por dividir ésta ecuación por 3. Podemos, entonces, omitir la tercera ecuación, porque es simplemente un múltiplo de la segunda. (en terminos de la matríz, los elementos (2,2) y (3,2) son cero. Y consideremos la tercera columna, reduce los elementos (2,3) a la unidad por multiplicar el renglón siguiente po 1/3, y restando el renglón resultante del último renglón, tal que el último renglón es reducido a un renglón de ceros).

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 2 \\ x_3 &= 3 \end{aligned} \qquad \begin{bmatrix} 1 & 2 & -1 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad (2\ ')$$

La solución general del conjunto original de ecuaciones está dado por

$$x_3 = 1 \ , \ x_2 = p \ , \ x_1 = 3 - 2p \qquad (2)$$

Donde tenemos el conjunto de $x_2 = p$, una constante arbitraria. Dado que de conjuntar x_2 igual a una constante

arbitraria, podriamos tener el conjunto de x_1 igual a una constante arbitraria, y entonces la solución general es

$$x_3 = 1 \quad , \quad x_1 = q \quad , \quad x_2 = 1/2(3 - q) \qquad (3)$$

Es claro que la soluciónes anteriores son esencialmente las mismas dado que, si asignamos cualquier valor a p en la primera solución general (2), la misma solución será dada al conjuntar q = 3 - 2p en (3). Sin embargo, en casos más complicados la equivalencia de formas diferentes de soluciónes puede no ser obvia. Deberia notarse que en el ejemplo de arriba no es posible asignar un valor arbitrario a x_3, dado que una ecuación es $x_3 = 1$

El procedimiento dado en la ecuación (2 ') corresponde a lo que hemos llamado previamente " reducción a la forma triangular ", con modificaciones apropiadas para el hecho de que los elementos cero en matrices sucesivas significa que no podemos tener una forma triangular estricta. Para ver la estructura de la solución de (2 ') más claramente, es conveniente usar la sustitución hacia atrás en las ecuaciones (en la matríz; usando el procedimiento análogo de Gauss - Jordann para reducir el elemento (1,3) a cero. Esto da

$$\begin{array}{l} x_1 + 2x_2 - \qquad = 3 \\ \qquad\qquad\quad x_3 = 1 \end{array} \qquad \left[\begin{array}{cccc} 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Esto nos dice inmediatamente que a cualquiera de x_1 ó x_2 (pero no ambos) puede asignarsele valores arbitrarios, pero a, x_3 no se le puede dar un valor arbitrario.

C A P I T U L O

V I.

INEXACTITUD INHERENTE EN SISTEMAS LINEALES (SOLUCIONES)

Dada una matríz no singular A y un vector b distinto de cero, sea $\bar{x}$ la solución del sistema.

$$A\,x = b$$

Supongamos que A y b son sujetos de incertidumbre. Cuál es la incertidumbre resultante en la solución x ?. Para cualquier vector columna y de orden n definimos $||\ y\ ||$, como la longitud euclidiana de y ;

$$||\ y\ ||_2 = ||\ y\ || = \sqrt{y_1^2 + y_2^2 , + \ . \ . \ . \ + y_n^2}$$

Para una matríz cuadrada A de n x n, definimos la norma espectral de $||\ A\ ||$, como.

$$||\ A\ || = \max_{||x|| = 1} ||\ Ax\ ||$$

Estas funciones dan por lo general, la medida ó tamaño de vectores y matrices respectivamente.
Para una matríz no singular A. Definimos la condición de A como cond(A), por la relación.

$$cond(\ A\ ) = ||\ A\ ||\ \cdot\ ||\ A^{-1}\ ||$$

El concepto de condición de una matríz, intenta medir la sensibilidad ó la " vulnerabilidad " de una solución : si cambiamos ligeramente A y b, cuán grande es el efecto en $x = A^{-1} b$? antes de comenzar con esta cuestión queremos señalar un obstaculo (que podemos salvar fácilmente). Tiene que haber una manera de medir el cambio δA y de calcular el tamaño de A misma.

Ya definimos la longitud de un vector; necesitamos ahora la norma de una matríz. Entonces el número de condición, y con él la sensibilidad de A, se podrán obtener fácilmente directamente de las normas A y A^{-1}

Uno de los principales usos del concepto de condición yace en respuesta, por ahora retornamos a la cuestión hecha al principio de este capítulo, Qué cambio produce en la solución x un cambio pequeño en b ó en A ?.

Comenzaremos con un cambio del lado derecho, de b a b + δb. Este erroe puede provenir de los datos experimentales ó de redondeo; podemos suponer que δb es pequeño pero que su dirección está fuera de nuestro control. La solución correspondiente cambia de x a x + δx :

$$A (x + \delta x) = b + \delta b ,$$

$$x + \delta x = A^{-1} b + A^{-1} \delta b$$

$$\delta x = A^{-1} \delta b$$

$$|| \delta || \leq || A^{-1} || \cdot || \delta b ||$$

Este es un caso particularmente sencillo; concideremos todas las perturbaciónes δb y calculamos la perturbación. resultante $\delta x = A^{-1} \delta b$. Habrá grandes cambios en la solución cuando A^{-1} sea grande (A es casi singular) y será particularmente grande cuando δb vaya en la dirección donde sea amplificado por A^{-1} .

y puesto que $A x = b$, tenemos ,

$$|| b || = || A x || \leq || A || \, || x ||$$

dividiendo

$$|| \delta x || \leq A^{-1} \cdot \delta b$$

por $|| b || = || A x || \leq || A || \cdot || x ||$

tenemos

$$\frac{|| \delta x ||}{|| x ||} \leq || A || \; || A^{-t} || \; \frac{|| \delta b ||}{|| b ||}$$

ó

$$\frac{\| \delta x \|}{\| x \|} \leq \text{cond}(A) \frac{\| \delta b \|}{\| b \|}$$

La última desigualdad muestra que la incertidumbre relativa a x es limitada por la cond(A). El límite en está expresión es alcanzable, para cualquier A no singular y b distinto de cero. Esto es fácil de ver, si llevamos a cabo un cambio de coordenadas en el cual A tome la forma diagonal.
Pero hay un serio inconveniente para medir así la sensibilidad. Supongamos que multiplicamos todas las entradas de A por 1,000; entonces γ (el error del tamaño γb se amplifica por el factor $1/\gamma$ que no es más que el valor propio mayor de A^{-1}. La simplificación es mayor cuando γ está cerca de cero, así que las matrices casi singulares (1 son las más sensibles) se multiplicará por 1,000 y la matríz será mucho menos singular. Pero se supone que estamos jugando limpio; no es posible que mediante un cambio de escala podamos convertir la matríz mal dispuesta en una bien colocada. Es cierto que δx será menor en un factor de 1,000, pero también lo será la solución $x = A^{-1} b$ y el error relativo $\| \delta x \| / \| x \|$ será el mismo. El factor x en el denominador normaliza el problema contra pequeños cambios de escala. Al mismo tiempo hay una normalización coorespondiente para δb; nuestro problema es comparar el cambio relativo $\| \delta b \| / \| b \|$ con el error relativo $\| \delta b \| / \| x \|$.
Lo peor sucede cuando el número δx es grande (las perturbaciones están en la dirección del vector propio x_1) y cuando el denominador x es pequeño. La solución sin perturbar x debería ser la más pequeña posible comparada con la b sin perturbar.

Ahora bien llevando a cabo un cambio de coordenadas, como una transformación lineal, a toma vectores x en vector b. Un teorema de fundamental importancia pero poco conocido establece que, por un cierto cambio ortogonal de coordenadas en el espacio x y por otro cambio ortogonal de coordenadas en el espacio de b, la matríz A puede ser puesta en la forma diagonal.

$$A = \begin{bmatrix} \mu_1 & & & 0 \\ & \mu_2 & & \\ & & \ddots & \\ 0 & & & \mu_n \end{bmatrix}$$

Aquí los números positivos $\mu_1, \mu_2, \ldots \mu_n$, son llamados los valores singulares de A además .

$$||A|| = \mu_1 \qquad ||A^{-1}|| = \mu_n^{-1}$$

Finalmente, las transformaciones ortogonales no cambian las normas de x y b, tenemos.

$$A^{-1} = \begin{bmatrix} \mu_1^{-1} & & & \\ & \mu_2^{-1} & & \\ & & \ddots & \\ & & & \mu_n^{-1} \end{bmatrix}$$

$$\text{si } b = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ 0 \end{bmatrix}, \quad \text{y} \quad \delta b = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ e \end{bmatrix}$$

$$x = A^{-1} b = \begin{bmatrix} \mu_1^{-1} \\ 0 \\ . \\ 0 \end{bmatrix}, \quad y \quad x = A^{-1} \delta b = \begin{bmatrix} 0 \\ . \\ 0 \\ e\mu_n^{-1} \end{bmatrix}$$

para estos vectores

$$\frac{||\delta x||}{||x||} = \frac{\varsigma\mu_1}{\mu_n} = \frac{||\delta b||}{||b||} \frac{\mu_1}{\mu_n} = \text{cond}(A) \frac{||\delta b||}{||b||}$$

Y está última expreción muestra que

$$\frac{||\delta x||}{||x||} \leq \text{cond}(A) \frac{||\delta b||}{||b||}$$

es una igualdad, como prometimos probar aunque.

$$\frac{||\delta x||}{||x||} = \frac{\varsigma\mu_1}{\mu_n} = \frac{||\delta b||}{||b||} \frac{\mu_1}{\mu_n} = \text{cond}(A) \frac{||\delta b||}{||b||}$$

es solo una igualdad exacta, bajo condiciones excepcionales es mejor cerrar la igualdad y por consiguiente damos por supuesto la igualdad aproximada. Si $\text{cond}(A) = 10^{p}$, y si b es correctamente conocida solo para 10 decimales, ahora x puede ser conocida para 10^{-p} decimales, entonces p puede tomar cualquier rango de 0 a x

El solo esperar tener alguna significancia para x, el sistema de computadora de 10 decimales, es que

$$\text{cond}(A)\ 10^{-10} < 1/2$$

En una computadora de base β con t digitos significativos necesitamos

$$\text{cond}(A)\,\beta^{-t} < 1/2$$

Para tener alguna significancia para una solución. Recordamos que todas las exposiciones en esta sección son independientes de cualquier método de resolver un sistema Ax = b. Son exposiciones acerca de los errores en x, los cuales son inherentes a la incertidumbre en los datos.

Sí A está sujeta a un cambio δA y b es conocido exactamente, entonces una desigual análoga a,

$$\frac{||\ \delta x\ ||}{||\ x\ ||} \leq \text{cond}(A)\ \frac{||\ \delta b\ ||}{||\ b\ ||}$$

es la siguiente

$$\frac{||\ \delta x\ ||}{||\ x + \delta x\ ||} = \text{cond}(A)\ \frac{||\ \delta b\ ||}{||\ A\ ||}$$

Sí x es pequeña comparado con || x ||, entonces podemos, considerar sin contratiempo alguno, el lado izquierdo de ésta última desigualdad, como un error relativo a x

Ejemplo.

Los valores propios de

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix}$$

Son aproximadamente $\gamma_1 = 10^{-4}/2$ y $\gamma_2 = 2$. Por lo tanto, su número de condición es alrededor de $c = 4 \cdot 10^4$ y debemos esperar un cambio violento en la solución debido a cambios nada extraños en los datos. En la solución.

$$u + v = 2 \qquad u + v = 2$$

$$u + 1.0001\, v = 2\ ; \qquad u + 1.0001\, v = 2.0001.$$

Los lados derechos se alteran solamente en $\delta b = 10^{-4}$. Al mismo tiempo, las soluciónes cambiaron de $u = 2$, $v = 0$ a $u = v = 1$. Este es un error relativo de

$$\frac{||\ \delta x\ ||}{||\ x\ ||} = \frac{||\ (-1,1)\ ||}{||\ (2,0)\ ||} = \frac{2}{2}$$

que es igual a

$$2 \cdot 10^{4}\ \frac{||\ \delta b\ ||}{||\ b\ ||}$$

Aun sin seleccionar alguna perturbación en especial (nuestra x y δb forma un ángulo de 45º en el peor de los casos, lo cual equivale a la pérdida del 2 en nuestro factor $2 \cdot 10^4$ comparado con ;a posibilidad extrema $c = 4 \cdot 10^4$) tuvimos un cambio tremendo en la solución.

