UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

FACULTAD DE CIENCIAS

METODOS ITERATIVOS PARA CALCULAR LA INVERSA DE UNA MATRIZ

T E S I S

QUE PARA OBTENER EL TITULO DE:

M A T E M A T I C O

P R E S E N T A:

Enrique Dueñas Blanquel

MEXICO, D. F. 1987

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

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

El objetivo de la tesis es presentar en forma precisa una recopilación de algunos de los métodos iterativos para calcular la inversa de una matriz.

Los prerrequisitos son elementales, basta sólo conocer un poco de Algebra Lineal, Análisis Numérico y tener alguna experiencia en programación, pero independientemente de esto, el segundo capítulo de encargará de dar una visión general de los conceptos que se necesitan.

Inicialmente discutiremos la importancia que tiene el problema en aplicaciones prácticas.

Argumentaremos que este problema matemáticamente encierra en sí una amplia riqueza de conceptos.

Es necesario decir que ya existen métodos para resolver el problema y esta es también una forma diferente de hacerlo que surge como una opción para totalmente mecanizar la resolución del problema.

Estableceremos la teoría necesaria para tener un marco adecuado con el cual manejemos todos los conceptos que se requieren para el desarrollo del tema, como por ejemplo la Estimación de los errores, Normas Vectoriales y Matriciales, Condición y Radio Espectral de una matriz, etc.

Mencionaremos algunas de las dificultades que se tienen al utilizarse este tipo de métodos en cuanto al número de operaciones, exactitud, condición y singularidad de una matriz, etc.

Se estimarán cotas absolutas y relativas para determinar la exactitud de la aproximación obtenida. En esta parte se verá que es básico analizar el concepto de norma de una matriz y lo que se entiende también por el número de condición de una matriz.

En base a lo anterior, desarrollaremos detalladamente los métodos generales, el de Newton y la variante planteada en el artículo de Phipps [12]. Las ajustes los mostraremos en terminos de ; la aceleración de la convergencia del método, pequeñas modificaciones y las correspondientes dificultades de implementación que ocasionan.

Será vital la importancia que tiene para un método la correcta elección de la matriz que será la aproximación inicial a la inversa y es por ello que nos detendremos bastante en este punto para justificar ampliamente el porqué de tal elección, y para el caso de matrices que tienen una estructura especial, determinar una aprovechando sus características.

Después mostraremos los detalles para implementar los métodos numericamente y cuál de ellos es más efectivo al implementarse en la maquina.

Mostraremos también comparaciones numericas con otros algorítmos ya desarrollados para este fin.

Finalmente exibiremos las conclusiones pertinentes obtenidas durante la elaboración del presente trabajo.

I N D I C E

C A P I T U L O I

DISCUSION DE LA IMPORTANCIA DEL PROBLEMA

a) Ejemplos de la necesidad de calcular la inversa de una matriz.

Es claro que todo planteamiento matemático viene dado en terminos de encontrar la solución de un problema y la rápidez con la que la obtengamos dependerá indudablemente de la importancia que tenga dicho problema.

Es asi como nos vemos en la necesidad de resolver un sistema de n ecuaciones con n incógnitas lineal y nos proponemos métodos para encontrar la solución. Nuestras ideas irán en el sentido de plantear métodos y seleccionar el óptimo, es decir, el que sea más fácil analíticamente (hablando teorícamente) ; si el número de incognitas es grande y va a ser resuelto numerícamente, entonces tomar en cuenta cual de ellos es más efectivo, sea por : el número de operaciones, la rápidez con que se obtiene la solución, la estabilidad, la exactitud que se alcanza, etc.

Si planteamos en forma general el problema decimos :

encontrar $X \in \mathbb{R}^{n \times n}$ tal que satisface

$$A\,X = M$$

donde A es una matriz no-singular ($A, M \in \mathbb{R}^{n \times n}$).

Si $M = I$ entonces $X = A^{-1}$, y este es exactamente el problema que queremos resolver en este trabajo.

Debemos decir que para resolver un sistema de ecuaciones no es necesario calcular la inversa de la matriz asociada al problema, ya que un método como la Eliminación Gaussiana es bastante efectivo, pero en otros casos si es importante tener explicítamente la inversa, como por ejemplo calcular la matriz simétrica positiva definida

$$C = (A^{t}\, A)^{-1}$$

o un múltiplo escalar de ella, sea σ C. Esta matriz tiene una interpretación estadística,bajo hipótesis apropiadas, de ser una estima de la matriz de covarianza para el vector solución del Problema de Mínimos Cuadrados [7].

En algunos otros problemas de Análisis de Regresión y de Análisis Multivariado se requiere también calcular en forma indispensable la inversa de una matriz [8].

Así el cálculo de la inversa de una matriz es un problema importante y que por lo tanto debe ser resuelto de la mejor manera posible.

I b) Comentarios Adicionales.

Antes de pasar a materia del tema hagamos algunas consideraciones.

Mencionamos en el inciso anterior que debemos de hacer para resolver un problema , y si este se plantea desde el punto de vista numérico tenemos que resolver además otras cuestiones interesantes.

Afirmamos que existen métodos para resolver tal problema, pero las dificultades prácticas que se tienen con todos estos métodos son :

i) El trabajo que se requiere para ejecutar un número elevado de operaciones.
ii) La indudable pérdida de exactitud en tales cálculos debido alSistema Punto Flotante que se este manejando y que será de un número finito de dígitos.

La primera dificultad viene como consecuencia de la complejidad del algorítmo que hayamos elegido y por supuesto si este requiere bastantes operaciones.

La segunda es consecuencia de trabajar en una computadora que aunque sea lo más efectiva posible , maneja sólo un número finito de dígitos que haran que se pierda exactitud en los cálculos.

Es entonces claro que para determinar que tan bueno es nuestro método para resolver un problema partícular en una maquina dada, debemos esclarecer ciertas preguntas:

1.- ¿ Cuántas operaciones son requeridas para aplicar nuestro método ? .
2.- ¿ Cual sera la exactitud de la solución encontrada por el método elegido ? .
Esto en terminos numericos significa establecer cotas a priori para los resultados que se van a obtener.
3.- ¿ Podemos hacer una estimación a posteriori ?, es decir, ¿ Podemos checar la exactitud de la respuesta obtenida ?.
4.- En el caso de que existan otros métodos, ¿ el elegido ejecuta menos operaciones ? , ¿ Es más exacto ? , ¿ Es más efectivo ? .

Se pueden establecer criterios para evaluarlas, uno de ellos para la primera y cuarta preguntas será llevando un contador de operaciones, o hablando en terminos computacionales cuanto tiempo de procesador requirió.

En relación a nuestro problema, la 2a. , 3a. y 4a. preguntas serán resueltas determinando cotas para los resultados obtenidos y criterios de comparación con los demas métodos o entre ellos mismos, puesto que algunos son el resultado de ajustar un método en la práctica viendo que con una pequeña modificación la convergencia y exactitud del método es mejor.

Recalcamos que todas estas cuestiones deben de tomarse muy en cuenta antes de iniciar un trabajo en estos terminos. Es así como quisimos incluirlo en este momento para reflejar el tipo de problemas a los que nos enfrentamos y también la forma de como llevar el análisis de los métodos estudiados.

C A P I T U L O II

P R E L I M I N A R E S

En este capítulo daremos los resultados más importantes que serán utilizados posteriormente en el análisis de los métodos.

Es importante señalar que muchas de las demostraciones se han omitido debido a que son resultados de dominio común en Algebra Lineal Numérica y que en la mayoria de los libros de esta área y en los que se dan de referencia vienen establecidos. Sin embargo, se demuestran algunos que son dejados como ejercicios ejercicios en algunos de ellos y que son importantes en el desarrollo del trabajo presente.

Los numeros que aparecen entre corchetes [ ] son las libros donde encontramos los resultados mencionados y que vienen incluidos en la bibliografía.

II a. Normas Vectoriales.

Definición. Una Norma Vectorial en $\mathbb{R}^n$ es una función $\| \cdot \| : \mathbb{R}^n \longrightarrow \mathbb{R}^+ \cup \{ 0 \}$ que satisface

a) $x \neq 0 \Rightarrow \| x \| > 0$ y $\| x \| = 0 \iff x = (0,0,\ldots,0)$

b) $\| \alpha x \| = | \alpha | \ \| x \|$

c) $\| x + y \| \leq \| x \| + \| y \|$

Ejemplos de estas normas son:

$$\| x \|_1 = \sum_{i=1}^{n} | x_i |$$

$$\| x \|_2 = \left(\sum_{i=1}^{n} | x_i |^2 \right)^{1/2}$$

$$\| x \|_\infty = \max_i | x_i |$$

Por otra parte, la noción de límite y norma estan relacionadas y nos es útil establecer la noción de límite en términos de una norma.

Teorema II.1 .- Sea $(x_n)_n$ una sucesión de n-vectores y $x \in \mathbb{R}^n$, entonces

$$\lim x_n = x \iff \lim \| x - x_n \| = 0 \quad , \quad [\ 14\ ].$$

II b. Normas Matriciales.

Consideraremos ahora el problema de extender la idea de una norma para las matrices.

Se sabe que el conjunto de matrices $\mathbb{R}^{m\times n}$ es un espacio vectorial que es esencialmente idéntico con $\mathbb{R}^{mn}$. De manera que cualquier norma vectorial en $\mathbb{R}^{m\times n}$ induce una función homogénea positiva definida que satisface la desigualdad del triángulo y por lo tanto es natural llamar a tal función una norma matricial.

Por ejemplo, $\|\cdot\|_2$ en $\mathbb{R}^{mn}$ induce la norma de Frobenius $\|\cdot\|_F$ en $\mathbb{R}^{m\times n}$ definida por

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2}$$

Sin embargo, algunas de las normas matriciales definidas en esta forma no son útiles porque esta definición no da una relación entre la norma de 2 matrices y la norma de su producto. Por lo tanto, restringiremos nuestra atención a normas matriciales que cumplan la condición :

Definición. Una norma matricial es consistente si
$\forall\ A \in \mathbb{R}^{l\times m}$, $B \in \mathbb{R}^{m\times n}$

$$\|AB\| \leq \|A\| \ \|B\| \qquad (**)$$

con $l,m,n = 1, 2, 3, \ldots$

con esta restricción , algunas de las normas vectoriales que dimos al principio no llegan a ser normas matriciales consistentes puesto que por ejemplo la extensión de la $\|\cdot\|_\infty$ es la función ν definida para $A \in \mathbb{R}^{m\times n}$ por

$$\nu(A) = \max\{\, |a_{ij}| \ ; \ i=1,\ldots,m \quad j=1,\ldots,n \,\}$$

la cual es una norma matricial pero para

$$A = B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

no cumple (**).

Sin embargo, hay una forma más geométrica y natural en que puede ser definida la norma de una matriz.
De esta manera, si $x \in \mathbb{R}^n$ y $\|\cdot\|$ es alguna norma vectorial en $\mathbb{R}^n$ entonces $\|x\|$ es la "longitud" de x y por lo tanto $\|Ax\|$ es la longitud de Ax , de donde definimos una norma para A $\|A\|$ por el máximo valor relativo :

$$\| A \| = \sup_{x \neq 0} \frac{\| A x \|}{\| x \|}$$

esta es la llamada norma natural o norma matricial inducida por la norma vectorial $\| \cdot \|$.

Se puede mostrar que la definición anterior es equivalente a

$$\| A \| = \max_{\| y \| = 1} \| A y \|$$

Por la definición de la norma matricial inducida tenemos que

$$\text{Si } x \neq 0, \quad \frac{\| Ax \|}{\| x \|} \leq \sup_{x \neq 0} \frac{\| Ax \|}{\| x \|} = \| A \|$$

$$\Longrightarrow \quad \| Ax \| \leq \| A \| \, \| x \|$$

esto sugiere la siguiente :

Afirmación .- Una norma matricial $\| \cdot \|$ inducida es consistente con su correspondiente norma vectorial, si $\forall A \in \mathbb{R}^{m \times n}$ y $\forall x \in \mathbb{R}^n$

$$\| Ax \| \leq \| A \| \, \| x \|$$

Ejemplos de otras Normas Matriciales :

$$\| A \|_1 = \max_j \sum_{i=1}^{n} | a_{ij} |$$

$$\| A \|_\infty = \max_i \sum_{j=1}^{n} | a_{ij} |$$

Interpretando las dos normas anteriores significa que la $\| \cdot \|_1$ es el máximo de las sumas de los elementos de A por columnas mientras que $\| \cdot \|_\infty$ es el máximo por renglones.

Definición.- Una matriz $A \in \mathbb{R}^{n \times n}$ es Simétrica si

$$A^t = A$$

Definición.- Una matriz $A \in \mathbb{R}^{n \times n}$ es Simétrica Positiva Definida si $\forall\ x \in \mathbb{R}^n,\ x \neq 0$

$$x^t A x > 0$$

Pasaremos ahora a definir lo que significa el polinomio característico asociado a una matriz.

Definición . La función

$$f_A(\lambda) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0$$

es un polinomio de grado n en λ que es llamado el polinomio caracteristico de la matriz A.

Es de notar que las raíces de esta ecuación son los valores especiales de λ para los cuales las ecuaciones simultáneas anteriores poseen soluciones distinta de la trivial.

Más formalmente podemos decir :

Definición.- Sea $A \in \mathbb{R}^{m \times n}$ entonces los eigenvalores de A son los escalares λ para los cuales la ecuación $f_A(\lambda) = 0$ posee soluciones distintas de cero. Los vectores x tal que $A x = \lambda x$ son llamados los eigenvalores de la matrix A.

Con la definición anterior podemos enunciar ahora los siguientes resultados.

Teorema II.2 .- A^t y A tienen los mismos eigenvalores.

Teorema II.3 .- $A^t A$ y $A A^t$ tienen los mismos eigenvalores.

Teorema II.4 .- Si λ_i es eigenvalor de A entonces λ_i^p es eigenvalor de A^p para $p \geq 1$.

Estos tres últimos resultados vienen comentados en [10].

Enseguida daremos una definición de lo que entenderemos por el radio espectral de una matriz.

Definición.- El Radio Espectral de una matriz $A \in \mathbb{R}^{m \times n}$ es

$$\rho (A) = \max_i \{ \lambda_i(A) : \lambda_i \text{ es eigenvalor de } A \}$$

Teorema II.5 .- Si $A \in \mathbb{R}^{m \times n}$ es no-singular entonces $B = A^t A$ es una matriz simétrica positiva definida y por lo tanto todos sus eigenvalores son no-negativos, esto es, si suponemos que son

$$\lambda_1 , \lambda_2 , \ldots , \lambda_n$$

dichos valores propios de B tendremos entonces

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0 \quad , [6].$$

Con este resultado estamos listos para establecer un teorema que nos determina la norma espectral de una matriz.

Teorema II.6 . $$\| A \|_2 = (\rho (A^t A))^{1/2}$$

$$= (\text{máximo eigenvalor de } A^t A)^{1/2} \quad , \quad [6].$$

En partícular si A es simétrica entonces

$$\| A \|_2 = \rho (A)$$

Esta norma y la 1 e ∞ serán las más usuales en el desarrollo del trabajo.

Enseguida daremos relaciones entre normas matriciales.

Teorema II.7 . Para cada par de normas matriciales, sean $\| \cdot \|$ y $\| \cdot \|'$, existen constantes M , m tal que $\forall$ $A \in \mathbb{R}^{n \times n}$ tenemos

$$m \| A \|' \leq \| A \| \leq M \| A \|' \quad , \quad [10].$$

más aún, si $\|\cdot\|$ y $\|\cdot\|'$ son las normas 1 , 2 e ∞ y $A \in \mathbb{R}^{m\times n}$ entonces

$$1/m^{1/2}\,\|A\|_2 \le \|A\|_\infty \le n^{1/2}\,\|A\|_2$$

$$1/n^{1/2}\,\|A\|_2 \le \|A\|_1 \le m^{1/2}\,\|A\|_2$$

$$1/n^{1/2}\,\|A\|_\infty \le \|A\|_1 \le m\,\|A\|_\infty$$

Finalmente, daremos la equivalencia entre límite de una sucesión de matrices en terminos de normas matriciales.

Definición.- Sea $(An)_n \in \mathbb{R}^{m\times n}$ entonces

$$\lim_{k\to\infty} A_k = A \quad <\!=\!=\!=\!=\!> \quad \lim_{k\to\infty} \|A_k - A\| = 0$$

Asi, la norma nos proporciona un dispositivo simple para discutir la convergencia de una sucesión de matrices (ver siguiente sección).

II c. Matrices Convergentes.

Esta sección nos permitirá preparar los conceptos bajo los cuales los métodos que daremos posteriormente serán considerados convergentes.

Definición.- Una matriz cuadrada es convergente si

$$\lim_{k\to\infty} A^k = 0$$

Condiciones de equivalencia estan contenidas en

Teorema II.8 .- Las siguientes 3 proposiciones son equivalentes :

i) A es convergente,

ii) $\lim_{k\to\infty} \| A^k \| = 0$ para alguna norma matricial ,

iii) $\rho (A) < 1$, [6].

Corolario II.9 .- A es convergente si para alguna norma matricial

$$\| A \| < 1 \quad , \quad [6].$$

La condición de convergencia tiene una justificación natural en el caso de que la matriz A sea simétrica ya que por la definición de similaridad de la sección e de este capítulo existe una matriz P invertible y una matriz D diagonal tal que

$$A = P^{-1} D P$$

en otras palabras, A y D son matrices similares y por lo tanto tienen los mismos eigenvalores.

De manera que las potencias de la matriz A quedan en la forma

$$A^n = P D^n P^{-1} \qquad n \in \mathbb{N} \qquad (1)$$

y como la matriz D tiene los eigenvalores de A en su diagonal y si estos son menores que uno entonces las potencias de la matriz D tenderán a parecerse a la matriz nula a condición que la n sea sea suficientemente grande. Este argumento y la relación (1) muestran porque A es una matriz convergente.

Otros resultados que serán importantes puesto que nos dan condiciones bajo las cuales una serie converge estan contenidos en los siguientes :

Teorema II.10 .- La serie geométrica

$$I + A + A^2 + A^3 \ldots$$

converge $\langle ==\blacksquare \rangle$ A es una matriz convergente.

Teorema II.11 .- Si A es convergente , entonces I - A es no-singular y

$$(I - A)^{-1} = I + A + A^2 + A^3 + \ldots$$

Agregaremos enseguida una consecuencia de la condición de consistencia dada anteriormente:

Teorema II.12 .- Si la matriz A es consistente entonces

$$\| A^n \| \leq \| A \|^n$$

Los últimos cinco resultados vienen establecidos en [6].

II d . Descomposición de Matrices en Formas Diagonales.

Daremos antes algunos resultados sobre matrices similares.

Definición .- Sean A,B matrices cuadradas del mismo orden, Entonces diremos que A es Similar a B si existe una matriz P no-singular para la cual

$$B = P^{-1} A P$$

Teorema II.13 .- Si A y B son matrices similares, entonces el polinomio característico de A es igual al polinimio caracteristico de B , es decir ,

$$f_A(\lambda) = f_B(\lambda) \quad , \quad [\ 10\ ].$$

Teorema II.14 .- Los eigenvalores de matrices similares son los mismos y existe una correspondencia uno a uno entre los eigenvectores , [10].
Si $A x = \lambda x$ y $P^{-1}A P = B$, entonces por la definición de eigenvector haciendo

$$z = P^{-1} x$$

tenemos

$$\begin{aligned} B z &= (P^{-1} A P) P^{-1} x \\ &= P^{-1} A x \\ &= P^{-1} \lambda x \\ &= \lambda P^{-1} x \\ &= \lambda z \end{aligned}$$

Esto muestra que un eigenvalor de A es eigenvalor de B con $z = P^{-1}x$ como su eigenvector correspondiente.
Argumentos análogos muestran que un eigenvalor de B es también un eigenvalor de A con la respectiva relación entre los eigenvectores.

Teorema II. 15 . Si $A \in \mathbb{R}^{n \times n}$ es una matriz simétrica entonces existe una matriz ortogonal Q , esto es, $Q\ Q^t = I$ tal que

$$S = Q^t A Q$$

es una matriz diagonal que tiene los eigenvalores de A

$$\lambda_i = \lambda_i(A) \quad , \quad i = 1 , 2 , \ldots , n$$

sobre su diagonal principal , [4].

Por lo tanto ,

$$S = \text{Diag} \{ \lambda_1, \lambda_2, \ldots, \lambda_n \}$$

y si

$$| \lambda_1 | = \max_i | \lambda_i |$$

entonces

$$|| A ||_1 = || S ||_1 = \lambda_1$$

Además ,

$$S^{-1} = \text{Diag} \{ \lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_n^{-1} \}$$

de lo cual se sigue que los valores propios de A^{-1} son

$$\lambda_1^{-1}, \quad \lambda_2^{-1}, \ldots, \quad \lambda_n^{-1}$$

si el rango de A es n y

$$\lambda_n = \min_i | \lambda_i |$$

entonces

$$|| A^{-1} ||_1 = || D^{-1} ||_1 = \lambda_n^{-1}$$

II e . Relaciones entre el Radio Espectral y las Normas Matriciales.

Es indudable el uso que le daremos en los siguientes capítulos al próximo teorema que nos garantizará desigualdades importantes.

Teorema II.16 .- Para las normas 1,2 e ∞ y $A \in \mathbb{R}^{n \times n}$

$$\rho (A) \leq || A ||$$, [6].

Por otra parte, para cada matriz existe una norma natural que es arbitrariamente cercana al radio espectral.

Teorema II.17 .- Para cada matriz A de orden n y $\varepsilon > 0$, existe una norma natural tal que

$$\rho (A) \leq \rho (A) + \varepsilon$$, [6].

II f . Relación Entre los Eigenvalores y la Traza de una Matriz.

Definición.- La traza de una matriz A , que se representa simbolicamente por Tr (A) es la suma de sus elementos de la diagonal , es decir ,si $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ entonces

$$Tr (A) = \sum_{i=1}^{n} a_{ii}$$

Teorema II.18 .- El polinomio característico de una matriz cuadrada de orden n es un polinomio de grado exacto n con el coeficiente principal $(-1)^n$ y el término constante igual al determinante de A . El coeficiente de λ^{n-1} es

$$(- 1)^{n-1} Tr (A)$$

Son n eigenvalores, y si estos son $\lambda_1, \lambda_2 , . . . , \lambda_n$ entonces

$$\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} a_{ii} = Tr (A)$$ y

$$\lambda_1 \lambda_2 . . . \lambda_n = Det A$$, [10].

II g .- Número de Condición de una Matriz.

Para tratar el número de condición de una matriz requiere tener idea de cuando una matriz es bien o mal condicionada.

Entenderemos que un problema es "bien planteado" o un procedimiento numérico es estable si a "pequeños" cambios en los datos ocurren "pequeños" cambios en la solución.

Por lo tanto, si en un proceso numérico tenemos estabilidad, podemos decir que los errores de redondeo que produce la maquina y el proceso en sí mismo serán acotados razonablemente.

Entenderemos que una matriz A se dice ser "bien condicionada" ó "mal condicionada" si los cálculos son o no , respectivamente, bien planteados.

Asi, con este antecedente,podemos decir que el número de condición de una matriz surge por la necesidad de medir el mal condicionamiento de una matriz.

Tomaremos nuevamente como referencia el problema

$$A\,x = b \qquad (2)$$

para obtener de manera natural este número de condición.

Supongamos primero que los datos A y b han sido perturbados, sea por la forma en que la maquina los representó o porque la medición que hicimos de ellos no fue exacta. Entonces en vez de en vez de plantear el problema original tendremos ahora :

$$A\,(x + h) = b + k$$

es decir, A fue modificada por Ah y b por un cierto valor k, que esperamos sean pequeños.

Por lo tanto,

$$Ax + Ah = b + k$$

por (2) $\qquad Ax - b = 0$

$$\Longrightarrow \qquad Ah = k$$

$$\Longrightarrow \qquad h = A^{-1} k$$

$$\Longrightarrow \qquad \| h \| = \| A^{-1} k \| \leq \| A^{-1} \| \, \| k \|$$

entonces

$$\frac{\| h \|}{\| k \|} \leq \| A^{-1} \| \qquad (3)$$

Esto significa que el error relativo resultado de la perturbación esta acotado por la norma de la matriz inversa.

Ahora para determinar el error relativo causado en la solución por el sumando Ah consideramos

$$A x = b$$

$$\Longrightarrow \quad \| b \| = \| Ax \| \leq \| A \| \, \| x \|$$

$$\Longrightarrow \quad \frac{\| b \|}{\| A \|} \leq \| x \|$$

$$\Longrightarrow \quad \frac{1}{\| x \|} \leq \frac{\| A \|}{\| b \|} \qquad (4)$$

usando (3) y (4)

$$\frac{\| h \|}{\| k \|} \; \frac{1}{\| x \|} \leq \frac{\| A^{-1} \| \, \| A \|}{\| b \|}$$

en consecuencia ,

$$\frac{\| h \|}{\| x \|} \leq \| A \| \, \| A^{-1} \| \, \frac{\| k \|}{\| b \|}$$

de manera que el error relativo en la solución esta acotado por el error relativo causado por la perturbación en el lado derecho de la ecuacion multiplicado por el factor

$$\| A \| \, \| A^{-1} \|$$

de lo cual se sigue

Definición.- El número de condición de una matriz es

$$\text{cond} (A) = \| A \| \, \| A^{-1} \|$$

Debe quedar claro que cuando el número de condición cond (A) no es muy grande, el sistema (2) es bien condicionado y alrevés, si es grande el sistema será mal condicionado.

Es de notar que no podemos esperar que cond (A) sea pequeño comparado con la unidad puesto que

$$1 = \| I \| = \| A \, A^{-1} \| \leq \| A \| \, \| A^{-1} \| = \text{cond} (A)$$

II h . Aclaraciones.

Durante el curso del presente trabajo hacemos un uso indistinto de la matriz error inicial en la forma

$$E_o = I - A B_o \qquad \text{ó}$$

$$E_o' = I - B_o A$$

donde B_o denota una primera aproximación a la inversa de la matriz A.

Quisieramos aclarar que esta bien justificado puesto que

$$I - A B_o = A^{-1} (I - A B_o) A$$

en otras palabras , E_o y E_o' son matrices similares , por supuesto supondremos que A es una matriz no singular , de manera que tienen los mismos eigenvalores y en consecuencia su radio espectral es el mismo.

Es interesante hacer notar algunos resultados más acerca de estas matrices. En Stewart (1973) pag. 199 los proponen como ejercicios.