Notese que el número de condición c, no se afecta directamente por el tamaño de la matríz; si $A = I$ ó aún si $A = I / 10$ el número de condición es $c = \gamma_{máx} / \gamma_{mín} = 1$

En comparación, el determinante es un criterio malísimo para probar la mala colocación de una matríz, ya que depende no sólo de la escala considerada, sino también del orden n ; si $A = I/10$ entonces el determinate de A es 10^{-n} De hecho, esta matríz " casi singular " está lo mejor colocada posible.

C A P I T U L O

V I I.

P R E C I S I O N R E A L I Z A B L E

E N

E L I M I N A C I O N G A U S S I A N A

En este capítulo daremos por hecho que el lector conoce la eliminación gaussiana, que es un método para resolver sistemas de ecuaciónes lineales. La principal decisión estratégica que contempla el diseñador del algoritmo es la elección de un elemento pivote único para cada uno de los n-1 pasos,en el cual una variable es eliminada de las ecuaciónes restantes.
Hay dos tipos de estrategias importantes :

i) Pivoteo completo.- En el cual cada paso uno selecciona como pivote algún elemento a_{ij} de valor absoluto máximo entre todos los elementos restantes de la matríz.

ii) Pivoteo parcial.- En el cual a cada paso uno selecciona como pivote algún elemento a_{ij} de valor absoluto máximo entre la primera columna de los elementos restantes de la matríz.

Así en el primer paso el pivoteo completo examinaría la sola matríz A, para un elemento maximal en valor absoluto, mientras que el pivoteo parcial examinaría solo la primera columna.

Algunas clases especiales de matrices permiten la eliminación para proceder sucesivamente sin algún pivote - por ejemplo, las matrices simétricas positivas definidas - pero generalmente el examen de pivote es esencial para garantizar el éxito. El siguiente ejemplo simple ilustra el desastre que produce al no examinar un pivote.

Consideremos una máquina de tres dígitos decimales en punto flotante.

El sistema es,

$$.0001\ x + 1.000\ y = 1.00$$

$$1.00\ \ x + 1.00\ \ y = 2.00$$

La verdadera solución, redondeando a cinco decimales es,

$$x = 1.00010 \qquad y = .99990$$

Si aceptamos el elemento .0001 como pivote, la eliminación de x en la segunda ecuación da la ecuación.

$$-10000\,y = -10000$$

Resolviendo por sustitución hacia atrás, encontramos que $y = 1.00$, mientras que $x = 0.00$, un claro desastre.
Por otra parte el pivoteo parcial seleccionaría el elemento pivote $a_{2,1} = 1.00$. La eliminación de x de la primera ecuación da la ecuación.

$$1.00\,y = 1.00$$

Resolviendo a la inversa encontramos que

$$y = 1.00$$

y entonces

$$x = 1.00$$

con éxito obvio.

Ahora diremos que estamos tratando con una computadora de punto flotante de base 2 y t digitos. Más bien que discutir la solución de un sistema lineal consideremos la computación de la inversa A^{-1} de una matriz dada. Deseamos establecer los límites de error de redondeo que habian sido probados por la eliminación gaussiana.
Asumiendo una estrategia de pivoteo completo y que la matríz A este razonablemente escalonada, al principio y en todos los pasos intermedios (ver cápitulo de Escalonamiento de Matrices) entonces, si todo $a_{ij} > 1$, un cierto algoritmo gaussiano da una matríz X tal que.

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} \leq (0.8)\, 2^{-t}\, 7/2\, n\, g(n)\, A^{-1}$$

Aquí g(n) es el máximo de todos los elementos de las matrices sucesivas encontadas durante la eliminación.
Para expresar el resultado anterior, en una forma para poder compararla con algunos resultados del cápitulo

anterior notemos que $1 < A < n$, tal que esperamos que $A = n^{\frac{1}{2}}$.
Entonces tenemos.

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} = n^{\frac{3}{2}} 2^{-t} \, cond(A) \, g(n)$$

Qué clase de límite podemos dar para g(n) ?.

Esto arroja una pregunta abierta, y lo que se sabe que resulta mejor es aproximadamente.

$$g(n) \leq 1.8 \, n^{(\frac{1}{2}) \log n}$$

Por otro lado para todas las matrices reales que han sido examinadas, siempre se ha observado que.

$$g(n) \leq n$$

el último límite es alcanzado para una n grande ilimitadamente por matrices relaciónadas a las matrices de Hadamard.

Para las demás matrices algúna ves se observa que.

$$g(n) \leq 8$$

Y se han encontrado matrices complejas A de una n ilimitadamente grande para lo cual.

$$g(n) \doteq 3.1 \, n$$

Debería ser más deseable tener un buen límite para g(n), tal que.

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} \leq (0.8) \, 2^{-t} \, n^{7/2} \, g(n) \, \| A^{-1} \|$$

Podría ser cambiado en un buen límite de error a - priori
para el cálculo de A^{-1}. Naturalmente para cualquier matriz A en particular, g(n) es facilmente observado en el curso de la eliminación, tal que en algún evento ésta relación

$$\frac{|| X - A^{-1} ||}{|| A^{-1} ||} \; (0.8)^{-t} \; 2^{7/2} \; n \; g(n) \; A^{-1}$$

se transforma en un límite de error a - posteriori.

Sin embargo todavía límites de error mejores queden estar dados a - posteriori, como será mostrado en el capitulo " Más Soluciónes Exactas ".

El factor 2^{-t} es esencialmente el nivel de insertidumbre inherente a los datos, y sería igualado a

$$\frac{|| \delta A ||}{|| A ||}$$

Entonces el límite en

$$\frac{|| X - A^{-1} ||}{|| A^{-1} ||} = n^{3} \; 2^{-t} \; cond(A) \; g(n)$$

es más grande que en

$$\frac{|| \delta x ||}{|| x + \delta x ||} \leq cond(A) \; \frac{|| \delta A ||}{|| A ||}$$

Por el factor n^{3} g(n). Tomando en cuenta el resultado empirico

$$g(n) \leq 8$$

que para más matrices reales, entonces interpretamos

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} \leq n^{3} 2^{-t} \, \mathrm{cond}(A) \, g(n)$$

como diciendo que la matriz X calculada, generalmente difiere de la verdadera inversa A^{-1}, enterminos relativos por no más de Sn^{3} veces el error inherente en el problema.

Así la eliminación gaussiana simple es razonblemente buena al tomar el límite, con el error de redondeo bajo control, para valores modestos de n. Mejores resultados pueden obtenerse con algunas ideas que mencionaremos en el siguiente capítulo.

El límite dado a.

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} (0.8) 2^{-t} n^{7/2} g(n) A^{-1}$$

fué

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} \leq (5.3 + 14.6 A^{2}) 2^{-t} n^{2} \| A^{-1} \|^{2}$$

$$\approx 15n^{2} 2^{-t} (\mathrm{cond}(A))^{2}$$

El factor $(\mathrm{cond}(A))^{2}$ surge de resolver

$$A^{t} Ax = A^{t} b \text{ , más bien que}$$

$$Ax = b$$

La prueba de

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} \leq (5.3 + 14.6\, A^{2})\, 2^{-t}\, n^{2}\, \| A^{-1} \|^{2}$$

$$\approx 15n^{2}\, 2^{-t}\, (\mathrm{cond}(A))^{2}$$

es mucho más dificil y tediosa que la prueba de

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} (0.8)\, 2^{-t}\, n^{7/2}\, g(n)\, A^{-1}$$

C A P I T U L O

V I I I.

M A S S O L U C I O N E S
E X A C T A S

Supongamos que la matríz A está dada como un dato en precisión simple y que deseamos conseguir soluciones que garanticen ser más precisas, que los límites,

$$\frac{\| X - TA^{-1} \|}{\| A^{-1} \|} = (5.3 + 14.6\ A^{2})\ 2^{-t}\ n^{2}\ \| A^{-1} \|^{2}$$

$$\approx 15n^{2}\ 2^{-t}\ (\mathrm{cond}(A))^{2}$$

donde el condicional $(\mathrm{cond}(A))^{2}$

surge de resolver $A^{t}Ax = A^{t}b$

más bien que de

$$Ax = b$$

Ahora la cuestión es, Cómo procederíamos ?. La elección más obvia es ejecutar los cálculos en doble precisión entonces t es sustituido por 2t en los límites de arriba y hasta 2^{-2t} es mucho más pequeño que 2^{-t}, con esto ganamos muchos órdenes de magnitud en precisión. El costo en tiempo de computación varía entre diferentes máquinas, pero es sólo un factor de cuatro en la IBM.

El costo del almacenaje es grande, hasta debemos duplicar el almacenaje reservado para la matríz desarrollada.

Cuando el costo del tiempo y almacenaje son muy grandes para justificar el completar la doble precisión, es posible aprovecharlo sustancialmete, por un uso más límitado de doble precisión. Muchas de las operaciones en la eliminación gaussiana pueden ser expresadas como pruductos internos de vectores de números de precisión simple. En muchas máquinas es posible acumular como un producto interno en doble precisión y entonces redondearlo para precisión simple, antes de almacenar. El resultado de esta acumulación es reducir al máximo el error de redondeo de un producto interno por un factor de n.

El efecto total arroja el reducir el límite del error de redondeo por un factor $n^{5/2}$. Así, en lugar del resultado.

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} = (0.8)\, 2^{t}\, n^{7/2}\, g(n)\, A^{-1}$$

Una eliminación con pivoteo y acumulación produce una inversa aproximada X tal que

$$\frac{\| X - A^{-1} \|}{\| A^{-1} \|} = 3.3\, n\, 2^{-t}\, A^{-1}$$

Bajo ciertas hipótesis adicionales, el provecho del factor $n^{5/2}$ es muy sustancial. aunque la experiencia muestra que los errores actuales en computación como precisión simple son usualmente un tanto más pequeños que los límites. Una desventaja teórica de la estrategia de pivoteo completo es que no se mescla bien con la acumulación de productos internos. Cuando los productos son acumulados, casi siempre se usa una estrategia de pivoteo parcial y acepta la posibilidad teórica de que estos pivoteos puedan crecer mucho.

Un tercero y el aprovechamiento más afortunado para incrementar la exactitud en soluciones de sistemas lineales densos almacenados, es el tan llamado método de refinamiento interativo. Describámoslo brevemente.

Cuando un sistema lineal Ax = b ha sido resuelto usando una factorización estable de A tal como el dado en la eliminación gaussiana con pivoteo ó una triangularización ortogonal entonces una solución más precisa puede encontrarse usando esta factorización dando el residual correspondiente a la solución aproximada, es calculada usando gran precisión. Este proceso puede ser usado de una manera interativa para dar una solución mejorada progresivamente. Digamos que el método directo es tal, que

corresponde a algún c del lado derecho que da la solución exacta de algunos

$$(A + E_c) x = c \qquad \text{donde} \quad \| E_c \| \leq \| A \|$$

Donde E_c es una función de c , pero el límite en E_c es independiente de c. Entonces si el residual pudiera ser exactamente un proceso iterativo. podría ser definido por.

$$x_0 = 0 , \quad r_k = b - Ax_k , \quad (A + E_k) \delta_{k+1} = r_k$$

$$x_{k+1} = x_k + \delta_k$$

Aquí damos por hecho que, no solamente cada una de las r_k está computada exactamente, pero también que este residual puede ser usado por el lado derecho y que la correción calculada δ_k puede ser agregada a la apriximación x_k sin error. Los únicos errores son aquellos en resolver los sistemas lineales.

si escribimos

$$e_k = x - x_k$$

entonces de las ecuaciónes

$$x_0 = 0$$

$$r_k = b - Ax_k$$

$$(A + E_k) \delta_{k+1} = r_k$$

y

$$e_k = x - x_k$$

Resulta,

$$e_{k+1} = (I - (A + E_k)^{-1} A) e_k = P_k e_k$$

Donde

$$P_k = I - (A + E_k)^{-1} A$$

y,

$$|| P_k || \leq \varepsilon_k /(1 - \varepsilon E_k)$$

De aquí si

$$\varepsilon_k /(1 - \varepsilon_k) \div \beta^{-p}$$

La solución gana al final p dígitos de precisión para la iteración.

En la práctica el residual r_k no puede ser calculado exactamente, pero si se puede acumular productos internos en doble precisión, el error puede ser mantenido a un nivel muy bajo. El residual entonces, puede ser redondeado en precisión simple y usado como el próximo lado derecho y δ_{k+1} puede ser obtenido usando la factorización de A.
Esta prcisión simple δ_{k+1}, es entonces sumada a x_k y el vector resultante redondeado a precisión simple de la x_{k+1} calculada. Dado que cada x_k es un vector en precisión simple, la precisión final obtenible está obviamente limitada a precisión simple.