Si definimos nuevamente la matriz error E_o por

$$E_o = I - A B_o$$

donde B_o diremos que es una aproximación derecha a la inversa de la matriz A , podremos decir que

$$\frac{\| A^{-1} - B_o \|}{\| A^{-1} \|} \leq \| E_o \|$$

Esto es consecuencia de

$$B_o = A^{-1} (I - E_o) = A^{-1} - A^{-1} E_o$$

y

$$A^{-1} - B_o = A^{-1} - A^{-1} + A^{-1} E_o = A^{-1} E_o$$

por lo tanto

$$\| A^{-1} - B_o \| = \| A^{-1} E_o \| \leq \| A^{-1} \| \| E_o \|$$

asi que

$$\frac{\| A^{-1} - B_o \|}{\| A^{-1} \|} \leq \| E_o \|$$

De la misma manera podemos decir que si B_o es una inversa izquierda aproximada , es decir ,

$$E_o' = I - B_o A$$

entonces

$$\| I - B_o A \| \leq \text{cond} (A) \| E_o \|$$

El resultado se sigue de

$$A^{-1} (I - A B_o) A = (I - A B_o)$$

$$\Longrightarrow \quad I - B_o A = A (I - A B_o) A^{-1}$$

$$\Longrightarrow \quad \| I - B_o A \| = \| A E_o A^{-1} \|$$

$$\leq \| A \| \| E_o \| \| A^{-1} \|$$

$$= \text{cond} (A) \| E_o \|$$

Podemos concluir en base a los ultimos resultados que si tenemos una buena aproximación a la inversa por la derecha,ello no nos garantiza que dicha matriz sea una buena inversa por la izquierda y viceversa.

En Wilkinson (1963) se menciona que en general los errores para los dos casos se comportan del mismo orden y dan algunos ejemplos y comentarios adicionales al respecto. En Househölder (1964) se hace un análisis pero tomando en cuenta los eigenvalores de las dos matrices error y concluye que la aproximación inversa izquierda tendrá mejores resultados que la derecha. Quiza valga la pena decir aqui que estos resultados dependen bastante de la matriz original y entonces ella determine el comportamiento de la matriz error.

II i . Métodos Directos y Métodos Iterativos.

Puesto que el propósito del trabajo es mostrar algunos métodos iterativos para resolver el problema de obtener la inversa de una matriz , es necesario entonces dar la definición de que es lo que entendemos por un método iterativo. De la misma manera, en el momento de efectuar comparaciones, las realizaremos con los métodos directos que son los más efectivos, de manera que también hace falta definir estos ultimos.

Para encontrar la inversa de una matriz , los métodos numericos se han dividido en dos grupos : métodos directos y métodos iterativos.

Métodos Directos.

Por métodos directos entenderemos métodos que dan la solución de un problema por medio de un número finito de operaciones aritmeticas elementales. El número de cálculos aritméticos necesario para la solución del problema depende sólo de la forma del esquema computacional y del orden de la matriz que define el problema dado. La inexactitud en la solución encontrada ocurre como resultado del inevitable redondeo de los números en el transcurso de los cálculos. Junto con esto, puede ocurrir la desaparición de los digitos significativos al efectuar operaciones como la substracción de dos numeros que difieren poco uno del otro. Esta perdida de dígitos significativos puede ocasionar una importante reducción en la exactitud de los resultados de manera que es a menudo necesario alterar nuestro esquema computacional , o rehacer el proceso con un mayor número de dígitos significativos en los cálculos intermedios.

El método fundamental de este grupo es el método basado en la idea de *eliminación* . El algorítmo de este método , el cual es llamado " La Eliminación Gaussiana " consiste , cuando se aplica a la solución de un sistema lineal no-homogéneo , de una cadena de eliminaciones sucesivas por medio de la cual el sistema dado es transformado en un sistema con una matriz triangular cuya solución no presenta dificultad.

Los métodos directos pueden ser impracticos cuando la matriz que tengamos sea bastante grande o sparse [4].

Métodos Iterativos.

El cálculo de la inversa por un método iterativo es obtenida como el límite de aproximaciones sucesivas calculadas por algún proceso uniforme. La convergencia de estas aproximaciones depende esencialmente de los elementos de la matriz. El radio de convergencia depende también de una correcta elección de la aproximación inicial en que el proceso iterativo se encuentra.

En partícular puede suceder que para algún proceso iterativo exista una matriz para la que el proceso converga lentamente o aún diverja . De este mismo problema pueden padecer los métodos directos aunque en este caso para ellos se tendrá que la solución obtenida no se parezca en nada a la original.

Podemos adelantar diciendo que la inmensa ventaja de los esquemas iterativos consiste en la simplicidad y uniformidad de las operaciones que van a ser efectuadas y por lo tanto en la posibilidad de completamente mecanizar el proceso de cálculo.

C A P I T U L O III

CALCULO DE UNA MATRIZ APROXIMACION INICIAL

III a. Introducción.

Este capítulo prepara los elementos para mostrar cómo calcular una apropiada aproximación inicial a la inversa de una matriz dada. Decimos que la aproximación inicial B_o es apropiada en el sentido de que produzca una matriz $E_o = I - AB_o$ convergente. Esto es importante pues los métodos que daremos en los siguientes capítulos requieren de una matriz error inicial convergente.

III b . Condición de Convergencia.

Para entrar en materia, supongamos que B_o es nuestra aproximación inicial y definimos la matriz E_o como el error obtenido, es decir,

$$E_o = I - AB_o$$

y supongamos que si el método que tenemos necesita que la matriz E_o sea convergente, y por el Teorema 8 de la sección C del capítulo anterior deben de cumplirse cualquiera de las las siguientes 2 condiciones:

i) $\lim_{k\to\infty} \| E_o^k \| = 0$

ii) $\rho (E_o) < 1$

El segundo criterio es el que vamos a utilizar para mostrar que para ciertas elecciones de la matriz aproximación inicial B_o, nos produce una matriz E_o convergente.

La importancia de la convergencia de E_o radica en el hecho de que en todos los métodos planteados el error cometido queda siempre en potencias de esta matriz, razón por la cual es necesario que E_o sea convergente.

III c . Formas de Elegir La Matriz Aproximación Inicial.

En esta sección supondremos que los eigenvalores de la matriz A son λ_1 , λ_2 , ... , λ_n.

Aproximación inicial en la forma Bo = w I , donde w > 0.

Supongamos entonces que nuestra aproximación inicial es de la forma

$$B_o = w\ I$$

donde w > 0.
Para esta elección la matriz error inicial toma la forma

$$E_o = I - w\ A$$

de manera que sus eigenvalores son

$$\mu_i = 1 - w\ \lambda_i$$

Supondremos además que las μ_i son reales.
De manera que si queremos que esta matriz E_o sea convergente necesitamos

$$\rho (E_o) = \max_i \mid 1 - w\ \lambda_i \mid < 1$$

esto se cumple si

$$-1 < 1 - w\ \lambda_i < 1 \qquad \forall\ i$$

$$< \text{===} > \qquad -2 < - w\ \lambda_i < 0 \qquad \forall\ i$$

$$< \text{====} > \qquad 0 < w\ \lambda_i < 2 \qquad \forall\ i$$

$$< \text{====} > \qquad 0 < w < \frac{2}{\lambda_i} \qquad \forall\ i$$

Por lo tanto , E_o será una matriz convergente cuando w cumpla la condición anterior.
Esta relación muestra porque inicialmente pedimos que w fuera mayor que cero. Ya que si w < 0 tendríamos que saber como son los productos $w\ \lambda_i$ para conocer como se comportan los eigenvalores de la matriz E_o , pues por ejemplo si alguno de los λ_j es positivo y w < 0 entonces 1 - $w\ \lambda_j$ será mayor que 1 de manera que E_o no será convergente.

En consecuencia si A tiene eigenvalores positivos y negativos entonces E_o no será una matriz convergente (cuando usamos por supuesto la aproximación inicial $B_o = w\ I$).

Pedimos inicialmente que $\mu_i \in \mathbb{R}$ $i=1, 2, \ldots, n$, si no ocurriera así, el análisis de las condiciones bajo las cuales E_o es convergente sería más complicado.

La matriz error inicial Eo es una matriz Simétrica.

Por lo que vimos anteriormente nos convendría tomar E_o tal que fuera simétrica,ya que de esta manera garantizamos que sus eigenvalores son reales. Aprovechando esta condición, Issacson y Keller [6], sugieren la elección

$$B_o = a\ A^t \qquad \text{con} \qquad a = 1 / Tr\ (\ A\ A^t)$$

En efecto, con esta aproximación inicial la matriz E_o es convergente.

Para mostrarlo, haremos algunas observaciones :

1) E_o es una matriz simétrica,

$$\begin{aligned} (\ E_o\ )^t &= (\ I - A\ B_o\ )^t \\ &= (\ I - A\ a\ A^t\ )^t \\ &= I^t - a\ (\ A\ A^t)^t \\ &= I - a\ A\ A^t \\ &= I - A\ B_o \\ &= E_o \end{aligned}$$

de manera que

$$|| E_o ||_2 = \rho\ (\ E_o\ )$$

2) Si λ es un eigenvalor de A , entonces λ^2 es un eigenvalor de la matriz simétrica $A\ A^t$, en consecuencia

$$\mu = 1 - a\ \lambda^2$$

es eigenvalor de

$$E_o = I - a\ A\ A^t$$

puesto que si x es eigenvector de $A A^t$ y por lo tanto $x \neq 0$, tenemos

$$\begin{aligned}(I - a A A^t)x &= Ix - a(A A^t)x \\ &= x - a(\lambda^2 x) \\ &= (1 - a\lambda^2)x \\ &= \mu x.\end{aligned}$$

Con estas dos observaciones podemos ahora decir que si queremos que

$$\rho(E_o) < 1$$

entonces debera ocurrir que

$$\max_i \; |1 - a\lambda_i^2| < 1$$

es decir,

$$\max_i \; |1 - 1/Tr(A A^t)\lambda_i^2| < 1$$

pero en el capítulo anterior mencionamos que

$$Tr(A) = \sum_{i=1}^{n} a_{ii} = \sum_{j=1}^{n} \lambda_j$$

donde los λ_j representan los eigenvalores de A, asi que para nuestro caso:

$$Tr(A A^t) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}a_{ji} = \sum_{j=1}^{n} \lambda_j^2$$

lo que implica que deseamos

$$\max_i \left| 1 - \frac{\lambda_i^2}{\sum_{j=1}^{n} \lambda_j^2} \right| < 1$$

y es claro que

asi,

$$0 < \frac{\lambda_i^2}{\sum_{j=1}^{n} \lambda_j^2} < 1 \qquad \forall\ i$$

$$\Longrightarrow \quad -1 < -\frac{\lambda_i^2}{\sum_{j=1}^{n} \lambda_j^2} < 0 \qquad \forall\ i$$

$$\Longrightarrow \quad 0 < 1 - \frac{\lambda_i^2}{\sum_{j=1}^{n} \lambda_j^2} < 1 \qquad \forall\ i$$

por lo tanto,

$$\max_i \left| 1 - \frac{\lambda_i^2}{\sum_{j=1}^{n} \lambda_j^2} \right| < 1$$

de manera que

$$\rho (E_o) < 1$$

en consecuencia, E_o es una matriz convergente.

Los pasos que seguimos muestran claramente cual es la idea de elegir asi la matriz aproximación inicial.

Primero, B_o asi escogida hace que E_o sea una matriz simétrica.

Y segundo, era conveniente que

$$\frac{\lambda_i^2}{\sum_{j=1}^{n} \lambda_j^2} < 1 \qquad \forall\ i$$

lo cual efectivamente ocurre , de manera que el resultado queda establecido.

En general , el resultado sería cierto para toda α tal que

$$\max_i \left| 1 - \alpha\ \lambda_i^2 \right| < 1$$

y como E_o es simétrica (suponiendo por supuesto que $B_o = \alpha A^t$) entonces

$$\text{máx} \mid 1 - \alpha \lambda_i^2 \mid < 1$$

$$< \text{====} > \quad -1 < 1 - \alpha \lambda_i^2 < 1 \qquad \forall\, i$$

$$< \text{====} > \quad -1 + \alpha \lambda_i^2 < 1 < 1 + \alpha \lambda_i^2 \qquad \forall\, i$$

$$< \text{====} > \quad 0 < \alpha \lambda_i^2 < 2 \qquad \forall\, i$$

En consecuencia , E_o será también convergente para las α que sean de la forma

$$0 < \alpha < \frac{2}{\lambda_i^2} \qquad \forall\, i$$

Una cuestión importante ahora y que será necesario considerar es el de elegir α tal que sea fácil de calcular operacional y computacionalmente.

Aproximaciones iniciales de Pan y Reif.

Pan y Reif [11] proponen la siguiente aproximación inicial

$$B_o = t\, A^t$$

considerando 2 valores para t que veremos enseguida. En ambos casos $E_o = I - A\, B_o$ es una matriz convergente.

Toman como primer caso

i)
$$t = \frac{1}{\| A \|_1 \, \| A^t \|_1}$$

de manera que con esta elección tendremos

$$\rho (E_o) < 1$$

$$< \text{====} > \quad \text{máx}_i \left| 1 - \frac{\lambda_i^2}{\| A \|_1 \, \| A^t \|_1} \right| < 1$$

pero por definición del radio espectral

$$\lambda_i^2 \leq \rho (A A^t) \qquad \forall \ i$$

además , por ser $\| \ \|_1$ una norma natural

$$\rho (A A^t) \leq \| A A^t \|_1$$

y por cumplir la condición de consistencia

$$\| A A^t \|_1 \leq \| A \|_1 \| A^t \|_1$$

tenemos

$$0 \leq \lambda_i^2 \leq \rho (A A^t) \leq \| A A^t \|_1 \leq \| A \|_1 \| A^t \|_1 \quad \forall \ i$$

así,

$$0 \leq \frac{\lambda_i^2}{\| A \|_1 \ \| A^t \|_1} < 1$$

de lo cual concluimos que

$$\rho (E_o) < 1$$

por lo tanto E_o es convergente.