Por este método, si la matríz A no esta demasiado mal condicionada, en la práctica se consiguen las soluciones las cuales son las aproximaciones correctas redondeadas a la verdadera respuesta. Describiendo este desarrollo en otras palabras tenemos : Supongamos que por la eliminación gaussiana se ha encontrado una primera solución aproximada x_0 del sistema lineal Ax = b. El próximo paso es formar el vector residual $r_0 = b - Ax_0$. Si x_0 era la

solución exacta del sistema podriamos tener $r_0 = 0$, el vector nulo.
Si no resolvemos el nuevo sistema lineal $Ay = r_0$, para obtener un vector y_1, Sea

$$x_1 = x_0 + y_1 .$$

El proceso es repetido iterativamente, es decir para $k = 0, 1,1,. . .,$ formamos el residual.

$$r_k = b - Ax_k$$

resolver el sistema $Ay = r_k$, para obtener un vector y_{k+1}, y entonces formar

$$x_{k+1} = x_k + y_{k+1}$$

bajo hipótesis apropiadas para ser especificado bajo la serie x_k que converge a la verdadera solución A^{-1} del sistema $Ax = b$.

Varias cuestiones se necesitan para esclarecer esta algoritmo.

PRIMERO.- Parece involucrar una gran cantidad de trabajo para resolver sistemas de la forma $Ay = r_k$, para muchos valores de k. De hecho esto no es suficiente.
La eliminación gaussiana para resolver un sistema $Ax = b$, involucra tres pasos principales;

i) Triangularización de la matríz A por transformaciones elementales de renglón.

ii) Aplicaciones de las mismas transformaciones por renglón al lado derecho b;

iii) La solución del sistema triangular por sustitución hacia atrás

$3/3$

El producir el paso i) requiere aproximadamente de n adiciones y multiplicaciones, pero los pasos ii) y iii) 2 requieren solo de n multiplicaciones y adiciones aproximadamente. El paso i) solo es necesario hacerlo una vez para todos los sistemas $Ay = r_k$. Si los multiplicadores definen que las transformaciones por renglón están listas, los pasos ii) y iii), entonces pueden ser realizados rapidamente para cada nuevo sistema $Ay = r_k$ en turno. Como un resultado, se ve que una serie suficientemente larga de vectores x_k puede por lo general, ser calculada en más o menos 20% más tiempo que el del cálculo en la primera solución x_0.

Es absolutamente indispensable que cada vector residual r_k sea calculado, para gran precisión. Esto es un hecho normal para la acumulación de productos internos en doble precisión, seguido por el redondeo de la respuesta en precisión simple en la forma de punto flotante. Si r_k es calcu;ado con un producto interior en precisión simple, tendrá errores de redondeo de varias unidades en los mínimos dígitos significativos de x_k. Entonces la desigualdad.

$$\frac{||\ \delta x\ ||}{||\ x\ ||} \leq \text{cond}(A)\ \frac{||\ \delta b\ ||}{||\ b\ ||}$$

será una igualdad aproximada en la práctica, digamos que x_k será fuerte en algunos casos, cuando el cond(A) sea el mínimo dígito significativo. Puesto que el cond(A) puede 4 5 bien ser 10 ó 10. La precisión resultante en x_k es

muy bajo y, en efecto, x_k es también tan precisa como cualquier x_0.

El siguiente teorema de las bases del método expuesto anteriormente del refinamiento iterativo.

TEOREMA.

Sea la matríz A, la cual tiene la propiedad de

$$(0.8)^{2-t}\, n^{7/2}\, g(n) \left\| A^{-1} \right\| < 1/2$$

Sea el algoritmo de arriba, llevándose a cabo, con cada sistema $Ay = r_k$, siendo resuelto en simple precisión en aritmética de punto flotante con base 2, pero con computaciónes de,

$$r_k = b - Ax_k$$

y,

$$x_{k+1} = x_k + y_{k+1}$$

ejecutando sin error de redondeo, entonces,

$$\left\| x_k - A^{-1} b \right\| \longrightarrow 0$$

cuando $k \longrightarrow \infty$

Si la solución de los sistemas $Ay = r_k$, están hechos con acumulaciones de producto internos en doble precisión, entonces el lado izquierdo de

$$(0.8)^{2-t}\, n^{7/2}\, g(n) \left\| A^{-1} \right\| < 1/2$$

puede ser reemplazado por el lado derecho de

$$\frac{\left\| X - A^{-1} \right\|}{\left\| A^{-1} \right\|} \leq 3.3\, n\, 2^{-t} \left\| A^{-1} \right\| g(n)$$

En la práctica, por supuesto, r_k es computado por una acumulación de productos internos en doble precisión y x_{k+1} es computado como la suma en punto flotante x_k y y_{k+1}.

Como un resultado la serie x_k no converge a $A^{-1} b$ en el sentido matemático.

En lugar de ello, x_k es observada para convertirse constantemente a un valor, el cual es normalmente el correcto redondeo en precisión simple. (de la aproximación $A^{-1} b$). En el uso actual del refinamiento iterativo, generalmente no se conoce el avance, si la hipótesis.

$$(0.8)^{2-t}\, n^{7/2}\, g(n) \left\| A^{-1} \right\| < 1/2$$

es ó no sastisfecha y puede no ser concluida después de uno u otro. En la práctica es por consiguiente normal contar con el siguiente resultado.

Sea el algoritmo de arriba, ejecutado en cada sistema $Ay = r_k$, siendo resuelto por la misma versión de eliminación gaussiana, con cada r_k, computado por una acumulación de productos internos en doble precisión, y con x_{k+1}, computado con la suma de punto flotante de x_k y y_{k+1}

Sí para $k \geq k_0$, todos los vectores x_k son iguales para

algún vector x *, en doble precisión, entonces x * es la aproximación correctamente redondeada en precisión simple a
 -1
A b.

Este resultado, no puede ser probado, y realmente la mayoría de las computadoras ocuparían todo su tiempo en la aplicación del algoritmo de arriba, en algún ejemplo práctico.

Normalmente cuando el cond(A), consigue acercarse a 2^t los vectores x_k obviamente divergen, entonces no hay remedio excepto para incrementar la precisión con lo cual la eliminación es llevada a cabo, a menos que escalonar A ayudara, el valor de k es generalmente 3 ó 4.

C A P I T U L O

I X.

E S C A L O N A M I E N T O

O P T I M O D E

M A T R I C E S

Un aspecto que fue mencionado en el capítulo referente a Sistemas de Ecuaciones Lineales, fué el escalonar una matríz A, antes de resolver un sistema Ax = b, los términos alternativos para escalonar una matríz, están precondicionados y equilibrados.

De nuevo veamos el condicional. Una definición de condición de una matríz, de manera más natural, es el radio de su mínimo límite superior y de su máximo límite inferior, con respecto a un par de normas.

Algunas veces, los renglones ó columnas de una matríz son un poco grandes para algunos factores; en cualquier caso escalonar una matríz es trivial, computacionalmente hablando, por consiguiente, el problema de minimizar la condición por escalinamiento tiene varias aplicaciónes y deseamos determinar el ínfimo ó mínimo para derivar los límites inferiores y superiores. Para esto queremos conocer la dirección del escalonamiento al mínimo ó al último que se le acerque y queremos mostrar condiciones para que una matríz esté optimamente escalonada.

Así el problema puede ser completamente resuelto por la condición subordinada a un par de normas máximas; y así
-1
cada una de las matrices A y A , tienen una distribución en bloque cuadriculares. Para derivar los límites superiores, se debe determinar un par de vectores mutuamente duales.

Sean B_1 y B_2 dos espacios vectoriales normados, con

$$\left|\left| x_i \right|\right|_i \qquad \text{para } x_i \in B_i$$

Para un mapeo A de B_2 en B_1 , el mínimo límite superior
(mín. lím. sup.)-el cual es una norma-y el máximo límite inferior (máx. lím. inf.) están definidos por.

$$\text{mín.lím.sup. } (A) = \sup_{x=/=0} \frac{||\ Ax\ ||}{||\ \ x\ ||}$$

$$\text{máx.lím.inf. } (A) = \inf_{x=/=0} \frac{|| Ax ||}{|| x ||}$$

Entonces el condicional subordinado a las normas $|| Ax ||$ y $|| x ||$ se puede definir por;

$$\text{cond}(A) = \text{mím. lím. sup } (A) \;/\; \text{máx. lím. inf } (A)$$

si tenemos que

$$\text{máx. lím. inf } (A) > 0$$

y

$$\text{máx. lím. inf } (A) = 0$$

podemos decir que $\text{cond}(A) = \infty$.

El condicional puede ser totalmente finito para una A rectangular, sin embargo.

$$\text{cond}(A) = \text{mím. lím. sup } (A) \;\; \text{mín. lím. sup } (A^{-1})$$

A^{-1} si A existe.

En lo sucesivo diremoa que A es un mapeo regular. Mencionaremos algunas propiedades básicas de la función cond(A) las cuales son válidas independientemente de las normas fundamentales.

1) $\text{cond}(A) \geq 1$

2) $\text{cond}(A) = 1$, implica que $|| Ax || = \delta || x ||$ para una cierta $\delta > 0$, independiente de x.

el cond(A) = 1, se ve inmediatamente de la definición

$$\text{cond}(A) = \text{mím. lím. sup } (A) \;/\; \text{máx. lím. inf } (A)$$

dado que el infímo no execede el supremo y significa igualmente que.

$$\text{mím. lím. sup } (A) \;\; \text{máx. lím. inf } (A) = \delta$$

ó

$$|| Ax || \leq \delta || x || \quad \text{y} \quad || Ax || = \delta || x ||$$

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

$$4)\ \text{cond}(A\,B) \leq \text{cond}(A)\ \text{cond}(B)$$

Se ve de la definición

$$\text{cond}(A) = \text{mím. lím. sup } (A)\ \ \text{mín. lím. sup } (A^{-1})$$

si A^{-1} existe que,

$$\text{cond}(A\,B) \leq \text{cond}(A)\ \text{cond}(B)$$

se cumple dado que las normas mín. lím. sup. son multiplicativas.

La importancia de los números condición está basada en el hecho de que dan los mejores límites posibles en las desigualdades.

$$\frac{1}{\text{cond}(A)}\ \frac{||\ x\ ||}{||\ y\ ||} \leq \frac{||\ Ax\ ||}{||\ Ay\ ||} \leq \text{cond}(A)\ \frac{||\ x\ ||}{||\ y\ ||}$$

Por consiguiente, las condiciones aparecen, por ejemplo en la estimación de eigen valores para matrices diagonalizables, donde P, puede ser un sistema de eigenvectores diagonales para A, entonces,

$$P^{-1}\,AP = \text{diag}(\gamma_i)\ \text{y por consiguiente}$$

$$\text{mín. lím. sup.}(A) = \text{cond}(A)\ \text{máx}\ |\ \gamma_i\ |$$

Para cualquier norma absoluta $||\ x\ ||$, donde:

$$\text{mím. lím. sup}\ (\ \text{diag}\ (\gamma_i)\ ) = \text{máx}\ |\ \gamma_i\ |$$

finalmente, se save que el análisis de error de métodos de eliminación para matrices positivas definidas, que la condición (por ejemplo el de uno subordinado a la norma

euclidiana) está caracterizada por la amplificación posible de errores de redondeo simple.

Si A es la matríz de un sistema de ecuaciónes líneales Ax = b, el escalonar los renglones de un sistema, suma a una trasformación de A por una matriz diagonal D_1, por la izquierda y el escalonar las sumas no conocidas a la transformación de A por una matríz diagonal D_2 por la derecha. Si A es hermitiana, ésta propiedad se preserva si $D_1 = D_2$. Dado que las transformaciónes diagonales de una matríz son modificaciónes triviales y dado que de otro modo la calidad de las computaciónes numéricas son generalmente mejoradas si la condición de la matríz concerniente es disminuida. Ha llavado a investigar los cuatro problemas de minimización, para conseguir el pré-escalonamiento de una matríz.

i) $\inf \operatorname{cond} D_1 D_2 (D_1 A D_2)$

ii) $\inf \operatorname{cond} D_2 (A D_2)$

iii) $\inf \operatorname{cond} D_1 (D_1 A)$

iv) $\inf_D \operatorname{cond} (D A D^H)$ para A hermitiana.

Aquí podemos asumir que D_1, D_2, D, son no singulares para el problema.

$$\inf. \operatorname{cond} D_1 D_2 (D_1 A D_2)$$

y en el caso hermitiano el problema $\inf_D \operatorname{cond} (D A D^H)$ para una A hermitiana, son presentados por ejemplo en el análisis de error de métodos directos para la solución de ecuaciónes líneales.

El problema del $\inf. \operatorname{cond} D_1 D_2 (D_1 A D_2)$, es

importante para obtener los mejores límites posibles para la inclusión de teoremas, finalmente la cantidad $\inf_{D_2} \text{cond}(AD_2)$ está también una medida natural para la independencia líneal de los vectores columna los cuales forman A.

En esta relación, A puede ser una matríz rectangular.

Ahora supongamos un ejemplo numérico de un sistema de 2 x 2 alterado por multiplicar la primera ecuación por 10^5. Entonces el sistema sería:

$$10.0\,x + 100{,}000\,y = 100{,}000.$$

$$1.0\,x + 1.0\,y = 2.00$$

El efecto de escalonar es hacer 10.0 el mayor pivote es la primera columna. Entonces la eliminación de x en la seguna ecuación del sistema de arriba en aritmética de punto flotante en tres dígitos, resultará en una nueva segunda ecuación.

$$-10{,}000\,y = -10{,}000.$$

La solución al revéz lleva a y = 1.00 y el resultado tremendo de x = 0.00.

Veamos que el escalonar pobremente con una buena estrategia de pivoteo nos forza a caer en el mismo error de redondeo enorme que obtenemos del conjunto original de ecuaciónes y una mala estrategia de pivoteo.

La conclusión de esto es que una buena estrategia de pivoteo es solo buena cuando la matríz está propiamente escalonada. Sin embargo puede ser omitido el método que no conocemos tanto por algoritmos garantizados para matrices bien escalonadas.

Es normal escalonar matrices por multiplicacion simplemente de renglones y columnas por factores. En efecto,se eligen matrices diagonales no singulares D_1 y D_2, y entonces se escalona A por la transformación.

$$A \longrightarrow D^{-1} A D_1 D_2$$

Porque el cond(A) es un factor de todos nuestros límites de error y teoremas de convergencia.

Es natural el desear seleccionar D_1 y D_2, tanto para reducir el $cond(D_1^{-1} A D_2)$ como para reducir un valor tan razonablemente como sea posible. Generalmente se usan potencias de la base de punto flotante para escalonar factores, para evitar la introducción de errores de redondeo en el escalonamiento. O alternativamente, se puede usar el escalonamiento sólo implícitamente, sin alterar actualmente los elementos de A.

TEOREMA (B. L. Bauer)

Sí el orden del conjunto de elementos pivoteados es seleccionado para avanzar en el escalonamiento de una matríz A, por potencias de punto flotante, no cambia un simple dígito el significado de algún número final intermedio en la solución Ax = b, por eliminación gaussiana.

Este teorema fué presentado por Bauer. Así el único efecto posible del escalonamiento de A, en los errores de redondeo, debe ocurrir a travez de cambiar el orden de los pivotes.

El ejemplo mostró que el cambio en los pivotes puede hacer una gran diferencia en el resultado.

Algunas veces se advierte al escoger D_1 y D_2, tal que la matríz resultante $D_1^{-1} A D_2$, tiene máximo elemento en cada renglón y cada columna (en valor absoluto) en el intervalo (.1 , 1), en cuanto al número base que se este usando.

Sin embargo, se ha mostrado que esto no siempre conduce a un buen escalonamiento,si.

$$A = \begin{bmatrix} 1 & 1 & 2 \times 10^{9} \\ 2 & -1 & 10^{9} \\ 1 & 2 & 0 \end{bmatrix}$$

Entonces ambas, de las siguientes matrices están escalonadas decimalmente y son equivalentes a A.

$$A_c = \begin{bmatrix} .1 & .1 & .2 \\ .2 & -.1 & .1 \\ .1 & .2 & 0 \end{bmatrix}$$

$$A_R = \begin{bmatrix} 10^{-10} & -10^{-10} & .2 \\ 10 & 10 & \\ 2 \times 10^{-10} & -10^{-10} & .1 \\ .1 & .2 & 0 \end{bmatrix}$$

Sin embargo, A_c es una matríz bien condicionada que no ofrece dificultades en la solución de un sistema de ecuaciónes mientra que A_R está mal condicionada y da muchos problemas para la eliminación.

Se ha estudiado el problema de encontrar D_1 y D_2 para minimizar el condicional $(D_1 A D_2)^{-1}$.

De esto se deduce que la solución depende de ciertas propiedades de las matrices no negativas $|A| \cdot |A^{-1}|$ y $|A^{-1}| \cdot |A|$

Claramente podemos calcular A^{-1} para encontrar un escalonamiento razonable, tal que podemos calcular A^{-1}, y está es una cuestión abierta, como encontrar un algoritmo para escalonar, demostrablemente bueno y conveniente. Los algoritmos existentes, son unos y otros muy superficiales ó potencialmente muy bajos. El lado bueno de ésta cuestión del escalonamiento, es que está visto que es para tratar con una matriz rara, la cual despues de un buen escalonamiento la cambia de intratable a tratable.

Así el escalonar optimamente para la norma máxima $(p_1 = p_2 = \infty)$ está caracterizada por $D_1 A$ ó $(A D_2)$ respectivamente, teniendo sumas de renglónes iguales a valores absolutos.

Ejemplo.

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 6 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} 3 & -3 & 1 \\ -3 & 5 & -2 \\ 1 & -2 & 1 \end{bmatrix}$$

Escalonar optimamente para la norma máxima.

$$D\,A\,D_{1\;\;2} = \begin{bmatrix} 3/5\ \sqrt{10} & 3 & 2/5\ \sqrt{10} \\ 1 & \sqrt{10} & 2 \\ 1/5\ \sqrt{10} & 3 & 4/5\ \sqrt{10} \end{bmatrix}$$

$$D_{2}^{-1}\,A\,D_{1}^{-1} = \begin{bmatrix} 1/2\ \sqrt{10} & -3 & 1/2\ \sqrt{10} \\ -1 & \sqrt{10} & -2 \\ 1/4\ \sqrt{10} & -3 & 3/4\ \sqrt{10} \end{bmatrix}$$

Donde el condicionál $(D_1\,A\,D_2) = (3 + \sqrt{10})^2 = 37.974$

$$A\,D_2 = \begin{bmatrix} 7 & 10 & 4 \\ 7 & 20 & 12 \\ 7 & 30 & 24 \end{bmatrix}$$

$$D_2^{-1}\,A^{-1} = \begin{bmatrix} 3/7 & -3/7 & 1/7 \\ -3/10 & 1/2 & -1/5 \\ 1/4 & -1/2 & 1/4 \end{bmatrix}$$

Donde el cond $(A\,D_2) = 61$

$$D_1 A = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/6 & 1/3 & 1/2 \\ 1/10 & 3/10 & 3/5 \end{bmatrix}$$

$$A^{-1} D_1^{-1} = \begin{bmatrix} 9 & -18 & 10 \\ -9 & 30 & -20 \\ 3 & -12 & 10 \end{bmatrix}$$

Cond (D_1 A) = 59

Ahora el escalonamiento optimo para la norma euclidiana

(Ambas A y A^{-1} tienen distribución de signo cuadricular)

$$D_1 A D_2 = \begin{bmatrix} 3/5\sqrt{10} & 3 & 2/5\sqrt{5} \\ 3 & \sqrt{10} & 6 \\ 2/5\sqrt{5} & 6 & 4/5\sqrt{10} \end{bmatrix}$$

$$D_2^{-1} A^{-1} D_1^{-1} = \begin{bmatrix} 1/2\sqrt{10} & -\sqrt{3} & 1/2\sqrt{5} \\ -\sqrt{3} & \sqrt{10} & -\sqrt{6} \\ 1/2\sqrt{5} & -\sqrt{6} & 3/4\sqrt{10} \end{bmatrix}$$

Donde el cond ($D_1 A D_2$) = $(3 + \sqrt{10})^2$ = 37.974

Que es el mismo del escalonamiento para la norma máxima.

C A P I T U L O

X.

ANALISIS DE ERROR DE REDONDEO

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 centrararon 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 límite para el error en la solución computada abandonando las más 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 sí 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 capítulo.

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 sí 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 matématico 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 operaciónes 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 implicito en el uso de esta notación que a y b son números y fl(a * b) es, por definición un número dígital. Es el resultado establecido por la computadora después de cualquier redondeo, corte, etc. y la renormalización puede ser necesaria. Aunque se le de gran importancia a los intentos que están hechos para asegurar que el futuro de las computadoras tendrá unidades aritméticas realmente sastifactorias, 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 operaciónes 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.

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

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 computado 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 implicaciónes, las cuales nos pueden ilustrar en el caso de la multiplicación. La relación

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

establece que el producto computado es el producto exacto de $a(1 + \epsilon)$ y b ó de a y, $b(1 + \epsilon)$, ó más simétricamente, de $a(1 + \epsilon/2)$ y $b(1 + \epsilon/2)$. dado que ϵ es pequeña podemos decir que el resultado computado 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 computados 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 podriamos tomar a/2 y $2b(1 + \epsilon)$ pero esto no involucra perturbaciónes

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.

Concideremos cualquier algoritmo computado 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 computados y_j no serán exactos, pero en general, será un número infinito del conjunto de elementos $x_i (1 + \epsilon)$ para los cuales los y_j computados 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 operaciónes aritméticas básicas los valores naturales de ϵ son sugeridos por las

relaciónes fundamentales sastifechas por los valores computados; 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 operaciónes involucradas también lleva de una manera equitativamente natural a un conjunto de datos apropiados.

$$x_i (1 + \epsilon_i)$$

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

arriba para $k\beta^{-t}$, donde k es la constante en la expresión (1). Para un algoritmo muy complicado sería razonable esperar el poder garantizar guardar el ε_i bajo este nivel.

Realmente si el número de operaciónes involucradas en el algoritmo es mp, es decir, un promedio de m operaciónes por elemento de datos, un algoritmo para el cual podríamos establecer límites a-priori de $mk\beta^{-t}$, para la ε_i podría ser considerado como extremadamente estable con respecto a errores de redondeo.

Un algoritmo para el cual los límites existen sastifactoriamente para ε_i, se dice que es inversamente 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_i correspondiendo a A 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 computada es decir, la distancia entre la y_j computada y la y_j exacta correspondiente a la x_i dada. La estabilidad inversa no garantiza que la solución computada 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

computado 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 rara 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 computada 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 algebraicos 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.

Análisis de Error de Multiplicaciones Anidadas.

Una ilustación de las tres caracteristicas que hemos mencionado es estipulando en un contexto de gran simplicidad formal, por la evaluación de un polinomio por multiplicaciones anidadas.

Sea,

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_0 \qquad (3)$$

Entonces f(x) puede ser evaluado por el algoritmo.

$$s_n = a_n$$

$$s_r = x\, s_{r+1} + a_r \qquad \text{donde } r = n-1, \ldots, 2, 1 \qquad (4)$$

$$f(x) = s_0$$

Podemos llevar a cabo el análisis de error inverso de este algoritmo en el cual, incluimos el efecto de cualquier error efectuado en las operaciones aritméticas incluidas, posiblemente grandes desatinos. Denotemos las cantidades computadas por $\bar{s}_r$, y tenemos.

$$\bar{s}_r = (x\, \bar{s}_{r+1} + a_r) = (x\, \bar{s}_{r+1} + e_r) + a_r + f_r \qquad (5)$$

computada

Donde e_r denota el error efectuado al calcular $x\, \bar{s}_{r+1}$ (esto podría ser meramente un error de redondeo ó un error sustancial debido a un mal funcionamiento de alguna clase) De aquí.

$$\bar{s}_r = x\, \bar{s}_{r+1} + (a_r + e_r + f_r) \qquad (6)$$

Muestra que las $\bar{s}_r$ computadas son exactas para el polinomio con coeficiente $a_r + e_r + f_r$. Nótese la gran simplicidad formal de este resultado; sin intersecciones entre los errores sucesivos que han sido conciderados. Supóngase, por ejemplo, que estamos trabajando con a entera y una x_i

entera los únicos errores cometidos son desatinos. Supóngase que solo dos errores son cometidos, y

que estos ocurrieron cuando se calculaban $\bar{s}_p$ y $\bar{s}_q$.