Ahora para cuando t es

ii)
$$t = \frac{1}{\| A A^t \|_1}$$

los argumentos son análogos y se llega a

$$\max_i \left| 1 - \frac{\lambda_i^2}{\| A A^t \|_1} \right| < 1$$

ya que

$$0 \leq \lambda_i^2 \leq \rho (A A^t) \leq \| A A^t \|_1$$

por lo tanto

$$0 \leq \frac{\lambda_i^2}{\| A A^t \|_1} \leq 1$$

de lo cual se sigue que nuevamente E_o es convergente.
Podemos todavia comparar estas 2 ultimas elecciones, dado que

$$\frac{\lambda_i^2}{\| A \|_1 \| A^t \|_1} \leq \frac{\lambda_i^2}{\| A A^t \|_1} \quad \forall \ i$$

asi ,

$$1 - \frac{\lambda_i^2}{\| A \|_1 \| A^t \|_1} \leq 1 - \frac{\lambda_i^2}{\| A A^t \|_1} \quad \forall \ i$$

en consecuencia ,

$$\max_i \left| 1 - \frac{\lambda_i^2}{\| A \|_1 \| A^t \|_1} \right| \leq \max_i \left| 1 - \frac{\lambda_i^2}{\| A A^t \|_1} \right|$$

por lo tanto,la primera elección de Pan y Reif es mejor que la segunda puesto que provocará que el radio de convergencia de la matriz aproximación inicial , o equivalentemente su norma, sea menor y en consecuencia el error cometido se reduce.

Comparaciones.

Debido a que en las últimas dos elecciones una de ellas resultó mejor que la otra , optaríamos por tomar la que obtiene mejores resultados ,pero debemos tener un poco de cuidado pues al momento de implementar numericamente esta elección podría ser más costosa . Por esta razón daremos un cuadro comparando el radio espectral de la matriz error considerando las tres elecciones anteriores y también el número de operaciones que requiere cada una de ellas. Entenderemos que una operación en la maquina representa una multiplicación o una división mientras que las sumas y las comparaciones entre los elementos de la matriz no son tomadas en cuenta por ser faciles de evaluar. Por ejemplo , sabemos que las entradas resultantes de una multiplicación de matrices $C = A B$ son de la forma

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Entonces cada c_{ij} representa n productos y n sumas de manera que en total son n operaciones para cada c_{ij}, de lo cual se sigue que para calcular todas las entradas de la matriz C nos llevará n^3 operaciones.

Las matrices que utilizamos fueron con entradas aleatorias y tenían una dimensión de 5, 10, 15, 20 y 30.
El cálculo fue hecho con el paquete MATLAB (1983) en una computadora compatible con la PC - XT de IBM marca Printaform modelo 5220 con procesador 8086.

VALOR INICIAL	N				
	5	10	15	20	30
$a = \frac{1}{Tr(A A^t)}$	.9895115	.9977254	.9998900	.9999982	.99999971
$t = \frac{1}{\|A\|_1 \|A^t\|_1}$	.9919945	.9999661	.9999130	.9999984	.99999974
$t' = \frac{1}{\|A A^t\|_1}$	.9901451	.9999606	.9998868	.9999980	.9999970

Debemos mencionar que estos números son muy cercanos a uno, es decir , estamos en el límite del radio en el que convergen las matrices, y esto veremos despues que provocará que la convergencia de los métodos planteados en los siguientes capítulos sea lenta. También podría suceder que si $\rho(E_o)$ es muy parecido a uno, entonces la matriz error no sea convergente debido a los errores de redondeo y entonces el proceso pueda diverger o converger lentamente.

Enseguida mostramos el número de operaciones que requiere cada una de las 3 aproximaciones iniciales.

VALOR INICIAL	TOTAL DE OPERACIONES
$a = 1 / Tr(A A^t)$	$n^2 + n - 1$
$t = 1/\|\| A \|\|_1 \quad \|\| A^t \|\|_1$	$2n^2 + 2n + 1$
$t' = 1 / \|\| A A^t \|\|_1$	$2n^3 + n^2 + n$

Viendo la última tabla podemos deducir que en terminos del número de operaciones, el primer valor es el más adecuado.

Analizando el último valor obtenido podemos notar que el sumando n^3 provocará que cuando la dimensión de la matriz sea grande resulte costoso tomarlo en cuenta para que sea factor de nuestra aproximación inicial.

Para enriquecer un poco más la discusión daremos otra elección también sugerida por Pan y Reif que a pesar de que reduce el radio espectral de la matriz error aumenta significativamente el número de operaciones.

Supongamos que $\lambda_1 , \lambda_2 , \ldots , \lambda_n$ son los eigenvalores de la matriz simétrica $A A^t$ de manera que

$$\lambda_1 \geq \lambda_2 \geq . \; . \; . \geq \lambda_n > 0$$

por lo tanto

$$\lambda_1 = \|\| A \|\|_2^2 \qquad y \qquad \lambda_n = \frac{1}{\|\| A^{-1} \|\|_2^2}$$

haciendo

$$B = t^* A^t \qquad con \qquad t^* = \frac{2}{\lambda_1 + \lambda_n}$$

Afirmamos que esta elección mejora las 2 aproximaciones de Pan y Reif dadas inicialmente y también la propuesta por Keller e Isaacson. Veamos, como

$$-\lambda_1 \leq -\lambda_2 \leq \ . \ . \ . \leq -\lambda_n < 0$$

$$\Longrightarrow \quad -\lambda_1 t^* \leq -\lambda_2 t^* \leq \ . \ . \ . \leq -\lambda_n t^* < 0$$

$$\Longrightarrow \quad 1 - \lambda_1 t^* \leq 1 - \lambda_2 t^* \leq \ . \ . \ . \leq 1 - \lambda_n t^* < 1$$

Esto muestra inicialmente que

$$\rho (E_o) = \rho (I - A\ t^* A^t) < 1$$

y por lo tanto E_o es convergente.
Mostremos como queda explícitamente su radio espectral.

$$\begin{aligned}
\rho (E_o) &= \rho (I - A\ t^*\ A^t) \\
&= \max_i \ | \ 1 - \lambda_i \ t^* \ | \\
&= | \ 1 - \lambda_n \ t^* \ | \\
&= \left| \ 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_n} \ \right| \\
&= \left| \ \frac{\lambda_1 + \lambda_n - 2\lambda_n}{\lambda_1 + \lambda_n} \ \right| \\
&= \left| \ \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \ \right| \\
&= \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}
\end{aligned}$$

notemos que aqui es bastante claro que $\rho (E_o) < 1$ pues

$$\lambda_1 \geq \lambda_n$$

$$\Longrightarrow \quad \lambda_1 - \lambda_n \geq 0$$

y

$$\lambda_1 + \lambda_n \geq \lambda_1 - \lambda_n$$

entonces

$$1 > \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \geq 0$$

Podemos también mostrar que $\rho (E_o)$ tiene una expresión bastante clara en terminos del número de condición de la matriz. Así,

$$\rho (E_o) = \| E_o \|_2 = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}$$

$$= \frac{\frac{\lambda_1}{\lambda_n} - 1}{\frac{\lambda_1}{\lambda_n} + 1}$$

$$= \frac{\frac{\| A \|_2^2}{\frac{1}{\| A^{-1} \|_2^2}} - 1}{\frac{\| A \|_2^2}{\frac{1}{\| A^{-1} \|_2^2}} + 1}$$

$$= \frac{\| A \|_2^2 \| A^{-1} \|_2^2 - 1}{\| A \|_2^2 \| A^{-1} \|_2^2 + 1}$$

$$= \frac{cond (A)^2 - 1}{cond (A)^2 + 1}$$

O también escribirlo en la forma

$$\rho (E_o) = \frac{cond (A)^2 - 1}{cond (A)^2 + 1}$$

$$= \frac{cond (A)^2 + 1 - 2}{cond (A)^2 + 1}$$

$$= 1 - \frac{2}{cond (A)^2 + 1}$$

Vale la pena comentar en relación a las ultimas expresiones de $\rho (E_o)$, que cuando la matriz A sea mal condicionada, el radio espectral de la matriz error será muy cercano a uno y entonces estaremos en el límite de convergencia para la matriz E_o y en caso contrario si por ejemplo nuestra matriz A se comporta como una rotación y conserva distancias, su número de condición será cercano a uno y por lo tanto $\rho (E_o)$ será muy cercano a cero lo cual optimiza la elección.

Es necesario mencionar cuán costoso es calcular el valor de t^* pues para conocerlo debemos tener el eigenvalor más grande y el más chico o equivalentemente la norma 2 de A y de A^{-1}.

Ahora falta mostrar que efectivamente esta elección mejora las 3 anteriormente mencionadas.

Primero veamos que sucede cuando se elige el factor t en la forma

$$t = \frac{1}{\| A \|_1 \| A^t \|_1}$$

para esta elección es fácil ver que

$$\rho (E_0) = \left| 1 - \frac{\lambda_n}{\| A \|_1 \| A^t \|_1} \right|$$

$$= 1 - \frac{\lambda_n}{\| A \|_1 \| A^t \|_1}$$

Entonces nos gustaría que sucediera

$$\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \leq 1 - \frac{\lambda_n}{\| A \|_1 \| A^t \|_1} = \frac{\| A \|_1 \| A^t \|_1 - \lambda_n}{\| A \|_1 \| A^t \|_1}$$

$$< \Longleftrightarrow > (\| A \|_1 \| A^t \|_1)(\lambda_1 - \lambda_n) \leq (\| A \|_1 \| A^t \|_1 - \lambda_n)(\lambda_1 + \lambda_n)$$

$$< \Longleftrightarrow > \| A \|_1 \| A^t \|_1 \lambda_1 - \| A \|_1 \| A^t \|_1 \lambda_n \leq$$

$$\| A \|_1 \| A^t \|_1 \lambda_1 + \| A \|_1 \| A^t \|_1 \lambda_n - \lambda_n \lambda_1 - \lambda_n^2$$

$$< \Longleftrightarrow > 0 \leq 2 \| A \|_1 \| A^t \|_1 \lambda_n - \lambda_n \lambda_1 - \lambda_n^2$$

$$< \Longleftrightarrow > \lambda_1 + \lambda_n \leq 2 \| A \|_1 \| A^t \|_1$$

lo cual efectivamente es cierto puesto que

$$\lambda_i \leq \rho (A) \leq \| A \|_1 \qquad \forall\, i$$

Para el otro valor que proponen Pan y Reif

$$t = \frac{1}{\| A\, A^t \|_1}$$

mostramos antes que

$$\max_i \left| 1 - \frac{\lambda_i^2}{\| A \|_1 \| A^t \|_1} \right| \leq \max_i \left| 1 - \frac{\lambda_i^2}{\| A\, A \|_1} \right|$$

de manera que por transitividad obtenemos también la mejora establecida.

Falta ver que con la elección

$$t = \frac{1}{Tr(A)} = \frac{1}{\sum_{i=1}^{n} \lambda_i}$$

también mejoramos la aproximación inicial. Entonces

$$1 - \frac{2\lambda_n}{\lambda_1 + \lambda_n} \leq 1 - \frac{\lambda_n}{\sum_{i=1}^{n} \lambda_i}$$

$$< \text{-----} > \qquad \frac{\lambda_n}{\sum_{i=1}^{n} \lambda_i} \leq \frac{2\lambda_n}{\lambda_1 + \lambda_n}$$

$$< \text{------} > \qquad \lambda_1 + \lambda_n \leq 2 \sum_{i=1}^{n} \lambda_i$$

y por supuesto que esto también se cumple.

Hay un comentario adicional, esta mejora es limitada aún teniendo de antemano λ_1 y λ_n (Pan y Reif [11]).

Notemos que esta mejora es para cualquier matriz A , pues solo usamos la propiedad de que A A^t fuera simétrica positiva definida.

Comparaciones Numericas.

Hicimos la comparación cuando el valor de t es

$$t = \frac{1}{\| A \|_1 \| A^t \|_1}$$

contra la mejora

$$t^* = \frac{2}{\lambda_1 + \lambda_n} .$$

Se usaron una matriz aleatoria de 5 x 5 , una matriz de rotación en la forma

$$A = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

después una matriz mal condicionada

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1+10^{-p} \end{bmatrix}$$

y finalmente una matriz ortogonal de 10 x 10.
Los resultados se muestran en la siguiente tabla.

Matriz	Orden	Número de Condición	$\rho(E_o)$ cuando $B_o = \frac{2}{\lambda_1 + \lambda_n} A^t$	$\rho(E_o)$ cuando $B_o = \frac{1}{\|A\|_1 \|A^t\|_1} A^t$
Aleatoria	5	9.02430561556	0.99990352767	0.999966119310
Rotación	3	1.0	2.444062 D-16	0.5
Mal Condicionada	2	3.999987 D+10	1.0	1.0
Ortogonal	10	1.0	1.045687 D-15	8.7668016 D-01

Concluimos esta sección diciendo que realmente si obtuvimos una mejora en ciertos casos pero recalcamos que esto es a costa de calcular previamente λ_1 y λ_n lo cual en general es caro de obtener.

III d . Minimización del Radio Espectral de la Matriz Error.

Para enriquecer un poco más la discusión daremos algunas consideraciones más acerca del problema de reducir el radio espectral de la matriz error.

Nuestro interés es ahora en el sentido de obtener que este número sea por supuesto menor que la unidad pero lo más pequeño posible,es decir,queremos que ρ (E_0) sea mínimo para cierta elección de B_0, donde B_0 es un múltiplo de A^t. Esto se traducirá por lo tanto en que obtendremos la mejor aproximación inicial a la inversa de la matriz dada.