Entonces $\bar{s}_o$ (el valor computado de f(x) sería el valor exacto de un polinomio, el cual difiere del polinomio dado solo en los elemntos a_p y a_q , los cuales serían argumentos precisamente por la magnitud de los errores en los pasos p-ésimos y q-ésimos respectivamente.

Como un ejemplo sencillo consideremos la evaluación del polinomio.

$$f(x) = 3x^4 + 4x^3 + 2x^2 + x + 5 \qquad \text{con } x = 3. \qquad (5)$$

La evaluación correcta está descrita obviamente por la tabla

3	4	2	1	5	
	9	39	123	372	
3	11	41	124	377	f(3) = 377

Ahora consideremos la evaluación incorrecta descrita por.

3	4	2	1	5
	10	45	142	435
3	15	47	145	440

Observemos que aquí los errores fueron cometidos en la segunda columna; un error de +1 en la multiplicación, un +1 en la adición. Estos son contados según por una

perturbación de +2 en el coeficiente de x^3. Los errores también son cometidos en la cuarta columna; un error de +1 en la multiplicación y un erroe de +2 en la adición. Estos son contados según por una perturbación de +3 en el coeficiente de x. De aquí el valor computado es el valor exacto de x = 3 de,

$$3x^4 + (4 + 1 + 1 + x^3 + 2x^2 + (1 + 1 + 2) x + 5$$

es decir,

$$3x^4 + 5x^3 + 2x^2 + 4x + 5 \qquad (6)$$

Nótese que la diferencia entre los elementos correspondientes en las columnas en la computación exacta y erronea son irrelevantes y no viene al caso el análisis de error.

Retornando ahora al análisis de error de redondeo es conveniente, espresar el efecto que hace el error de redondeo en la adición y sustracción de la forma,

$$fl(a \pm b) = (a \pm b)/(1 + \epsilon) \text{ , donde } | \epsilon | \leq k\beta^{-t} \qquad (7)$$

Mejor que la expresión

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

La computación se $\bar{s}_r$ toma lugar en los pasos

$$b_r = fl(x\bar{s}_{r+1}) = x\bar{s}_{r+1} (1 + \epsilon_r) , \quad | \epsilon_r | \leq k\beta^{-t} \qquad (8)$$

$$\bar{s}_r = fl(b_r + a_{r+1}) = (b_r + a_{r+1})/(1 + \eta_r) , \quad | \eta_r | \leq k\beta^{-t} \qquad (9)$$

dando

$$\bar{s}_r = b_r + a_{r+1} - \bar{s}_r \eta_r$$

$$= \bar{x}\bar{s}_{r+1}(1 + \epsilon_r) + a_{r+1} - \bar{s}_r \Phi_r$$

$$= \bar{x}\bar{s}_{r+1}(a_{r+1} + \bar{x}\bar{s}_{r+1}\epsilon_r - \bar{s}_r \Phi_r) \qquad (10)$$

De aquí la f(x) computada es exacta para el polinomio con coeficientes $a_{r+1} + \bar{x}\bar{s}_{r+1}\epsilon_r - \bar{s}_r \Phi_r$. El análisis descrito aquí es mencionado como un análisis de error inverso corrido; para las perturbaciónes es el conjunto de datos en términos de los valores computados intermedios actuales (aquí la $\bar{s}_r$) tal que, cuando la computación ha sido completada tenemos a la mano límites para las perturbaciónes equivalentes en el conjunto de datos. Nótese que si ni $|\bar{x}\bar{s}_{r+1}|$ ni $|\bar{s}_r|$ es significativamente más grande que $|a_{r+1}|$ la perturbación relativa equivalente es por si misma el nivel de β^{-t} y Φ_r, el orden del polinomio no está involucrado. Y ésta es una situación muy común.

Para este algoritmo simple podemos dar un error de análisis muy sastifactorio a-priori. De las relaciónes,

$$b_r = fl(\bar{x}\bar{s}_{r+1})\,\bar{x}\bar{s}_{r+1}(1 + \epsilon_r), \qquad |\epsilon_r| \leq k\beta^{-t}$$

y,

$$\bar{s}_r = fl(b_r + a_{r+1}) = (b_r + a_{r+1}) \div (1 + \Phi_r), \quad |\Phi_r| \leq k\beta^{-t}$$

Tenemos sucesivamente

$$\bar{s}_n \approx a_n$$

$$\bar{s}_n \approx a_n \tag{11}$$

$$\bar{s}_{n-1} = \frac{a_n\,(1+\in_{n-1})\times}{(1+\Phi_{n-1})} + \frac{a_{n-1}}{(1+\Phi_{n-1})} \tag{12}$$

$$\bar{s}_{n-1} = \frac{a_n\,(1+\in_{n-1})(1+\in_{n-2})\times^2}{(1+\Phi_{n-1})(1+\Phi_{n-2})} + \frac{a_{n-1}\,(1+\in_{n-2})}{(1+\Phi_{n-1})(1+\Phi_{n-2})} + \frac{a_{n-2}}{(1+\Phi_{n-2})} + \tag{13}$$

Y se deduce que, (14)

$$s_0 = a_n(1+E_n)\times^n + a_{n-1}(1+E_{n-1})\times^{n-1} + \ldots + a_0(1+E_0)$$

donde

$$1+E_n = \frac{(1+\in_{n-1})(1+\in_{n-2})\ldots(1+\in_0)}{(1+\Phi_{n-1})(1+\Phi_{n-2})\ldots(1+\Phi_0)} \tag{15}$$

$$1+E_r = \frac{(1+\in_{r-1})(1+\in_{r-2})\ldots(1+\in_0)}{(1+\Phi_{r-1})(1+\Phi_{r-2})\ldots(1+\Phi_0)} \tag{16}$$

donde $r = n-1,\ldots 1,0$

De los límites en las ϵ_i y ϕ_i, tenemos.

$$\frac{(1-\alpha)^n}{(1+\alpha)^n} \leq 1 + E_n \leq \frac{(1+\alpha)^n}{(1-\alpha)^n},$$

$$\frac{(1-\alpha)^r}{(1+\alpha)^{r+1}} \leq 1 + E_r \leq \frac{(1+\alpha)^r}{(1-\alpha)^{r-1}} \qquad (17)$$

con $r = n-1,\ldots 0$.

donde $\alpha = k\beta^{-t}$

Igual en una computadora para la cual $k\beta^{-t} = 10^{-7}$ (que es un tanto baja precisión) esto representa un error relativo bajo a no ser que n sea ordinariamente larga. Nótese que tenemos un error relativo pequeño en cada coeficiente indoividual y esto es verdadero, sin embargo muchas cancelaciónes pueden tomar lugar en la computación. Hay una diferencia interesante entre este análisis a-priori y el análisis corrido. En el análisis corrido encontramos expresiones para perturbaciónes en los datos y este conjunto de datos perturbados da todos los $\bar{s}_r$ computados.

En el análisis de error a-priori los datos perturbados dan la $\bar{s}_0$ computado pero diferentes perturbaciónes (en general más pequeño) son requeridos para cada una de las $\bar{s}_r$. Esto es evidente de las relaciónes,

$$\bar{s}_r = x\bar{s}_{r+1} + (a_{r+1} + x\bar{s}_{r+1})\epsilon_r - \bar{s}_r\phi_r$$

$$\bar{s}_n \approx a_n$$

$$\bar{s}_{n-1} = \frac{a_n\,(1 + \epsilon_{n-1}) \times}{(1 + \phi_{n-1})} + \frac{a_{n-1}}{(1 + \phi_{n-1})}$$

Y en sus sucesores.

El análisis muestra que la multiplicación anidada es muy estable inversamente. En el caso cuando las $x\bar{s}_{r+1}$ y $\bar{s}_r$ son comparables con a_r (una situación extremadamente común) el análisis de error corrido muestra que la multiplicación anidada es más ó menos tan estable como posiblemente se espera.

Además cuando los coeficientes no son números digitales, como que las sola representación de ellos en la computadora involucra un redondeo en cada una, el probable efecto de esto es casi tan grande como el efecto de los errores de redondeo cometidos en la computación.

Ahora consideremos los efectos de estas pequeñas perturbaciónes relativas en los coeficientes de los ceros del polinomio. Si asumimos que las perturbaciónes en los coeficientes y en los ceros podemos considerar la perturbación en cada coeficiente por separado. Si γ_i, con $i = 1, \ldots n$, son los ceros del polinomio dado (dando por supuesto que es simple) entonces,

$$\frac{\delta\gamma_i}{\delta a_r} = \frac{-\gamma_i^r}{a_n \prod_{j=/=i} (\gamma_i - \gamma_j)} \qquad (18)$$

Consideremos el polinomio de orden 20 con $\gamma_i = i$, con $i = 1,\ldots,20$ y se verificará que,

$$\frac{\delta\gamma_i}{\delta a_r}$$

es extremadamente grande para algunos valores de i y r. De hecho.

$$\frac{\left|\delta\gamma_{16}\right|}{\left|\delta a_{15}\right|} = 0.4 \times 10^{5}$$

Y dado que $a_{15} = 10^{-10}$, una perturbación en a_{15}, da una perturbación.

$$\delta\gamma_{16} = (0.4 \times 10^{15}) \in$$

De aquí cuando $\in = 0 (\beta^{-t})$ vemos que, a no ser que $\beta^{-t} << 10^{-15}$ la perturbación inducida en γ_{16} no será en ningún sentido pequeña. En una computadora de diez dígitos decimales algúnos de los ceros del polinomio perturbado no llevará relación con esos del polinomio verdadero.

Sí bien los resultados de arriba son triviales sus implicaciones para el análista numérico son totalmente fundamentales los eigen-valores de una matríz de orden n son los ceros del polinomio característico. Es atractivo superficialmente para determinar los eigen-valores. Nuestro análisis muestra que cualquier algoritmo basado en tal técnica esta predispuesto a fracasar como un algoritmo general. La probabilidad de que el polinomio explícito tenga números dígitales en sus coeficientes es pequeño y de aquí los errores cometidos en la representación del polinomio será muchas veces fatal.

Esta simple observación escapó al análista numérico por muchos años y la mayoría de los algoritmos propuestos para resolver el problema de los eigen-valores para matrices no normales estabán basados en el cálculo del polinomio explícito.

C A P I T U L O

XI.

V A L O R E S C A R A C T E R I S T I C O S

D E

M A T R I C E S S I M E T R I C A S

El espacio no permite un tratamiento tan extenso del problema de los valores característicos, como el dado para el problema de las ecuaciónes líneales, podemos solamente mencionar algunos métodos.

Continuaremos resoviendo los nuevos problemas mediante la simplificación de una matríz, convirtiendola en diagonal ó en triangualr superior, pero el paso fundamental ya no será sustraer de una fila el múltiplo de otra. Ya no nos interesará conservar el espacio fila de una matríz, sino perservar sus valores propios. Las operaciones elementales en la filas no lo hacen.

El determinante conduce a una solución formal; a la regla de Cramer en el caso Ax = b y al polinomio det(A - γI) cuyas raíces serán los valores propios. Como siempre, si n = 2 ó 3, puede realmente usarse el determinante para resolver el problema; el cálculo de los valores propios para una n grande es una tarea más larga y difícil que resolver Ax = b, e incluso el mismo Gauss no aportó mucho.

El primer paso es entender que son los valores propios y como pueden ser útiles. Por tanto estamos interersados sólo en aquellos valores particulares γ para los cuales existe algún vector propio x distinto de cero. Para poder ser útil el espacio nulo A - γI debe contener algún vector diferente de cero. Esto es que A - γI debe ser singular.
Para esto el determinante nos da un criterio definitivo.

El número γ será un valor propio de A, con un vector propio coorespondiente distinto de cero, si y sólo si

$$\det(A - \gamma I) = 0$$

esta es la ecuación característica de la matríz A.

En la ecuación Ax = γx, la mayoría de los vectores x no satisfacen esta ecuación, ya sea γ un valor propio ó no. Una x típica cambia de dirección cuando se multiplica por A, así que Ax no es múltiplo de x. Esto significa que sólo ciertos números especiales γ son valores propios y que sólo ciertos vectores especiales son vectores propios. Claro que si A fuera un múltiplo de la matríz identidad, entonces ningún vector cambiaría de dirección y todos los vectores serían vectores propios. Pero en general los vectores

propios son pocos y separados.