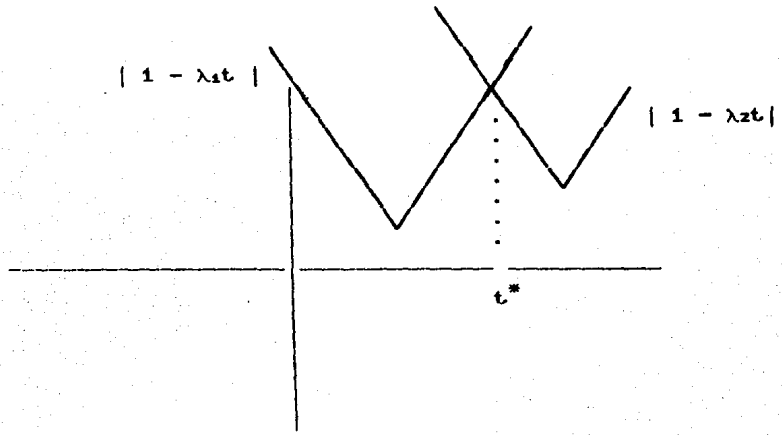
Consideremos el caso cuando la matriz es de 2 x 2 y por supuesto que sea no-singular. En tales condiciones nuestro problema es

$$\text{mín} \quad \left\{ \begin{array}{c} | \, 1 - \lambda_1 \, t \, | \\ | \, 1 - \lambda_n \, t \, | \end{array} \right\}$$

donde λ_1 y λ_2 son los eigenvalores de la matriz A. Supongamos además que λ_1 y λ_2 son reales.

Como λ_1 y λ_2 son fijos para una matriz dada entonces el problema de mínimizar queda en términos de t, que es el factor por el cual multiplicamos a la transpuesta de la matriz original para obtener la primera aproximación a la inversa.

Visto en el plano el problema queda



los puntos que satisfacen nuestros requerimientos son los comunes de las dos funciones y es claro que el punto t^* minimiza nuestro problema.

Para determinar el valor de t^*, el cual debe cumplir que

$$| 1 - \lambda_1 t | = | 1 - \lambda_2 t |$$

por lo tanto, dicho valor deberá satisfacer

$$\begin{aligned} (1 - \lambda_1 t) &= (1 - \lambda_2 t) \quad \text{ó} \\ -(1 - \lambda_1 t) &= (1 - \lambda_2 t) \quad \text{ó} \\ (1 - \lambda_1 t) &= -(1 - \lambda_2 t) \quad \text{ó} \\ -(1 - \lambda_1 t) &= -(1 - \lambda_2 t) . \end{aligned}$$

que se reducen a los casos

$$\begin{aligned} (1 - \lambda_1 t) &= (1 - \lambda_2 t) \quad \text{y} \\ -(1 - \lambda_1 t) &= (1 - \lambda_2 t) . \end{aligned}$$

Analizando el primero de ellos tenemos

$$\begin{aligned} & 1 - \lambda_1 t = 1 - \lambda_2 t \\ <\!\!=\!\!=\!\!=\!\!=\!\!> \quad & \lambda_1 t = \lambda_2 t \\ <\!\!=\!\!=\!\!=\!\!=\!\!> \quad & t (\lambda_1 - \lambda_2) = 0 \end{aligned}$$

si ocurre que $\lambda_1 = \lambda_2$ entonces es el caso trivial , pues tendremos que la matriz A dada es un múltiplo de la idéntica; si $t = 0$, nos conduce a que la aproximación inicial es la matriz nula $B_o \cong 0$ lo cual no tiene sentido.

Asi que el caso interesante es cuando

$$\begin{aligned} & 1 - \lambda_1 t = -1 + \lambda_2 t \\ <\!\!=\!\!=\!\!=\!\!=\!\!> \quad & t (\lambda_1 + \lambda_n) = 2 \\ <\!\!=\!\!=\!\!=\!\!=\!\!> \quad & t = \frac{2}{\lambda_1 + \lambda_n} = t^* \end{aligned}$$

asi que este valor mínimiza el problema.

Generalizando esta idea , tendremos para $A \in \mathbb{R}^{n \times n}$

$$\min \begin{matrix} | 1 - \lambda_1 t | \\ | 1 - \lambda_2 t | \\ \vdots \\ | 1 - \lambda_n t | \end{matrix}$$

el cual visto en $\mathbb{R}^n$ sera un conjunto de n-hiperplanos que contendran el punto óptimo t^* que buscamos el cual ahora deberá cumplir

$$| 1 - \lambda_1 t | = | 1 - \lambda_2 t | = \ldots = | 1 - \lambda_n t |$$

pero estas igualdades sólo las podemos manejar en pares, lo cual nos lleva entonces a considerar todas las combinaciones por pares de la igualdad anterior, de manera que primero debemos encontrar

$$\text{máx} \left\{ \left| 1 - \frac{2\lambda_1}{\lambda_1 + \lambda_2} \right|, \left| 1 - \frac{2\lambda_2}{\lambda_1 + \lambda_2} \right|, \ldots, \left| 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_2} \right| \right\}$$

$$\vdots$$

$$\text{máx} \left\{ \left| 1 - \frac{2\lambda_1}{\lambda_1 + \lambda_n} \right|, \left| 1 - \frac{2\lambda_2}{\lambda_1 + \lambda_n} \right|, \ldots, \left| 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_n} \right| \right\}$$

$$\vdots$$

$$\text{máx} \left\{ \left| 1 - \frac{2\lambda_1}{\lambda_{n-1}+\lambda_n} \right|, \left| 1 - \frac{2\lambda_2}{\lambda_{n-1}+\lambda_n} \right|, \ldots, \left| 1 - \frac{2\lambda_n}{\lambda_{n-1}+\lambda_n} \right| \right\}$$

y despues de tener este máximo por cada conjunto tomarnos el mínimo de ellos y tendremos resuelto el problema.

Podemos decir quien es este mínimo , puesto que los eigenvalores de $A A^t$ los podemos suponer

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$$

Entonces

$$\lambda_i + \lambda_j \leq \lambda_1 + \lambda_2 \qquad \forall\ i,j = 1, \ldots, n$$

y

$$\lambda_n \leq \lambda_k \qquad \forall\ k = 1, \ldots, n$$

$$\Rightarrow \quad 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_2} \leq 1 - \frac{2\lambda_k}{\lambda_i + \lambda_j} \qquad \forall\ i,j,k$$

puesto que los máximos son en cada conjunto

$$1 - \frac{2\lambda_n}{\lambda_1 + \lambda_2}\ ,\ 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_3}\ ,\ 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_4}\ ,\ \ldots\ ,\ 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_n}$$

$$1 - \frac{2\lambda_n}{\lambda_2 + \lambda_3}\ ,\ \ldots\ ,\ 1 - \frac{2\lambda_n}{\lambda_2 + \lambda_n}\ ,\ \ldots\ \ldots\ ,\ 1 - \frac{2\lambda_n}{\lambda_{n-1} + \lambda_n}$$

De manera que el valor que minimiza el problema es

$$1 - \frac{2\lambda_n}{\lambda_1 + \lambda_2}$$

Nuevamente hacemos hincapie en lo costoso que es obtener este valor, ya que envuelve explicítamente el problema de calcular algunos de los eigenvalores de $A A^t$.

Bajo este mismo orden de ideas podríamos preguntarnos también cuando

$$\rho (E_o) = 0$$

$$< ===== > \max_i \ | 1 - \lambda_i t | = 0 \qquad \text{con } t > 0$$

$$< ===== > \quad 1 - \lambda_i t = 0 \qquad \forall\, i$$

$$< ===== > \quad t = 1/\lambda_i \qquad \forall\, i$$

$$< ===== > \quad \lambda_1 = \lambda_2 = \ldots = \lambda_n = 1/t$$

$$< ===== > \quad \lambda_i = \lambda_j \qquad \forall\, i, j$$

$$===== > \quad A A^t = 1/t\ I \qquad t > 0$$

$$===== > \quad (\sqrt{t}\ A) (\sqrt{t}\ A^t) = I$$

$$===== > \quad \sqrt{t}\ A \quad \text{es una matriz ortogonal.}$$

En conclusión podemos decir que el radio espectral de la matriz error es cero solamente cuando la matriz original es un múltiplo de una ortogonal. Esto no nos debe de extrañar dado que si A es ortogonal entonces

$$A^{-1} = A^t$$

En la tabla anterior se muestra una aproximación inicial a una matriz ortogonal para la cual ρ (E_o) dió 1.045687 D-15 y no fue cero debido a la precisión de la maquina.

Como conclusión podemos decir que tenemos formas de calcular la aproximación inicial, pero debemos de considerarlas junto con alguno de los métodos que se verán en el capítulo V .

C A P I T U L O I V

CALCULO DE UNA MATRIZ APROXIMACION INICIAL EN ALGUNOS CASOS ESPECIALES

IV . a Introducción.

La idea principal de este capítulo es mostrar que podemos mejorar la aproximación inicial a la inversa cuando las matrices tienen una estructura partícular.

Estas mejoras van en el sentido de que el radio espectral de la matriz error se reduce más y por supuesto indica que obtuvimos una aproximación inicial más ajustada a la inversa.

Haremos las respectivas comparaciones teoricas entre los resultados obtenidos cuando tomamos el valor de t que usaron Pan y Reif en su artículo y las nuevas modificamos propuestas.

También se muestran algunas implementaciones numericas que comprueban las resultados obtenidos.

Todos los programas que se desarrollaron fueron hechos con el paquete Matlab y la computadora mencionada anteriormente.
Muchos de los resultados estan considerados en [11].

IV b . Matrices Simétricas Positivas Definidas.

Supongamos ahora que A es una matriz simétrica positiva definida cuyos eigenvalores se encuentran ordenados de manera que

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$$

Tomemos entonces la aproximación inicial en la forma

$$B_0 = t\, I \qquad \text{con} \qquad t = \frac{1}{\| A \|_1}$$

Cabe señalar que E_0 en estas condiciones hereda el ser también una matriz simétrica positiva definida y por lo tanto ya no es necesario que B_0 sea un factor de la transpuesta de la matriz original.

De esta manera el radio espectral queda en lo forma

$$\rho (E_0) = \max_i | 1 - \lambda_i t |$$

$$= | 1 - \lambda_n t |$$

$$= \left| 1 - \frac{\lambda_n}{\| A \|_1} \right|$$

$$= 1 - \frac{\lambda_n}{\| A \|_1} < 1$$

pues $0 < \lambda_n \leq \rho(A) \leq \| A \|_1$, de manera que también E_o es una matriz convergente.

Veremos que si logramos una mejora respecto a las elecciones que proponen Pan y Reif. Primeramente con

$$t = \frac{1}{\| A \|_1 \| A^t \|_1}$$

Dado que A es ya simétrica positiva definida entonces cuando elegimos inicialmente

$$B_o = t A^t$$

obtenemos en estas condiciones que los eigenvalores de la matriz error son de la forma

$$1 - \lambda_i^2 t$$

De manera que

$$\rho(E_o) = \max_i \left| 1 - \frac{\lambda_i^2}{\| A \|_1 \| A^t \|_1} \right|$$

$$= 1 - \frac{\lambda_n^2}{\| A \|_1^2}$$

Así que debemos ver que se cumple

$$1 - \frac{\lambda_n}{\| A \|_1} \leq 1 - \frac{\lambda_n^2}{\| A \|_1^2}$$

$$<\!\!=\!\!=\!\!=\!\!=\!\!> \quad \frac{\lambda_n^2}{\| A \|_1^2} \leq \frac{\lambda_n}{\| A \|_1}$$

$$<\!\!=\!\!=\!\!=\!\!=\!\!> \quad \frac{\lambda_n}{\| A \|_1} \leq 1$$

$$<\!\!=\!\!=\!\!=\!\!=\!\!> \quad \lambda_n \leq \| A \|_1$$

lo cual es cierto.

Nuevamente por la relación entre los 2 radios espectrales de las elecciones de Pan y Reif se obtiene que para el valor de

$$t = \frac{1}{\| A A^t \|_1}$$

se llega a

$$1 - \frac{\lambda_n}{\| A \|_1} \leq 1 - \frac{\lambda_n^2}{\| A^2 \|_1}$$

por lo tanto , también logramos una mejora.
Para esta elección

$$B_o = t\, I \qquad \text{con} \qquad t = \frac{1}{\| A \|_1}$$

tenemos

$$\rho (E_o) \leq 1 - \frac{1}{n^{1/2}\ \text{cond}(A)}$$

ya que,

$$\rho (E_o) = \max_i \, | 1 - \lambda_i t |$$

$$= 1 - \lambda_n t$$

además

$$\lambda_n = \frac{1}{\| A^{-1} \|_2}$$

y por las equivalencias entre las normas matriciales

$$1/n^{1/2} \ \| A \|_2 \leq \| A \|_1$$

$$\Longrightarrow \quad \frac{1}{n^{1/2} \| A \|_1} \leq \frac{1}{\| A \|_2}$$

$$\Longrightarrow \quad - \frac{1}{\| A \|_2} \leq - \frac{1}{n^{1/2} \| A \|_1}$$

$$\rho(A) \geq 1 - \frac{t}{\|A\|_2} \leq 1 - \frac{t}{n^{1/2}\|A\|_1}$$

reemplazando A por A^{-1}

$$1 - \frac{t}{\|A^{-1}\|_2} \leq 1 - \frac{t}{n^{1/2}\|A^{-1}\|_1}$$

$$\rho(A) \geq 1 - \frac{t}{\|A^{-1}\|_2} \leq 1 - \frac{1/\|A\|_2}{n^{1/2}\|A^{-1}\|_1}$$

por lo tanto,

$$\rho(E_o) \leq 1 - \frac{1}{n^{1/2}\,\mathrm{Cond}(A)}$$

Se hizo una prueba con una matriz simétrica positiva definida la cual fue el producto de una matriz aleatoria de 5x5 por su transpuesta. Obtuvimos que con

$$t = \frac{1}{\|A\|_1}$$

$\rho(E_o)$ = 1.765781100982471D-15 y la cota que acabamos de obtener quedo 0.552786404500043 mientras que con

$$t = \frac{1}{\|A\|_1 \|A^t\|_1}$$

$\rho(E_o)$ = 3.531562201964940D-15

Con otra matriz aleatoria $\rho(E_o)$ = 0.998124734429357 y la cota del error dió 0.999032426240966 mientras que con la elección original de t $\rho(E_o)$ = 0.999996483379039 .

Por otra parte, si ahora consideramos

$$t = \frac{1}{\|A^tA\|_1}$$

y elegimos

$$B_o = t\,I$$

llegaremos a

$$\rho (E_o) \leq 1 - \frac{1}{n^{1/2} \text{ cond } (A A^t)}$$

Otra elección puede ser, usando

$$t = \frac{2}{\lambda_1 + \lambda_n}$$

y recordando que

$$\lambda_1 = || A ||_2 \qquad y \qquad \lambda_n = \frac{1}{|| A^{-1} ||_2}$$

a considerar ahora

$$B_o = \frac{2 I}{|| A ||_2 + \frac{1}{|| A^{-1} ||_2}}$$

y obtener una expresión para el radio espectral de la matriz error en la forma

$$\rho (E_o) = \left| 1 - \frac{2 \lambda_n}{|| A ||_2 + 1/|| A^{-1} ||_2} \right|$$

$$= \left| 1 - \frac{\frac{2}{|| A^{-1} ||_2}}{\frac{|| A ||_2 || A^{-1} ||_2 + 1}{|| A^{-1} ||_2}} \right|$$

$$= \left| 1 - \frac{2}{|| A ||_2 || A^{-1} ||_2 + 1} \right|$$

$$= 1 - \frac{2}{\text{cond } (A) + 1}$$

y efectivamente

$$\rho (E_o) < 1.$$

IV c . Matrices Fuerte Diagonalmente Dominantes.

Consideremos que $A \in \mathbb{R}^{n \times n}$ es fuerte diagonalmente dominante por renglones , es decir , existe $c \in \mathbb{R}$ tal que

$$\left(2 - 1/n^{c} \right) \left| a_{ii} \right| > \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| a_{ij} \right| \qquad \forall\, i$$

Para este tipo de matrices podemos considerar que la aproximación inicial puede tomar la forma

$$B_o = \operatorname{diag} \left\{ 1/a_{11} , 1/a_{22} , \ldots , 1/a_{nn} \right\}$$

$$= \begin{bmatrix} 1/a_{11} & 0 & \cdots & 0 \\ 0 & 1/a_{22} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & & 1/a_{nn} \end{bmatrix}$$

de manera que la matriz error tendra la forma

$$E_o = \begin{bmatrix} 0 & -a_{12}/a_{22} & \cdots & 0 \\ -a_{21}/a_{11} & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ -a_{n1}/a_{11} & -a_{n2}/a_{22} & \cdots & 0 \end{bmatrix}$$

Entonces

$$\| E_o \|_{\infty} = \max_{i} \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| a_{ji}/a_{jj} \right| \right\}$$

Por definición,

$$\left(2 - 1/n^{c} \right) \left| a_{ii} \right| > \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| a_{ij} \right| \qquad \forall\, i$$

$$\Rightarrow \quad 2 \mid a_{ii} \mid - 1/n \mid a_{ii} \mid > \sum_{\substack{j=1 \\ j \neq i}}^{n} \mid a_{ij} \mid + \mid a_{ii} \mid \qquad \forall\, i$$

$$\Rightarrow \qquad \mid a_{ii} \mid (1 - 1/n^{c}) > \sum_{\substack{j=1 \\ j \neq i}}^{n} \mid a_{ij} \mid \qquad \forall\, i$$

$$\Rightarrow \qquad (1 - 1/n^{c}) > \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \qquad \forall\, i$$

de aqui que

$$|| E_o ||_{\infty} \leq 1 - 1/n^{c}$$

Analogamente podemos mostrar que si nuestra definición fuera

$$(2 - 1/n^{c}) \mid a_{jj} \mid > \sum_{i=1}^{n} \mid a_{ij} \mid \qquad \forall\ j$$

esto es,si tuvieramos una matriz fuerte diagonalmente dominante por columnas llegaríamos con el mismo análisis a

$$|| E_o ||_{1} \leq 1 - 1/n^{c}$$

es decir, en los dos casos tendriamos que la matriz error inicial es convergente.

Vale la pena hacer notar lo simple que fue la primera aproximación inicial a la inversa de la matriz.

Terminamos mostrando un ejemplo de la modificación.

La matriz que tomamos para que cumpliera la definición es de orden 10 y tiene la forma

$$A = \begin{bmatrix} 20 & 1 & . & . & . & 0 & 0 \\ 1 & 20 & & & & 0 & 0 \\ 0 & 1 & & & & 0 & 0 \\ . & . & & & & & \\ . & . & & & & & \\ . & . & & & & & 1 \\ 0 & 0 & . & . & . & 1 & 20 \end{bmatrix}$$

de manera que cuando usamos la mejora en la aproximación inicial obtuvimos que $\rho (E_o) =$ 0.086602540378444 mientras que con

$$t = \frac{1}{\| A \|_1 \| A^t \|_1}$$

se obtiene $\rho (E_o) =$ 0.998112815166071.

IV d . Matrices Triangulares.

Supongamos ahora que la matriz A es triangular , ya sea superior o inferior , entonces afirmamos que la misma elección de la matriz aproximación inicial de la sección anterior funciona también para este tipo de matrices.

Sean

$$B_o = \text{diag} \{ 1/a_{11}, 1/a_{22}, \ldots , 1/a_{nn} \}$$

y

$$E_o = I - A B_o$$

Podemos calcular A^{-1} en un número finito de pasos usando B_o y E_o, aunque E_o no sea una matriz convergente.

En efecto,

$$A^{-1} = (A^{-1} B_o^{-1}) B_o$$

$$= (B_o A)^{-1} B_o$$

$$= (I - E_o)^{-1} B_o$$

por la estructura de la matriz A el error inicial toma la forma

$$E_o = \begin{bmatrix} 0 & 0 & & 0 \\ -a_{21}/a_{11} & 0 & \ldots & 0 \\ -a_{31}/a_{11} & -a_{32}/a_{22} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ -a_{n1}/a_{11} & -a_{n2}/a_{22} & \ldots & 0 \end{bmatrix}$$

de manera que E_o es una matriz nilpotente de orden n , es decir ,

$$E_o^i = 0 \qquad \text{si } i \geq n$$

Ahora

$$(I - E_o)^{-1} \sum_{i=0}^{n-1} E_o^i = I + E_o^n$$

$$= I$$

$$\text{------} \rangle \qquad (I - E_o)^{-1} = \sum_{i=0}^{n-1} E_o^i$$

de manera que

$$A^{-1} = \sum_{i=0}^{n-1} (E_o^i) B_o$$

Este es un caso importante que hay que tener en cuenta puesto que la inversión de cualquier matriz se reduce a la inversión de matrices triangulares si disponemos de la descomposicion LU ya que

$$A = L\,U$$

donde L es una matriz triangular inferior y U es triangular superior , y en consecuencia

$$A^{-1} = (L\,U)^{-1}$$
$$= U^{-1}\ L^{-1}$$

Por último , las pruebas que se efectuaron tomando como base una matriz aleatoria de $5x5$ triangular la cual con la primer aproximación en la forma

$$B_o = \text{Diag}\ \{\ 1/a_{11},\ 1/a_{22},\ \ldots,\ 1/a_{nn}\ \}$$

resultó $\rho(E_o) = 5.406303524256341\ D\text{-}04$ mientras que con

$$t = \frac{1}{\|A\|_1 \|A^t\|_1}$$

obtuvimos $\rho(E_o) = 1.002659537476453$.

Debemos hacer notar que los errores de redondeo influyeron en el ultimo cálculo del radio espectral dado que nos salió mayor que uno.

Finalmente daremos una tabla esquematizando las elecciones para las matrices de estructura partícular.

TIPO DE MATRIZ	APROXIMACION INICIAL
Hermitianas Positivas Definidas	$B_0 = \frac{1}{\|\| A \|\|} I$
Fuerte Diagonalmente Dominantes	$B_0 = Diag\ (1/a_{11}, \ldots, 1/a_{nn})$
Triangulares	$B_0 = Diag\ (1/a_{11}, \ldots, 1/a_{nn})$

Consideramos conveniente mostrar los resultados obtenidos también por medio de una tabla para tener una visión más completa en cuanto a las mejoras establecidas.

Matriz	Orden	Número de Condición	$\rho\ (E_0)$ con mejora respectiva	$\rho\ (E_0)$ cuando $B_0 = \frac{1}{\|\| A \|\|_1 \|\| A \|\|_\infty} A^t$
Hermitianas Positivas Definidas	5	659.9262515	1.7657811 D-15	3.53156220 D-15
	5	81.43809184	0.998124734429	0.9999948337903
Fuerte Diagonalmente Dominante	10	1.212265301	0.086602540378	0.9981128151660
Triangular	5	29.30241137	5.4063035 D-04	1.0026595374764

C A P I T U L O V

METODOS DE ORDEN P

A partir de este capítulo y los siguientes 3 daremos con detalle algunos de los métodos iterativos que existen para calcular la inversa de una matriz.

Iniciaremos con el caso más general de los métodos para después mostrar los que tienen una mayor rápidez de convergencia o los que son el óptimo del caso general para finalmente dar algunas variantes y ajustes a estos métodos a fin de que sean más efectivos numericamente.

V a .- Caso general.

La motivación de los métodos es la siguiente :

Supongamos que B_o es una aproximación inicial a la inversa de la matriz A. Sea

$$E_o = I - A B_o$$

el error cometido en esta aproximación. Reescribiendo esta expresión tendremos :

$$A^{-1} E_o = A^{-1} - B_o$$

$$A^{-1} - A^{-1} E_o = B_o$$

$$A^{-1} (I - E_o) = B_o$$

$$A^{-1} = (I - E_o)^{-1} B_o$$

ahora, si $|| E_o || < 1$ (E_o es una matriz convergente) entonces

$$A^{-1} = \sum_{i=o}^{\infty} E_o^i \; B_o$$

El cálculo de la inversa de la matriz en terminos de la última expresión es difícil de implementar numericamente, de manera que en vez de tomar la serie usaremos la aproximación

$$A^{-1} \cong \sum_{i=o}^{n} E_o^i \; B_o$$

Con este antecedente podemos ahora definir en base a nuestra aproximación inicial una sucesión de inversas aproximadas en la forma :

i) B_o aproximación inicial tal que $E_o = I - A B_o$ cumpla

$$|| E_o || < 1,$$

ii) $B_n = B_{n-1} (I + E_{n-1} + E_{n-1} + . . . + E_{n-1})$,

para $n,p \in \mathbb{N}$ $p \geq 2$

iii) $E_n = I - A B_n$, $n \in \mathbb{N}$

Notemos que iii) es el correspondiente ajuste del error cometido.

Mostremos ahora porque decimos que estos métodos son de orden p.

Utilizando ii) , iii) y la definición del error inicial tenemos

$$E_n = I - A B_n$$

$$= I - A B_{n-1} (I + E_{n-1} + E_{n-1}^2 + . . . + E_{n-1}^{p-1})$$

como

$$E_{n-1} = I - A B_{n-1}$$

$$\Longrightarrow \quad E_n = I - (I - E_{n-1}) (I + E_{n-1} + E_{n-1}^2 + ... + E_{n-1}^{p-1})$$

$$= I - I - E_{n-1} - E_{n-1}^2 - ... - E_{n-1}^{p-1} + E_{n-1} + E_{n-1}^2 + ... + E_{n-1}^{p-1} + E_{n-1}^p$$

$$= E_{n-1}^p$$

En otras palabras , el error cometido en la n-ésima iteración es la p-ésima potencia de la matriz error de la iteración anterior.

Esto motiva la siguiente :

Definición .- Diremos que un método iterativo para calcular la inversa de una matriz es de orden p si con las definiciones i), ii) y iii) se cumple

$$E_n = E_{n-1}^p$$

Siguiendo la misma idea anterior podemos deducir facilmente que

$$E_{n-1} = E_{n-2}^{P}$$

de lo cual se sigue

$$E_n = (E_n E_1) = (E_n E_2) = \overset{P}{\ldots} = (E_1^P) \overset{(n-1)}{} = (E_o^P)^n$$

$$n \in \mathbb{N}$$

Esto significa que el error cometido en cualquier iteración queda siempre en terminos del error original.

Enseguida queda plantear una pregunta :

¿ Los métodos que acabamos de definir convergen ? , es decir, ¿ Las aproximaciones sucesivas en realidad tienden a la inversa de la matriz original ? .

Para que esto suceda deberá ocurrir que el residuo

$$S_n = A^{-1} - B_n \quad \overset{n \to \infty}{\longrightarrow} \quad 0$$

Veremos que en efecto esto ocurre. Supongamos que tenemos una aproximación inicial tal que E_o es convergente. En partícular esto significa que $I - E_o$ es no-singular. Entonces

$$E_n = I - A B_n$$

$$\Longrightarrow \quad -(E_n - I) = A B_n$$

$$\Longrightarrow \quad A B_n = I - E_n$$

$$\Longrightarrow \quad B_n = A^{-1} - A^{-1} E_n$$

y además ,

$$E_o = I - A B_o$$

$$\Longrightarrow \quad A B_o = I - E_o$$

$$\Longrightarrow \quad A = (I - E_o) B_o^{-1}$$

$$\Longrightarrow \quad A^{-1} = B_o (I - E_o)^{-1}$$

Asi que,utilizando las igualdades anteriores tenemos:

$$S_n = A^{-1} - B_n$$

$$= A^{-1} - A^{-1} + A^{-1} E_n$$

$$= A^{-1} E_n$$

$$= A^{-1} E_o^{p^n}$$

$$= B_o (I - E_o)^{-1} E_o^{p^n}$$

Debemos darnos cuenta que B_o y $(I - E_o)^{-1}$ son matrices fijas, por lo tanto , la convergencia del método queda dependiente de una potencia de la matriz error inicial E_o de lo cual podemos decir entonces que nuestros métodos convergen si la matriz error inicial es convergente. De manera que si escogemos una matriz aproximación inicial a la inversa como una del tercer o cuarto capítulos aseguraremos que estos métodos serán convergentes.