Sin embargo la sustancia de este capitulo está en otro lado. La cuestión más importante es explicar cómo se descompone un sistema de ecuaciones al ir hallando los vectores propios. Estos vectores propios son los modos normales del sistema y actúan independientemente. Podemos observar el comportamiento de cada vector propio por separado y después combinar estos modos normales para encontrar la solución. Para decir lo mismo en otras palabras, la matríz subyacente ha sido diagonalizada.

Para concluir resumiremos algunos de los hechos básicos acerca de los valores propios y los vectores propios:
$Ax = \gamma x$. La matríz A de n x n, está dada y el problema es encontrar aquellos vectores especiales x para los cuales A actúa como una multiplicación simple; Ax apunta en la misma deirección que x. La cuestión es encontrar primero los valores propios, reescibiendo γx como γI x y pasando este término a lado izquierdo : $(A - \gamma I) x = 0$. Un vector propio de A es un vector nulo de $A - \gamma I$. En otras palabras, la cuestión clave es la siguiente : si modificamos A por varios múltiplos de la matríz identidad, Qué cambio la hará singular ?. Estas modificaciones revelarán los valores propios y entonces podremos calcular los vectores propios.

Si A es singular, entonces una posibilidad es no modificarla. Uno de los valores propios de una matríz singular es $\gamma = 0$ y el espacio nulo contiene los vectores propios correspondientes. Pero el que A sea singular no significa nada en especial; significa solo que $\gamma = 0$ es un valor propio. Ya sea que A sea singular ó no, todos sus valores propios están basados en lo mismo: la combinación $A - \gamma I$ es singular y el espacio nulo de esa combinación es el espacio de los vectores propios correspondientes a γ.

Para saber cuando $A - \gamma I$ es singular, calculamos su determinante. Este determinante es un polinomio en γ de grado n y se conoce como polinomio característico de A. La ecuación $\det(A - \gamma I)$ es la ecuación característica y sus raíces γ_1 , γ_n (que pueden ser números reales ó no, y

pueden incluir ó no algunas repeticiones de la misma γ) son los valores propios de A. Esto es:

Cada una de las siguientes condiciónes es necesaria y suficiente para que el número γ sea un valor propio de A :

1).- Hay algún vector x diferente de cero tal que $Ax = \gamma x$

2).- La matríz $A - \gamma I$ es singular

3).- $\det.(A - \gamma I) = 0$

Como un ejemplo ilustrativo, concideremos la matríz simétrica.

$$A = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

Su polinomio característico es

$$\det(A - \gamma I) = \begin{bmatrix} 1-\gamma & -1 & 0 \\ -1 & 2-\gamma & -1 \\ 0 & -1 & 1-\gamma \end{bmatrix}$$

$$= - \gamma^3 + 4\gamma^2 - 3\gamma$$

Por lo tanto, la ecuación característica es

$$-\gamma^3 + 4\gamma^2 - 3\gamma = - \gamma (\gamma - 1)(\gamma - 3) = 0$$

y los valores propios son reales y distintos : $\gamma_1 = 0$, $\gamma_2 = 1$, $\gamma_3 = 3$. Para cada valor propio γ_i buscamos un vector propio corespondiente x_i .

Así $\gamma_1 = 0$

$$(A - 0I)\,x_1 = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} x_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

y,

$$x_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$\gamma_2 = 1$:

$$(A - I)\,x_2 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 1 & -1 \\ 0 & -1 & 0 \end{bmatrix} x_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

y,

$$x_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

$\gamma_3 = 3$:

$$(A - 3I)\, x_3 = \begin{bmatrix} -2 & -1 & 0 \\ -1 & -1 & -1 \\ 0 & -1 & -2 \end{bmatrix} x_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

y,

$$x_3 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

Esta cantidad de cálculos es inevitable y desde luego la mayoría de los polinomios cúbicos no se factorizarán tan facilmente como sucedió con este. No cabe duda que algebraicamente y desde el punto de vista de los cálculos, el problema de los valores propios es mucho más difícil que Ax = b. Para un sistema líneal, un número finito de pasos de eliminación produjó la respuesta exacta en un tiempo finito (0 similarmente la regla de Cramer nos dió una fórmula exacta para la solución) en el caso de los valores propios no pueden existir tales pasos ni tal fórmula.

Ahora bien, la suma de los n valores propios es igual a la suma de las n entradas de la diagonal de A:

$$\gamma_1 + \ldots + \gamma_n = a_{11} + \ldots a_{nn}$$

Esta suma es la traza de A. Además, el producto de los n valores propios es igual al determinate de A.

Esto se confirma con el ejemplo anterior, donde la traza de A es 0 + 1 + 3 = 1 + 2 + 1 = 4 y el determinante es 0.1.3 = 0.

No deben confundirse los valores propios de una matríz y sus entradas diagonales. Normalmente son completamente diferentes. Sin embargo, señalaremos una situación donde esos números coinciden.

Si la matríz A es triangular (puede ser superior ó inferior y en particular diagonal) entonces los valores propios $\gamma_1, \ldots, \gamma_n$, son exactamente los mismos que las entradas de la diagonal $a_{11}, \ldots, a_{nn}$

Por ejemplo si.

$$A = \begin{bmatrix} 1 & 1/4 & 0 \\ 0 & 3/4 & 1/2 \\ 0 & 0 & 1/2 \end{bmatrix}$$

entonces su polinomo característico es

$$\det \begin{bmatrix} 1-\gamma & 1/4 & 0 \\ 0 & 3/4-\gamma & 1/2 \\ 0 & 0 & 1/2-\gamma \end{bmatrix}$$

$$= (1 - \gamma) (3/4 - \gamma) (1/2 - \gamma)$$

El determinante es el producto de las entradas diagonales. Obviamente las raíces son $\gamma = 1$, $\gamma = 3/4$ y $\gamma = 1/2$; los valores propios ya estában colocados a lo largo de la diagonal principal.
Hay otra situación en que los cálculos son fáciles. Supongamos que hemos encontrado ya los valores propios de A y tambien los vectores propios. Entonces los valores propios de A son exactamente $\gamma_1^2, \ldots, \gamma_n^2$, y cada vector propio de A es también un vector propio de A^2. La demostración es tipica en matemáticas, si comenzamos con

$Ax = \gamma x$, es obvio. Multiplicando nuevamente por A,

$$A^2 x = A\gamma x = \gamma Ax = \gamma^2 x.$$

Entonces γ^2 es un valor propio de A^2 con el mismo vector propio x. Si la primera multiplicación por A deja intacta la dirección de x, lo mismo hará la segunda.

Este razonamiento no se aplica cuando estén involucradas dos matrices diferentes. Supongamos que γ es un valor propio de A y que γ es un valor propio de B. Entonces, en general, γ no es valor propio de AB. Dado que A y B no comparten el mismo vector propio.

En relación al problema de las ecuaciónes lineales, el cálculo de los valores característicos de matrices se divide en dos casos, de acuerdo a la naturaleza de las matrices. Para matrices escasas ó huecas grandes, los métodos son para mejorar iteraciones infinitas y no serán concideradas aquí. Para matrices densas almacenadas, los métodos son algoritmos finitos. Si una matríz A es simetrica, sus valores característicos estan muy bien determinados por los datos, en efecto, sea la matríz simétrica $B = A + E$, que tiene valores característicos β_i sea A (tambien simétrica) con valores característicos α_i
Entonces los valores característicos pueden ser numérados tal que.

$$\left| \alpha_i - \beta_i \right| \leq \left|\left| E \right|\right|, \text{ para toda } i$$

El anális de error inverso se refiere a los eigen valores cálculados (valores característicos) de una matríz A, al revez para una matríz B = A + E. Sí puede probarce que E es pequeña entonces,

$$\left| \alpha_i - \beta_i \right| \leq \left|\left| E \right|\right|, \text{ para toda } i$$

muestra como son los errores de los eigen valores pequeños. En efecto los métodos actuales pueden llegar a los eigen valores.

que estan en error por solo unos pocos dígitos, en los mismos dígitos significativos de los valores característicos grandes.

El método de Jacobi es una iteración infinita para matrices densas almacenadas, produce una serie de matrices ortogonalmente congruente con A.

$$A_k = U_k^t A_k U_k$$

Además, A_k converge a una matríz diagonal D, cuyas entradas diagonales son, por supuesto, los eigen-valores de A.

En efecto, cada A_{k+1} es calculada de una A_k previa por una rotación de coordenadas en el espacio $2^{(k+1)}$ de algunos 2 índices i y j, una rotación elegida tal que a_{ij}

Para algún k tal que A_k sea casi diagonal, las columnas de la matríz ortogonal correspondiente U_k son aproximadamente los vectores característicos columna de A. Además las columnas son ortogonales entre sí. Así el método de Jacobi da los eigen-vectores aproximados de calidad fina como un producto de la iteración básica. El solo programa es fácil de escribir, y es difícil que se haga mal. Hay algunos problemas teóricos sobre que tan buenos son los eigen-vectores y si el U_k converge actualmente.

El algoritmo original de Jacobi elige i y j para máximizar el valor absoluto de las a_{ij} de A_k. Los algoritmos de hoy en día modifican este criterio en una de dos formas;

i).- En los métodos cíclicos de Jacobi, los elemntos fuera de la diagonal a_{ij}, se hacen cero en algún orden cíclico.

ii).- En los métodos de entrada de Jacobi, un elemento a_{ij} es seleccionado por eliminación solo cuando su valor absoluto está arriba de una cierta medida de entrada, la cual consigue la más pequeña, como la iteración progrese

Ejemplo:

Sea

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad S = \begin{bmatrix} 2 & \\ & 2 \end{bmatrix}, \quad T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$S^{-1} T = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix}$$

Si las componentes de x son v y w, el paso de Jacobi

$$Sx_{k+1} = Tx_k + b \quad \text{es}$$

$$\begin{aligned} 2v_{k+1} &= w_k + b_1 \\ 2w_{k+1} &= v_k + b_2 \end{aligned}$$

ó

$$\begin{bmatrix} v \\ w \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix}_k + \begin{bmatrix} b_1/2 \\ b_2/2 \end{bmatrix}$$

Donde tenemos que la matríz A se descompone como $A = S - T$ entonces la ecuación $Ax = b$ es la misma que $Sx = Tx + b$, por lo tanto obtenemos el paso de Jacobi anteriormente descrito.

Por supuesto que nada garantiza que el método sea bueno y una buena descomposición debe satisfacer dos condiciones:

i).- Debe ser fácil calcular el nuevo vector x_{k+1}. Por lo tanto S debe se una matríz sencilla (! e invertible !); puede ser diagonal ó triangular.

ii).- La sucesión x_k debe converger a la solución x.

Si sustraemos la iteración $Sx_{k+1} = Tx_k + b$, de la ecuación original $Sx = Tx + b$ el resultado es una formula que involucra solamente los errores $e_k = x - x_k$:

$$Se_{k+1} = Te_k$$

Esta es una ecuación en diferencias. Comienza con el error inicial e_0 y después de k pasos produce un nuevo

$$e_k = (S^{-1} T)^k e_0$$

La cuestión de la convergencia es exactamente la misma que la cuestión de la estabilidad : $x_k \longrightarrow x$ precisamente cuando $e_k \longrightarrow 0$

Este método interativo es convergente si y solo si cada valor propio de $S^{-1} T$ satisface $| \gamma | < 1$. Su razón de convergencia depende del tamaño máximo de $| \gamma |$, que se conoce como el radio espectral de $S^{-1} T$:

$$\rho (S^{-1} T) = \text{máx}_i | \gamma_i |$$

Recordando que una solución tipica de $e_{k+1} = S^{-1} Te_k$ es

$$e_k = C_{1-1} x_1^k + \ldots + C_{n-n} x_n^k .$$

Obviamnete el mayor de los $| \gamma_i |$ será finalmente el dominante y gobernará la razón con la caul e_k converge a cero

Volviendo al ejemplo anterior tenemos la matriz decisiva
 -1
S T tiene valores propios ± 1/2, lo que significa que en cada paso el error está partido por la mitad (un dígito binario más es el correcto) En este ejemplo, demasiado pequeño para ser típico, la convergencia es más rápida.