Haremos algunas aclaraciones pertinentes. Es de suponer que a a la hora de implementar numericamente los métodos , la expresión

$$B_n = B_{n-1} (I + E_{n-1} + E_{n-1}^2 . . . + E_{n-1}^{p-1})$$

sea muy costosa de calcular si p es suficientemente grande, y además puede suceder que a la hora de calcular E_{n-1}^{p-1} los errores de redondeo se acumulen bastante de manera que influyan notablemente en el resultado y perdamos en consecuencia exactitud. Esto nos obliga entonces a tratar de eliminar este problema buscando un orden adecuado para el método , lo cual haremos en la siguiente parte del capítulo.

Para terminar esta sección mostraremos los resultados de aplicar los métodos de orden 2,3 y del 6 al 11 a una matriz aleatoria de orden 10 y terminamos el proceso cuando las entradas de E_n son menores que 1×10^{-5}.

O R D E N	Número de Iteraciones
2	20
3	13
6	9
7	8
8	8
9	7
10	7
11	7

Notemos que a pesar de que en algunos de los métodos el número de iteraciones coincide,debemos tomar en cuenta que al pasar de un método de orden p a uno de orden p+1 se requieren n^3 operaciones extras. A pesar de esto, la exactitud de la solución es mejor conforme aumenta el orden del método.

V b. Elección del Método Optimo.

En esta sección nos hacemos la pregunta de cuál es el mejor valor para p , es decir , cuál es el mejor método de entre todos los planteados.

El problema también consiste en decir qué significa que un método sea mejor que los otros.

Diremos que un método es óptimo cuando para una determinada cantidad de operaciones se obtiene una mejor aproximación a la inversa de la matriz en comparación con los otros métodos.

Tratando de encontrar el método óptimo y partiendo ya del hecho de que E_o es una matriz convergente , supongamos que $\lambda_1, \lambda_2, \ldots, \lambda_n$ son sus correspondientes eigenvalores y

$$\xi = \max_i \; | \lambda_i |$$

entonces dado que los eigenvalores de E_o^p son λ_i^p y

$$E_n = E_o^{p^n}$$

en consecuencia el polinomio característico de E_n se anula en

$$\xi^{p^n}$$

Ahora bien,tomando en cuenta que el producto de 2 matrices requiere n^3 operaciones,entonces m iteraciones requieren $m\ n^3$ operaciones , de manera que para efectuar m iteraciones de un método de orden p se necesitan $m\ n^3 p$ operaciones.

Supongamos que restringimos nuestras operaciones por limitaciones de tiempo o espacio a k.

Asi que el número de iteraciones que nos vemos obligados a efectuar son solamente

$$n = \frac{k}{p\ m^3}$$

suponiendo por supuesto que $k/pm^3 \in \mathbb{N}$.

Sustituyendo el valor de n en el eigenvalor principal tenemos

$$\xi^{p^n} = \xi^{p^{k^3/(pm)}} = \xi^{(p^{1/p})^{k/m^3}}$$

y es obvio que k,m y ξ son independientes del número p.

Si consideramos la última expresión como una función entonces el error se mínimiza cuando

$$p^{1/p}$$

es máximo, y sabemos que la funcion $X^{1/x}$ alcanza su máximo en $x = 2.7118... = e$.

El problema es que nuestros métodos son de orden $p \geq 2$ y por lo tanto este valor obtenido no encaja en el desarrollo anterior. Pero un resultado debido a M. Altman [6], demuestra que para enteros p el máximo se alcanza en $p = 3$. Por lo tanto, desde un punto de vista teórico, el problema queda resuelto.

En resumidas cuentas, con estos argumentos obtuvimos que el mejor método es

i') B_0 tal que $E_0 = I - A B_0$ cumpla que $\| E_0 \| < 1$,

ii') $B_n = B_{n-1} (I + E_{n-1} + E_{n-1})$,

iii') $E_n = I - A B_n$

Como se observa en la tabla anterior, el método óptimo requirió de 13 iteraciones y aunque los métodos de orden 6 al 11 se llevan menos iteraciones, lo cierto es que necesitan más operaciones.

C A P I T U L O V I

METODO DE NEWTON

En el capítulo anterior se vio que el mejor método para aproximar la inversa de una matriz es cuando $p = 3$.

Pasaremos ahora a describir con más detalle el método de orden 2 llamado Método de Newton. Dicho método tiene una motivación natural que vale la pena mencionar.

Definamos nuevamente para $i = 0, 1, 2, \ldots$ la matriz error correspondiente a la aproximación B_i a la inversa

$$E_i = I - A B_i$$

Entonces si la i-ésima aproximación B_i a A^{-1} es buena esperamos que $B_i E_i$ sea cercano a $A^{-1} E_i$, y como

$$A^{-1} E_i = A^{-1} - B_i$$

por lo tanto, esta última expresión es justo lo que debemos sumar a B_i para llegar a la inversa de la matriz A, ya que

$$B_i + B_i E_i \cong B_i + A^{-1}E_i = B_i + A^{-1} - B_i = A^{-1}$$

De manera que si nosotros usamos $B_i + B_i E_i = B_i (I + E_i)$ como la siguiente inversa aproximada, llamemosla B_{i+1}, podremos esperar que esta sea una mejor aproximación a A^{-1}.

En suma, la iteración

$$B_{i+1} = B_i + B_i E_i$$

$$= B_i + B_i (I - A B_i)$$

parece un buen método para aproximar cada vez mejor la inversa de una matriz dada.

Este método generalmente se le conoce como el Método de Newton, porque su caso escalar

$$b_{i+1} = b_i + b_i (1 - a b_i)$$

es precisamente el proceso de Newton para resolver la ecuación

$$f (x) = a - 1/x$$

La matriz error en la n-ésima iteración es de la forma

$$E_n = E_o^{n^2}$$

ya que es el caso $p = 2$ de los métodos definidos en el capítulo anterior.

Mostremos una forma alterna de ver cuán rápida es la convergencia con este orden del método.

Dado que

$$E_n = I - A B_n$$

$$\Longrightarrow \quad E_o^{n^2} = I - A B_n$$

$$\Longrightarrow \quad A B_n = I - E_o^{n^2}$$

asi ,

$$B_n = A^{-1} (I - E_o^{n^2})$$

Esto muestra que B_n se aproxima bastante bien a A^{-1} siendo la convergencia del proceso del tipo cuadrático.

Hagamos una estimación del error que se comete al aproximar A^{-1} por B_n. Primeramente

$$E_o = I - A B_o$$

$$\Longrightarrow \quad A B_o = I - E_o$$

$$\Longrightarrow \quad A^{-1} = B_o (I - E_o)^{-1}$$

entonces,

$$A^{-1} E_o^{n^2} = B_o (I - E_o)^{-1} E_o^{n^2}$$

Por otro lado , si

$$|| E_o || \leq k$$

entonces

$$|| I - E_o || \leq 1 - k$$

por lo tanto

$$|| (I - E_o)^{-1} || \leq || (I - E_o) ||^{-1} \leq \frac{1}{1 - k} .$$

De manera que

$$\| B_n - A^{-1} \| = \| A^{-1} (I - E_o^{n^2}) - A^{-1} \|$$

$$= \| - A^{-1} E_o^{n^2} \|$$

$$= \| - B_o (I - E_o)^{-1} E_o^{n^2} \|$$

$$\leq \| B_o \| \| (I - E_o)^{-1} \| \| E_o^{n^2} \|$$

$$\leq \| B_o \| \frac{k^{n^2}}{1 - k}$$

en consecuencia, si hacemos

$$\| E_o \| = \| I - A B_o \| \leq k < 1$$

entonces el número de digitos significativos va coincidiendo de manera geométrica en cada iteración.

Estas aproximaciones sucesivas deben de ser calculadas de la siguiente manera :

$$B_i = B_{i-1} (I + E_{i-1})$$

$$= B_{i-1} (2I - A B_{i-1})$$

$$= B_{i-1} + B_{i-1} (I - A B_{i-1})$$

Es de notar que el segundo sumando de la última expresión juega el papel de una pequeña correción a la aproximación del paso anterior.

Consultando la tabla del capítulo anterior, vemos que el método de Newton se llevó 20 iteraciones y por lo tanto respecto al método óptimo se requirieron 7 iteraciones de más.

Vale la pena mencionar que como tenemos convergencia cuadrática , si el método ya tiene 1 cifra significativa correcta entonces en la siguiente tendremos el doble , y asi sucesivamente. Al momento de estar realizando las pruebas , notamos que si ya teníamos una cifra significativa correcta , en 3 ó 4 iteraciones más llegabamos al máximo de exactitud posible. Para confirmar esta última aseveración aplicamos el método de Newton a una matriz fuerte diagonalmente dominante.

La aproximación $B_o = Diag (1/a_{11}, 1/a_{22}, \ldots, 1/a_{nn})$ dió lugar a $\rho (E_o) = 0.0866025403$ y para dicha matriz obtuvimos el máximo de exactitud posible para A^{-1} en sólo 4 iteraciones siendo $\| E_4 \|_\infty = 2.6404347 \quad D-16$.

C A P I T U L O V I I

AJUSTES NUMERICOS A LOS METODOS Y COMPARACIONES

En este capítulo daremos los detalles de una implementación debida a Thomas E. Phipps Jr. que es una variante de los métodos p=5 y p=6. También daremos algunas comparaciones entre el método del refinamiento iterativo, el método de Newton y el método óptimo de orden 3.

VII a. Modificaciones de Thomas E. Phipps Jr.

En su artículo "The Inversion of Large Matrices " [12] el autor propone la primera aproximación inicial que dan Pan y Reif [11]

$$B_0 = t\,A^t \quad , \qquad t = \frac{1}{\|A\|_1 \, \|A^t\|_1}$$

la cual da lugar a una matriz E_0 convergente.

Phipps propone una reordenación de los métodos de orden p de la siguiente manera

$$\begin{aligned} B_n &= (I + E_n + E_n^2 + \ldots + E_n^{p-1})\, B_{n-1} \\ &= (I + E_n + E_n^2 + \ldots + E_n^{m+1})\, B_{n-1} \quad , \; m = p + 2 \\ &= (I + (I + E_n + E_n^2 + \ldots + E_n^m)\, E_n)\, B_{n-1} \end{aligned}$$

De manera que podemos definir

$$Q_n = (I + E_n + E_n^2 + \ldots + E_n^m) \, , \quad m \in \mathbb{N} \qquad (5)$$

y entonces escribir la iteración en la forma

$$B_n = (I + Q_n E_n)\, B_{n-1}$$

Una forma de acelerar la convergencia del método definido antes es : podemos suponer que eventualmente los elementos de la matriz E_0 fuera de la diagonal tenderan a cero, de manera que una mejora para aproximar Q_n es considerar dichos elementos de E_n como cero , y en consecuencia Q_n será una matriz diagonal fácil de calcular numericamente. Además, podemos escribir los elementos de la matriz diagonal Q_n en la forma

$$(Q_n)_{ii} = 1 + (E_n)_{ii} + (E_n)_{ii}^2 + \ldots + (E_n)_{ii}^m = \frac{1 - (E_n)_{ii}^{m+1}}{1 - (E_n)_{ii}}$$

Un problema con esta modificación es la elección del entero m en la ecuación (5) . Debemos escogerlo con cuidado puesto que con m grande , digamos mayor que 10 , tenemos un método costoso operacionalmente y si m es 1 ó 2 la convergencia puede ser lenta. Ya se vió en el capítulo V que el óptimo teórico para un método es p = 3. Diremos al respecto que Phipps encontró desde el punto de vista numérico que los mejores valores para m son 4 ó 5 y para estos casos los métodos obtenidos son más efectivos que el de Newton.

VII b) Comparación entre el Refinamiento Iterativo , el Método de Newton y el Método de Orden 3.

La idea de esta sección consiste en mostrar cuál de los métodos es más efectivo para obtener una inversa tan aproximada como sea posible , esto es , el máximo de precisión que nos permita la aritmética de la maquina en que estemos trabajando.

Primero , para el método del refinamiento iterativo obtenemos una aproximación a la inversa de la matriz en precisión sencilla resolviendo los sistemas

$$A\,x = e_i \qquad i = 1 , \ldots , n$$

donde los e_i son los vectores canónicos.

y después aplicamos el refinamiento iterativo a cada uno de estos sistemas para encontrar la inversa más exacta posible.

Cabe añadir que por supuesto el residual $r_i = b - A\,x_i$ fue evaluado en doble precisión y los demás cálculos en precisión sencilla, tal como debe implementarse dicho método (Ver Apéndice).

El criterio para detener el proceso es :

Si la precisión de la maquina no nos permite mejorar la solución y si la aritmética que tenemos es de t dígitos significativos entonces pararemos cuando

$$\frac{\| r_i \|}{\| x_i \|} \leq 10^{-t} .$$

Segundo , para el método de Newton calculamos todas las aproximaciones en precisión sencilla mientras que los errores fueron calculados en doble precisión para así estar en igualdad de circunstancias respecto al método anterior.

El criterio para terminar este proceso es el siguiente :

como $B_n = (I + E_{n-1}) B_{n-1}$

y $E_{n-1} = I - A * B_{n-1}$

$\Longrightarrow$ $\| B_n - B_{n-1} \| = \| B_{n-1} + E_n B_{n-1} - B_{n-1} \|$

$= \| E_{n-1} B_{n-1} \|$

además

$\| E_{n-1} B_{n-1} \| \leq \| E_{n-1} \| \| B_{n-1} \|$

por lo tanto

$$\frac{\| B_n - B_{n-1} \|}{\| B_{n-1} \|} \leq \| E_{n-1} \|$$

de manera que este criterio y el correspondiente para el refinamiento iterativo son analogos.