Con una matriz más grande A , existe una dificultad inmediata y muy práctica con la interación de Jacobi expresada como

$$a_{11}(x_1)_{k+1} = (-a_{12}x_2 - a_{13}x_3 - \ldots - a_{1n}x_n)_k + b_1$$
$$\vdots$$
$$a_{nn}(x_n)_{k+1} = (-a_{n1}x_1 - a_{n2}x_2 - \ldots - a_{nn-1}x_{n-1})_k + b_n$$

Necesitamos conservar todas la componentes de x_k hasta terminar el cálculo de x_{k+1}. Una idea más natural, que requiere la mitad del almacenamiento, es comenzar a usar cada componente del nuevo vector x_{k+1} apenas se calcule; x_{k+1} ocupa el lugar de x_k una componente cada vez y, por tanto, podemos destruir x_k tan rápidamente como vamos creando x_{k+1}. Esto significa que la primera ecuación permanece como antes

$$a_{11}(x_1)_{k+1} = (-a_{12}x_2 - a_{13}x_3 - \ldots - a_{1n}x_n)_k + b_1$$

La siguiente ecuación funciona inmediatamente con este nuevo valor de x_1,

$$a_{22}(x_2)_{k+1} = -a_{21}(x_1)_{k+1} + (-a_{23}x_3 - \ldots - a_{2n}x_n)_k + b_2$$

Y la última ecuación usará exclusivamente valores nuevos.

$$a_{nn}(x_n)_{k+1} = (-a_{n1}x_1 - a_{n2}x_2 - \ldots - a_{nn-1}x_{n-1})_{k+1} + b_n$$

Este es el método de Gauss-Seidel, aun cuando aparentemente Gauss lo desconocía y Seidel no lo recomendaba.

Sorprendentemente detalle histórico, ya que no es un mal método. Nótese que cuando se mueven todos los términos x_{k+1} hacia el lado izquierdo, la matriz S es la parte triangular inferior de A. En el lado derecho, la otra matriz de la descomposición, T, es estrictamente triangular superior.

Ejemplo

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix},\ S = \begin{bmatrix} 2 & 0 \\ -1 & 2 \end{bmatrix},\ T = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix},\ S^{-1}T = \begin{bmatrix} 0 & 1/2 \\ 0 & 1/4 \end{bmatrix}$$

Un solo paso de Gauss Seidel transforma las componentes v_k y w_k en

$$2v_{k+1} = w_k + b_1$$

$$2w_{k+1} = v_{k+1} + b_2$$

ó

$$\begin{bmatrix} 2 & 0 \\ -1 & 2 \end{bmatrix} x_{k+1} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x_k + b$$

Nuevamente son decisivos los valores propios de $S^{-1}T$ y desde luego son fáciles de encontrar : son 1/4 y 0. El error está dividido cada vez en 4, así que un solo paso de Gauss-Seidel equivale a dos pasos de Jacobi. * Como ambos métoddos requieren del mismo número de operaciones (usamos el nuevo valor en lugar del anterior y ahorramos en almacenamiento), es mejor el método de Gauss-Seidel.

* Esta regla es válida en multitud de aplicaciones, aun cuando es posible construir otros ejemplos donde Jacobi converja y Gauss-Seidel falle (ó viceverza). El caso simétrico es el más director : si todas las a_{ii} son mayores

que cero, entonces Gauss- Seidel converge si y sólo si A es positivamente definida.

Por otro lado Givens obsevó que, aunque tome una serie infinita de rotaciones lleva a A_k a la forma diagonal y, unas simples rotaciones de $1/2 (n - 1) (n - 2)$ pueden llevar a A_k a la forma tridiagonal. Esto reduce el problema para encontrar los eigen-valores de matrices tridiagonales, y en algun caso la idea de Givens acorta el tiempo práctico de encontrar los valores característicos, por un factor de alrededor de 9 en la práctica. Pocos años despues Hauseholder introduce un nuevo método para tridiagonalizar una matriz simétrica, usando $n - 2$ rotaciones en lugar de $1/2 (n - 1) (n - 2)$ esto reduce el tiempo por otro factor de 2 y efectivamente anula el método de Givens para negocios.

Muchos programas contemporáneos utilizan el método de Hausholder, pero difieren ampliamente en como son encontrados los eigen-valores de matrices tridiagonales. Consiguiendo los eigen-vectores es demásiado complicado, y la falta de conocimiento de como hacerlo es una razón para el uso ocasional continuo de los métodos de Jacobi.

Veamos brevementa la transformación de Hausholder, es una matríz de la forma.

$$H = I - 2 \frac{v v^T}{|| v ||^2}$$

A menudo se normaliza v para obtener un vector unitario $u = v / || v ||$ y entonces H es $I - 2 uu^T$. En ambos casos H es tanto simétrica como ortogonal :

$$H^T H = (I - 2 uu^T) (I - 2 uu^T) =$$

$$= I - 4\, uu^T + 4\, uu^T uu^T = I$$

Así $H = H^T = H^{-1}$. En el caso complejo la matríz correspondiente $I - 2\, uu^T$ es hermitiana y unitaria.

El plan de Hausholder es producir ceros con estas matrices y el buen exito depende de la siguiente identidad.

Supongamos que z es el vector columna $(1,0,\ldots,0)^T$ y

$\sigma = \| x \|$ y $v = x + \sigma z$. Entonces $Hx = -\sigma z = (-\sigma, 0 \ldots 0)^T$

Comenzando con la primera columna de A y recordando que $U^{-1} A U$ final debe ser una forma tridiagonal ó de Hessenberg. Tendremos entonces que estarán involucradas solamente las n - 1 entradas debajo de la diagonal :

$$x = \begin{bmatrix} a_{21} \\ a_{31} \\ \cdot \\ a_{n1} \end{bmatrix} \quad z = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ 0 \end{bmatrix} \quad Hx = \begin{bmatrix} -\sigma \\ 0 \\ \cdot \\ 0 \end{bmatrix} \qquad (a)$$

En este paso la matríz de Hausholder es solamente de orden n - 1, así que está sumegida en la esquina inferior derecha de una matríz U del tamaño original :

$$U_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & & & & \\ 0 & & & & \\ 0 & & & & \end{bmatrix} = U_1^{-1}, \text{ y,}$$

$$U_1^{-1} A U_1 = \begin{bmatrix} a & * & * & * & * \\ \sigma & & & & \\ 0 & & & & \\ 0 & & & & \end{bmatrix}$$

La matriz U_1 no altera la entrada a debido al 1 de la esquina superior izquierda ; más aún, no toca los ceros que aparecen en la expresion (a). Se ha completado la primera parte y $U_1^{-1} A U_1$ tiene la primera columna deseada.

La segunda parte es análoga : x consta de las n - 2 entradas en la segunda columna, z es el vector coordenado unitario de longitud correspondiente y H es de orden n - 2. Cuando se inserta en U_2, produce.

$$U_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & & & \\ 0 & 0 & & & \\ 0 & 0 & & H_2 & \end{bmatrix} = U_2^{-1}$$

$$U_2^{-1}\,(U_1^{-1}\,AU_1)\,U_2 = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{bmatrix}$$

Finalmente U_3 se encargará de la tercera columna y se alcanzará la forma de Hessenberg para una matríz de 5 por 5. En general U es el producto de todas las matrices $U_1\ U_2\ ,\ldots U_{n-2}$ y el número de operaciones requeridas es del orden de n^3.

Ejemplo :

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad x = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad v = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad H = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$$

el resultado es tridiagonal cuando se incerta H en U.

$$U = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix} \qquad U^{-1}\,AU = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$U^{-1}\,AU$ es una matríz lista para obtener los valores propios.

C A P I T U L O

XII.

VALORES CARACTERISTICOS
DE
MATRICES NO SIMETRICAS

Un area de gran actividad en las décadas pasadas en la investigación del álgebra lineal computacional, ha sido el problema del valor característico para matrices no siméticas y solo prevalece un método de antes de la era de la computadora - - el método potencia - - y tiene solo aplicaciones limitadas hoy en dia. Describamos brevemente el método de las potencias, cuyas propiedades de convergencia son fáciles de entender, comienza con una base en el primcipio de una ecuación de diferencias. Comienza con una elección inicial u_0 y forma sucesivamente

$u_1 = Au_0$, $u_2 = Au_1$ y en general $u_{k+1} = Au_k$.

Cada paso es una multiplicación matríz vector y después de k - pasos se produce $u_k = A^k u_0$, a unque la matríz A^k nunca aparecerá. El punto fundamental es que la multiplicación por la matríz A debería ser fácil de realizarse (si la matríz es grande, es mejor que sea rala ó hueca), ya que a menudo la convergencia al vector propio es muy lenta. Suponiendo que A tiene un conjunto completo de vectores propios $x_1, \ldots, x_n$, el vector u_k estará dado por la fórmula usual de una ecuación de diferencias:

$$u_k = c_1 \gamma_1^k x_1 + \ldots + c_n \gamma_n^k x_n .$$

Imaginemos los valores propios numérados en orden creciente y supongamos que hay solo un valor propio mayor; no existe otro valor propio con la misma magnitud y γ_n no se repite.

Así, $|\gamma_1| \leq \ldots \leq |\gamma_{n-1}| < |\gamma_n|$. Entonces, en la medida en que la elección inicial u_0 contenga alguna componente del vector propio x_n , de modo que $c_n \neq 0$, esta componente será domonante gradualmente:

$$\frac{u_k}{\gamma_n^k} = c_1 \left|\frac{\gamma_1}{\gamma_n}\right|^k x_1 + \ldots + c_{n-1}\left|\frac{\gamma_{n-1}}{\gamma_n}\right|^k x_{n-1} + c_n x_n$$

Los vectores u_k apuntan con más y más precisión hacia la dirección de x_n y el factor de convergencia es la razon

$$r = \left|\frac{\gamma_{n-1}}{\gamma_n}\right|.$$

Es precisamente como la convergencia en un estado estacionario que estudiamos para los procesos de Markov, excepto que ahora el valor propio mayor de γ_n puede ser igual a 1. De hecho no conocemos el factor de escala; de otra manera debe introducirse algún factor de escala, pues u_k puede crecer mucho ó poco, dependiendo de si $|\gamma_n| > 1$ ó $|\gamma_n| < 1$. Normalmente podemos dividir cada u_k por su primer componente γ_k antes de realizar el paso siguiente; con esta escala tan simple el método de las potencias se transforma en $u_{k+1} = Au_k / \alpha_k$ y converge a un múltiplo de x_n *

Ejemplo :

U_n teniendo al vector propio

$$\begin{bmatrix} 0.667 \\ 0.333 \end{bmatrix}$$

* los factores de escala α_k tambien convergen; tienden a γ_n

$$A = \begin{bmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{bmatrix}$$ matriz de cambios de población

$$u_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, u_1 = \begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix}, u_2 = \begin{bmatrix} 0.83 \\ 0.17 \end{bmatrix}, u_3 = \begin{bmatrix} 0.781 \\ 0.219 \end{bmatrix}$$

$$u_4 = \begin{bmatrix} 0.747 \\ 0.253 \end{bmatrix}$$

La limitación más seria se ve claramente en este ejemplo : si r está cerca de 1, entonces la convergencia es muy lenta. En muchas aplicaciones $r > 0.9$, lo que significa que se tiene que hacer más de 20 interaciones para reducir $(\gamma_2 / \gamma_1)^k$ solo en un factor de 10. (El ejemplo tenia $r = 0.7$ y fué aún muy lento). Claro que si r es igual a uno significa que $|\gamma_{n-1}| = |\gamma_n|$, entonces no puede haber convergencia. Hay varias maneras de evitar esta límitación y mencionamos tres de ellas.

El método de las potencias por bloques, trabaja simultaneamente con varias vectores en lugar de con uno solo u_k .

El método de las potencias inversas.- trabaja con A^{-1} en lugar de A.

El método de las potencias inversa trasladado.- es el mejor

Supongamos que reemplazamos A por $A - \alpha I$. Entonces todos los valores propios γ_i se trasladan en la misma cantidad α y el factor de convergencia para el método inverso cambiara

a $r'''_i = |\gamma_1 - \alpha_i| / |\gamma_2 - \alpha_i|$. Por lo tanto, si se elige como una buena aproximación a γ_1, r'''_i será muy pequeña y la convergencia se verá muy acelerada.
Así, muchos métodos en uso hoy en día fueron desconocidos por más de 15 años.
Es esencial realizar la inestabilidad inherente al problema de los valores característicos, para matrices no simétricas. En contraste a los límites cerrados,

$$|\alpha_i - \beta_i| \leq ||E|| \quad \text{para toda } i$$

vistos en el capítulo anterior, para matrices no simétricas. En contraste el resultado correspondiente, debido a Ostrowski es ;

$$|\alpha_i - \beta_i| = (\text{polinomio en } n) \times ||E||^{1/n} \quad \text{para toda } i$$

Este resultado es muy débil y no obstante es el mejor resultado general de su clase.