Tercero , el método de orden 3 fue implementado de la misma manera que el de Newton asi que nada más falta decir en que forma terminamos las iteraciones para este caso.

Ahora el método es de la forma

$$B_n = (I + E_{n-1} + E_{n-1}^2) B_{n-1}$$

$$E_n = I - A B_n$$

entonces

$$\| B_n - B_{n-1} \| = \| B_{n-1} + E_{n-1} B_{n-1} + E_{n-1}^2 B_{n-1} - B_{n-1} \|$$

$$= \| (E_{n-1} + E_{n-1}^2) B_{n-1} \|$$

como

$$\| (E_{n-1} + E_{n-1}^2) B_{n-1} \| \leq \| E_{n-1} + E_{n-1}^2 \| \| B_{n-1} \|$$

y

$$\| E_{n-1} + E_{n-1}^2 \| \leq \| E_{n-1} \| + \| E_{n-1}^2 \|$$

$$\leq \| E_{n-1} \| + \| E_{n-1} \|^2$$

por lo tanto ,

$$\frac{\| B_n - B_{n-1} \|}{\| B_n \|} \leq \| E_{n-1} \| + \| E_{n-1} \|^2$$

Si $\| E_{n-1} \| \xrightarrow[n \to \infty]{} 0$, entonces en $\| E_{n-1} \| + \| E_{n-1} \|^2$ el sumando $\| E_n \|$ pesará más en el resultado de manera que podemos decir que este es un criterio equivalente al de los dos metodos anteriores.

Pasemos ahora a calcular el número de operaciones (multiplicaciones) que se lleva cada uno de los métodos.

Calcular la inversa por algún método directo lleva alrededor de n^3 operaciones mientras que el refinamiento iterativo por cada nueva iteración, que es basicamente 1 sustitución hacia atrás y otra hacia adelante pues ya tenemos a la matriz A descompuesta en la forma LU, nos lleva entonces $n(n+1)$ operaciones, pero esto con cada uno de los vectores canónicos siendo ellos el lado derecho de la ecuación, asi que en total tenemos $n^3 + n^2$ operaciones para efectuar cada iteración del método.

Usando el método de Newton con la primera aproximación en la forma

$$B_o = \frac{1}{\|A\|_1 \, \|A^t\|_1} A^t$$

se lleva $3n^2 - 2n + 1$ operaciones , y el número de operaciones por cada iteración son $2n^3$.

Para el método de orden 3 no es díficil concluir que se llevará por cada iteración $3n^3$ operaciones.

Para terminar daremos los resultados de las pruebas establecidas por medio de una tabla.

Nuevamente los cálculos fueron hechos con Matlab y con la misma computadora que antes , siendo la unidad de redondeo, para doble precisión de 2.220446049250313 D-16, mientras que en precisión simple fue de 9.536743164062500 E-07.

Matriz	Orden	Número de Condición	Refinamiento Iterativo	Método Orden 3	Método Newton
Aleatoria					
	3	6.20820870950	3	7	10
	5	9.02430561556	3	8	13
	8	21.0730743408	3	10	14
	10	112.730346679	3	12	20
	12	67.7683715820	3	13	20
	15	317.665527343	3	11	17
	18	126.612304687	3	12	18
	20	378.272798570	4	14	23
	25	248.671630859	3	15	22
Wilkinson (pag.132)	3	274908.3	7	* 26	* 42
Mágica	5	5.46185302734	3	8	10
Hilbert	3	524.009765623	3	14	23
Hessenberg	3	221.476074218	3	13	20
Ortogonal	5	1.00001430511	3	5	6
Schur	5	16.3526306152	5	8	11

* Cálculos efectuados en doble precisión

La matriz mágica que se hace mención es aquella cuya suma por columnas es igual que por renglones.

La matriz de Hilbert es la inversa de la matriz con elementos

$$\frac{1}{(\ i + j - 1\ )}$$

la cual es un famoso ejemplo de una matriz mal condicionada.

La matriz en la forma de Hessenberg tiene ceros debajo de la primera subdiagonal.

Finalmente la matriz con el nombre es Schur es el resultado de la descomposición de Schur.

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

CONCLUSIONES Y COMENTARIOS

1. Esperabamos que los métodos iterativos implementados en maquinas en serie fueran eficientes como lo son los métodos directos,tomando en cuenta los comentarios de [11] y [12], pero los experimentos numericos de los capítulos V, VI y VII mostraron lo contrario dado que el número de operaciones que requieren los metodos iterativos es mucho mayor que las que necesita un método directo, como la eliminación gaussiana, para invertir una matriz. Esta conclusión en cuestiones prácticas significa que los métodos iterativos planteados no compiten con los directos.
2. Dado un método de orden p, si tenemos una aproximación que coincida en al menos 1 dígito significativo con la solución exacta, en la siguiente iteración la matriz resultante coincidirá con la solución exacta en otros p dígitos significativos. Pero el problema es que una aproximación con tales características requiere de un cierto número de iteraciones y es aquí donde el proceso se hace lento. Esto es consecuencia de las pruebas efectuadas con el método de Newton que con una buena aproximación inicial sólo necesito 4 iteraciones en obtener el máximo de precisión posible.
3. Debemos hacer notar que el número de operaciones que estos métodos requieren para obtener una buena aproximación son comparativamente más elevadas que el de la eliminación gaussiana por ejemplo , pues este se lleva alrededor de n^3 operaciones en obtener la inversa mientras que el método de Newton, por citar alguno de ellos, tan sólo en el cálculo de la primera aproximación se lleva $3n^2 - 2n + 1$ operaciones cuando

$$B_0 = \frac{1}{\| A \|_1 \, \| A^t \|_1} A^t$$

y por cada nueva iteración se efectuan $2n^3$ operaciones más.
A pesar de esto, el esquema de los métodos es muy sencillo.
4. Podemos decir que los métodos expuestos y sus ajustes numéricos pueden ser utilizados para dar una mayor exactitud a la inversa de una matriz cuando ella es obtenida por algun otro procedimiento y en algunos problemas resulta importante tener la mejor aproximación posible a la inversa exacta.
5. Los métodos de un orden grande no son efectivos numéricamente puesto que el número de operaciones crece considerablemente y puede suceder que no convergan debido a los errores de redondeo acarreados en los calculos o también por el mal condicionamiento de la matriz original.

6. Debemos también tomar en cuenta que si queremos una convergencia razonable del método elegido es necesario una elección efectiva de la matriz aproximación inicial , para lo cual hemos planteado varias formas de calcular fácilmente esta matriz, las que garantizan que el proceso tendrá el éxito esperado y si deseamos acelerar la convergencia debemos tomar en cuenta la estructura de la matriz y el número de operaciones que requiere tal aceleración.
7. En cuanto a los métodos establecidos , notemos que la mejora efectuada a los métodos de orden p son consecuencia de las pruebas numericas aplicadas a los métodos en cuestión,asi que coincidimos con Phipps al afirmar que uno mismo podría lograr otras modificaciones o ajustes que aceleren la convergencia de los métodos. Lo ideal sería encontrar una aproximación inicial que coincida con la inversa exacta en al menos 1 dígito significativo en cada entrada.
8. No tuvimos forma de comprobar las afirmaciones de Pan y Reif [10] acerca de la efectividad de los métodos iterativos en maquinas en paralelo debido a la falta de una de ellas, pero si debemos señalar que en el mencionado artículo afirman que dichos métodos son efectivos aún en maquinas con un solo procesador , lo cual ya comprobamos que no es cierto.

A P E N D I C E

METODO DEL REFINAMIENTO ITERATIVO

En algunas ocasiones es importante que se tenga la solución más exacta posible (en la maquina que estemos usando) del sistema

$$A \; x = b$$

La clave para obtener una mejor exactitud de una primer solución x_1 obtenida por algún método directo es el cálculo en doble precisión del residual

$$r_1 = b - A\,x$$

Conociendo r_1, resolvemos el sistema

$$A\,d_1 = r_1$$

de manera que

$$x_2 = x_1 + d_1$$

debería resolver el sistema $A\,x = b$ dado que

$$A\,x_2 = A\,(\,x_1 + d_1\,) = A\,x_2 + A\,d_1 = A\,x_1 + r_1 = b$$

por lo tanto, x_2 proporciona una solución más exacta que x_1.

En la práctica y debido a los errores de redondeo , x_2 no resuelve exactamente el sistema pero si podemos repetir el argumento para encontrar cada vez soluciones más exactas.

Formando ahora $r_2 = b - A\,x_2$, podemos calcular la solución del sistema $A\,d_2 = r_2$ y obtendremos $x_3 = x_2 + r_2$.

Podemos seguir el proceso y encontrar los vectores $x_1, x_2, \ldots$ que forman una sucesión la cual rápidamente converge al vector que es la solución exacta de $A\,x = b$ en precisión sencilla.

Notemos que cada sistema

$$A\,d_i = r_i$$

tiene la misma matriz A , y si inicialmente usamos, por ejemplo, la descomposición L U para calcular la primera solución x_1 entonces se puede hacer uso de dicha descomposición reduciendose así el número de operaciones.

Por lo tanto, el proceso del refinamiento iterativo consiste en :

i) $A\,x_1 = b$

ii) $r_i = b - A\,x_i$, calculado en doble precisión.

iii) $A\,d_i = r_i$

A P E N D I C E

METODO DEL REFINAMIENTO ITERATIVO

En algunas ocasiones es importante que se tenga la solución más exacta posible (en la maquina que estemos usando) del sistema

$$A \; x = b$$

La clave para obtener una mejor exactitud de una primer solución x_1 obtenida por algún método directo es el cálculo en doble precisión del residual

$$r_1 = b - A \; x$$

Conociendo r_1, resolvemos el sistema

$$A \; d_1 = r_1$$

de manera que

$$x_2 = x_1 + d_1$$

debería resolver el sistema $A \; x = b$ dado que

$$A \; x_2 = A \; (\; x_1 + d_1 \;) = A \; x_1 + A \; d_1 = A \; x_1 + r_1 = b$$

por lo tanto, x_2 proporciona una solución más exacta que x_1.

En la práctica y debido a los errores de redondeo , x_2 no resuelve exactamente el sistema pero si podemos repetir el argumento para encontrar cada vez soluciones más exactas.

Formando ahora $r_2 = b - A \; x_2$, podemos calcular la solución del sistema $A \; d_2 = r_2$ y obtendremos $x_3 = x_2 + r_2$.

Podemos seguir el proceso y encontrar los vectores $x_1, x_2, \ldots$ que forman una sucesión la cual rápidamente converge al vector que es la solución exacta de $A \; x = b$ en precisión sencilla.

Notemos que cada sistema

$$A \; d_i = r_i$$

tiene la misma matriz A , y si inicialmente usamos, por ejemplo, la descomposición L U para calcular la primera solución x_1 entonces se puede hacer uso de dicha descomposición reduciendose así el número de operaciones.

Por lo tanto, el proceso del refinamiento iterativo consiste en :

i) $A \; x_1 = b$

ii) $r_i = b - A \; x_i$, calculado en doble precisión.

iii) $A \; d_i = r_i$

Los pasos ii) y iii) deben repetirse hasta que la precisión de la maquina no nos permita mejorar la solución. De manera que si nuestra aritmetica es de t dígitos significativos , el criterio para detener el proceso será cuando

$$\frac{\| d_{i+1} \|}{\| d_i \|} \leq 10^{-t}$$

Debemos recalcar que es esencial que el residuo r_i sea calculado con una precisión más alta que la del resto de los cálculos. Esto es un principio general en toda resolución de ecuaciones : el cálculo del residual es la operación crítica y debe ser realizada con la mayor exactitud. Para más detalles consultar [3] y [14].

B I B L I O G R A F I A

1. Atkinson, K. E. , An Introduction to Numerical Analysis , Wiley,New York (1978).

2. Fadeevva, V. N. , Computational Methods of Linear Algebra ,

3. Forsythe, G. E. , Moler, C. B. , Computer Solution Of Linear Algebraic Systems , Prentice-Hall,New Jersey (1967).

4. Golub, G. , Van Loan, C.F. , Matrix Computations , The Johns Hopkins University , Baltimore , Maryland (1983).

5. Householder, A. S. , The Teory Of Matrices In Numerical Analysis , Dover ,New York (1964).

6. Issacson, E., Keller, H. B. , Analysis Of Numerical Methods , Wiley,New York (1966).

7. Lawson, Ch. L. , Hanson, R. J. , Solving Least Squares Problems, Prentice Hall, Englewood Cliffs, New Jersey (1974).

8. Mardia, K. V., Kent, J. T., Bibby, J. M. , Multivariate Analysis, Academic Press, London (1979).

9. Moler , Cleve , Matlab User' Guide , 1983 , Department of Computer Science, University of New México.

10. Noble, B, Van Loan C. F. , Applied Linear Algebra , Prentice Hall, New Jersey (1977).

11. Pan, V. , Reif, J., " Efficient Parallel Solution Of Linear Systems " , Technical Report TR-02-85 , Center For Research in Computing Technology , Harvard University (1985).

12. Phipps, T. E., " The Inversion of Large Matrices ", Byte, April 1986 Vol.11, No.4.

13. Schendel , U. , Introduction To Numerical Methods For Parallel Computers , Ellis Horwood , England (1984).

14. Stewart, G. W. , Introduction to Matrix Computations , Academic Press,New York (1973).

15. Varga, R. S. , Matrix Iterative Analysis , Prentice-Hall, (1962).

16. Wilkinson, J. H. , Rounding Errors in Algebraic Processes, Prentice-Hall (1963).