Para la matriz

$$\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \\ & & & & \end{bmatrix}$$

(los blancos denotan ceros)

n tiene todos los valores característicos cero para $\epsilon = 0$, pero todos los valores característicos son distintos y tienen modulos $||\epsilon||^{1/n}$ para $\epsilon \neq 0$. Así, si $n = 40$ y

$\varepsilon = 10^{-40}$, todos los valores característicos tienen modulo 0.1.

Afortunadamente, los valores característicos no son usualmente tan sensibles. De hecho diferentes valores caracteróstticos de una matríz A pueden diferir enormemente en su sensitividad para perturbaciones de A, existen muchos resultados utiles. Son por lo general resultados a-posteriori, dando uno límites para los cambios en valores característicos como funciones de perturbaciones en una matríz e información acerca de otros valores característicos y vectores característicos.

La gran potencia de método de Jacobi para matrices simétricas, y las características extremadamente favorables del redondeo en matrices unitarias, llevaron a un deseo de usarlas para el problema de valores característicos en matrices no simétricas. El siguiente teorema básico es debido a Schur.

TEOREMA

Para una matríz arbitraria $\bar{A}$ existe una matríz unitaria U tal que,

$$T = U^{H} A U \qquad \text{triangular}$$

Donde U^{H} denota la transpuesta conjugada de U

Puesto que los valores característicos de A son los elementos diagonales de T, la espera ha sido para encontrar matrices unitarias las cuales llevan a A a una forma cercanamente triangular, y entonces deja que los elementos diagonales sirvan como valores característicos aproximados de A.

En la mayoría de métodos de atacar el problema de los valores característicos, el primer paso es condensar los datos para ahorrar tiempo y almacenaje adicionales. La forma condensada más aceptada por ahora es la matríz de Hesaenberg, en la cual $a_{ij} = 0$ para $i-j$ 0 (ó su transpuesta) es decir una matríz de la forma.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & 0 & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

Es posible transformar A por congruencias ortogonales del tipo de Hausholder en una forma de Hessenberg, con solo errores de redondeo muy pequeños. Alguna condensación adicional (se dice en una forma tridiagonal) es sujeto de serias perdidas de dígitos. Una transformación a la forma de la matriz acompañante es particularmente desastrosa en la práctica, y normalmente requiere incrementos muy sustanciales en precisión para producir sucesivamente los valores característicos de A.

Como una alternativa, se puede trasnformar A a a su forma de Hessenberg por eliminación gaussiana, con pivoteo parcial o una transformacion similar.

La siguiente etapa en el problema de los eigen-valores no simétricos, es conseguir los valores característicos de una matríz de Hessenberg H. Una variedad de métodos han sido usados. Entre los cuales están.

i).- Se puede examinar los ceros del det(H - zI) por el método de raices encontradas, para z compleja.
El método más satisfactorio parece ser el propuesto por Hyman en un número finito de programas, él hizo uso del método de Laguerre para encontrar los ceros de f(z), seguido por una forma de supreción del cero.

Muchas recurrencias son usadas satisfactoriamente, para evaluar f(z), f'(z), y f"(z), tan usadas en el proceso de Laguerre. Después que los valores característicos $\gamma_1, \ldots, \gamma_r$ han sido encontrados, son suprimidos al aplicar el proceso de Laguerre a.

$$f(z) \Big/ \prod_{i=1}^{r} (\gamma - \gamma_i)$$

ii).- El algoritmo de Ruthishauser, fué un importante desarrollo. Puesto que ahora ha sido suplantado por el algoritmo QR, el cual omitiremos mencionarlo.

iii).- Francis en Inglaterra y Kublanovskaja en la Union Soviética idearon el algoritmo muy interesante QR, y es considerado ampliamente como el algoritmo más satisfactorio para eigen-valores de matrices no-simétricas densas almacenadas.

El teorema básico es que una matríz real cuadrada arbitraria A puede ser factorizada en la forma A = QR, donde Q es una matríz ortogonal y R es una matríz triangular superior con todos los elementos diagonales r_{ij} no negativos. Si A es no-singular, entonces ambas Q y R, son únicas, de hecho, el cálculo esta hecho para construir una matríz Q^T ortogonal tal que $Q^T A = R$, donde R tiene las propiedades mencionadas.

Como una aparte, para A no-singular, el lector estará más familiarizado con la determinación del paso de una matríz triangular superior R^{-1}, tal que AR^{-1}, es una matríz ortogonal Q. Esta es la expresión matricial del proceso familiar del análisis de Gram-Schmidt.

Acaso sorprenderá al lector, que la matríz Q, que resulta del algoritmo de Gram-Schmidt, normalmente está lejos de ser ortogonal, porque por otro lado, si la misma Q está determinada tal que $Q^t A = R$, los errores de redondeo son muy pequeños.

El algoritmo básico QR, procede com sigue.
Sea $H = H_0$ una matríz de Hessenberg, para k = 0,1,2... el

factor H_k en la forma

$$H_k = Q_k R_k$$

y entonces la forma

$$H_k = R_k Q_k$$

Se prueba facilmente que H_{k+1}, es también una matríz de Hessenberg.
Veamos el siguiente teorema básico.

TEOREMA.

Sea H , cuyos eigen-valores $\gamma_1, \ldots \gamma_n$, con.

$$|\gamma_1| < |\gamma_2| < \ldots < |\gamma_n|$$

Entonces las matrices H_k convergen a una matríz triangular superior en la que los elementos de la diagonal son los valores característicos de H.

El caso más común donde el teprema anterior no satisface, es donde encontramos que H converge a una matríz triangular (esto significa que fuera de una matríz triangular todos los elementos de H_k tienden a cero, cuando k tiende a infinito, pero algnos elementos de la forma triangular pueden no converger). Además los grupos diagonales de 2 por 2 y 1 por 1 de la matríz H_k tienen valores característicos, los cuales en su totalidad convergen a los valores característicos de H.

Por simplicidad, concideremos una matríz H con valores característicos,

$$0 < \gamma_1 < \gamma_2 <, \ldots < \gamma_n ;$$

El método QR, para tales matrices converge con un error el cua es

$$0 ((\gamma_1 / \gamma_2)^k)$$

La convergencia podría ser más rápida si, en lugar de H trataramos con la matríz H - pI, donde $0 < p < \gamma_1$, si p fuera practicamente igual a γ_1, la convergencia podría ser extremadamente rápida. Han sido ideadas modificaciones del algoritmo, para el tan llamado " cambio original ", el cual introduce a p cerca de γ. Después que un valor característico γ_1 ha sido aislado, el método QR, puede entonces ser aplicado a una matríz de n - 1, con valores característicos $\gamma_2, \ldots \gamma_n$. Nuevos cambios originales están entonces introducidos para sacar γ_2 tan rapidamente como sea posible etc. Con cambios originales bien ideados, el proceso ha sido observado para converger con un promedio de no menos de 2 pasos interativos para valores característicos.

Algunas investigaciones profundizan en la investigación de cambios originales cuando alguno de los valores característicos son complejos y de modulos iguales, de lo cual no intentaremos dar ideas.

Normalmente los valores característicos son obtenidos para incrementar modulos.

Si H es una matríz simétrica banda, entonces el algoritmo QR preserva la banda de amplitud durante la iteración y es muy satisfactorio. En particular QR es un algoritmo posible para calcular valores característicos de una matríz

simétrica tridiagonal. Si H es una matríz banda no-simétrica, el algoritmo pierde las bandas cero sobre la diagonal.

Aún no hemos mencionado como conseguir los vectores característicos de una matríz de Hessenberg H. Este es el problema más dificil que mencionaremos.

El método prevaleciente es el de la iteración inversa.
Los valores característicos están tomados, siendo conocidos, para algún valor característico γ, se selecciona un vector x_0 de alguna manera. Entonces, se lleva a cabo una iteración de la siguiente forma ;
para cada k = 0, 1,2 . . . encontramos x_{k+1}
para resolver el sistema :

$$(H - \gamma I) x_{k+1} = x_k$$

Se continua hasta que x_k es completamente grande. En algunos casos x_k es facilmente cercano al vector característico correspondiente a γ.
Si H es una matríz de Hessenberg real, γ pero es un valor característico complejo, se tiene que elegir entre uno u otro haciendo aritmética compleja ó algún proceso de selección juicioso con aritmética real.

Finalmente se transforma x_k al revéz del sistema coordenado original de A por deshacer las transformaciones de A a H.

Si alguno de los valores característicos de H son muy cercanos, los problemas reales comienzan. Un par de valores característicos cercanos pueden en casos afortunados tener distintos vectores característicos columna que están muy lejos de ser paralelos ésto representa una aproximación a un valor característico doble con divisores elementalmente lineales. Esta muy lejos como que un par de vectores característicos cercanos tendrán vectores característicos columna, estos también son paralelos. Esto representa una aproximación a los casos infinitamente más probables de un

valor característico doble con divisores no lineales.

En el caso precedente, no es difícil computar dos valores característicos que están lejos de ser paralelos. Es sólo llevar necesariamente la iteración

$$(H - \gamma I) x_{k+1} = x_k$$

con un x_0 diferente ó con valores escasamente diferentes de A_0.

En el último caso parece difícil obtener mucho de la iteración, pero un vector característico simple, perteneciente a A_0. Qué sería dado proximamente ? En parte, no se sabe que el problema propuesto gustaría, pero no se sabe que es posible.

C O N C L U S I O N .

Los métodos de cálculo del álgebra líneal, se están moviendo en una etapa, donde existen métodos razonablmente satisfactorios para matrices densas almacenadas A. La principal excepción es el problema de encontrar vectores característicos con límites de error para matrices no simétricas. Los algoritmos han sido refinados varias veces y ya estan siendo publicados.

Los usuarios casuales del álgebra líneal solo harán uso de tales algoritmos para sus problemas en particular. Los mejores algoritmos están escritos en algol 60. Aún cuando se pueda usar otro lenguaje, vale la pena poder leer este, con el fin justamente de poder leer estos algoritmos y adaptarlos a sus propios problemas.

Ningún método de reslover un problema computacional es realmente aprovechable para el usuario hasta que esté completamente descrito en un lenguaje computacional algebraico y hecho completamente formal. Hay aspectos indeterminados en todo logaritmo. Frecuentemente el provecho entero de un cierto proceso de computación consiste en el tratamiento de ciertas sutilezas, las cuales pueden dificilmente ser suspendidas hasta que están programados. Esta es la razón de porque un novato consultaría a uno u otro experto ó pasa grandes penalidades para retomar un algoritmo totalmente probado. Y esta es la razón de porque los profesionales dificilmente se concretarían en completar pruebas obvias de los algoritmos.

Hoy en día y desde 1979 está disponible un conjunto de programas en Fortran, llamado LINPACK, para resolver varios de los problemas básicos del álgebra líneal numérica. Y tiende a reemplazar gradualmente las miles de subrutinas existentes algunas buenas y otras no, que se ha escrito y utilizado en multitud de medios de computación. El LINPACK se diseñó independientemente de la máquina, eficiente y sencillo. El exito que obtuvo el EISPACK para resolver problemas de valores propios estimuló el desarrollo de estos programas; gracias al EISPACK apreciamos el valor de un software matemático bien escrito.

B I B L I O G R A F I A.

1.-BAUER,F.L.- Optimally Scaled Matrices.-Numerische Mathematik.Vol.8,1963.- Pp.79 - 87.

2.-BURDEN,RICHARD L.& DOUGLAS FAIRES.-Analisis Numérico.- Grupo Editorial Iberoamericano.- 1985 .

3.-FORSYTHE.G.W.L.- Today's Computational Methods of Linear Algebra.- Siam Rev.9.- 1967.

4.-GOLUB,G.H.& J.H.WILKINSON.- Note on the Iterative Refinement of Least Squares Solution.- Numerische Mathematik ,1960 p.139 - 146 .

5.-NOBLE,BEN & DANIEL JAMES F.-Applied Linear Algebra.- Prentice Hall,Inc.-1977.

6.-STAIR,RALPH,JR.-Principles of Data Processing,Concepts Applications and Cases.-University of Florida,1981.

7.-STRANG,GILBERT.-Linear Algebra.-Academic Press,New York 1980.

8.-WILKINSON,J.H.-Rounding Errors in Algebraic Processes London.E.M.S.O.-1963.