CURSO DE METODOS NUMERICOS

TERCERA PARTE

METODOS PARA LA RESOLUCION

DE SISTEMAS LINEALES

CAPITULO XIII. METODOS PARA LA RESOLUCION DE SISTEMAS LINEALES: PRELIMINARES

1. SISTEMAS LINEALES DE ECUACIONES

En esta tercera parte se consideran técnicas para resolver el sistema de ecuaciones lineales:

$$\begin{aligned} E_1: &\quad a_{11}\,x_1 + a_{12}\,x_2 + \ldots + a_{1n}\,x_n = b_1\ , \\ E_2: &\quad a_{21}\,x_1 + a_{22}\,x_2 + \ldots + a_{2n}\,x_n = b_2\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad a_{n1}\,x_1 + a_{n2}\,x_2 + \ldots + a_{nn}\,x_n = b_n\ , \end{aligned} \tag{$XIII$.1}$$

para $x_1, \ldots, x_n$, dadas las a_{ij} para cada $i, j = 1, 2, \ldots, n$, y las b_i, para cada $i = 1, 2, \ldots, n$.

Los procedimientos de resolución de sistemas de ecuaciones lineales se dividen fundamentalmente en dos grupos:

(1) **procedimientos exactos o técnicas directas**, que son algoritmos finitos para cálculo de las raíces de un sistema (tales como la regla de Cramer, el método de Gauss, etc.);

(2) **procedimientos iterativos**, los cuales permiten obtener las raíces de un sistema con una exactitud dada mediante procesos infinitos convergentes (éstos incluyen el método de iteración, el de Seidel, el de relajación, etc.).

Debido al inevitable redondeo, incluso los resultados de procedimientos exactos son *aproximados*, viéndose comprometida, en el caso general, la estimación del error de las raíces. En el caso de procesos iterativos ha de añadirse el error del método.

Para resolver un sistema lineal como el de ($XIII$.1) están permitidas tres operaciones en las ecuaciones:

(1) la ecuación E_i puede multiplicarse por cualquier constante λ diferente de cero y se puede usar la ecuación resultante en lugar de E_i. Esta operación se denotará por $(\lambda E_i) \to (E_i)$;

(2) la ecuación E_j puede multiplicarse por cualquier constante λ diferente de cero, sumarla a la ecuación E_i, y usar la ecuación resultante en lugar de E_i. Esta operación se denotará por $(E_i + \lambda E_j) \to (E_i)$;

(3) las ecuaciones E_i y E_j se pueden intercambiar. Esta operación se denotará por $(E_i) \leftrightarrow (E_j)$.

Por medio de una secuencia de las operaciones anteriores, un sistema lineal se puede transformar a un sistema lineal más fácil de resolver y teniendo el mismo conjunto de soluciones. La secuencia de operaciones se ilustrará en el ejemplo siguiente.

Ejemplo. Resolver las cuatro ecuaciones:

$$\begin{array}{lrcrcrcrcr} E_1: & x_1 & + & x_2 & & & + & 3\,x_4 & = & 4\ , \\ E_2: & 2\,x_1 & + & x_2 & - & x_3 & + & x_4 & = & 1\ , \\ E_3: & 3\,x_1 & - & x_2 & - & x_3 & + & 2\,x_4 & = & -3\ , \\ E_4: & -\,x_1 & + & 2\,x_2 & + & 3\,x_3 & - & x_4 & = & 4\ , \end{array} \tag{$XIII$.2}$$

para las incógnitas x_1, x_2, x_3, x_4. Un primer paso puede ser usar la ecuación E_1 para eliminar la incógnita x_1 de E_2, E_3 y E_4 efectuando $(E_2 - 2E_1) \to (E_2)$, $(E_3 - 3E_1) \to (E_3)$, y $(E_4 + E_1) \to (E_4)$. El sistema resultante es:

$$\begin{array}{lrcrcrcrcr} E_1: & x_1 & + & x_2 & & & + & 3\,x_4 & = & 4\,, \\ E_2: & & - & x_2 & - & x_3 & - & 5\,x_4 & = & -7\,, \\ E_3: & & - & 4\,x_2 & - & x_3 & - & 7\,x_4 & = & -15\,, \\ E_4: & & & 3\,x_2 & + & 3\,x_3 & + & 2\,x_4 & = & 8\,. \end{array} \qquad (XIII.3)$$

En el nuevo sistema, se usa E_2 para eliminar x_2 de E_3 y E_4 por medio de las operaciones $(E_3 - 4E_2) \to (E_3)$ y $(E_4 + 3E_2) \to (E_4)$, resultando el sistema:

$$\begin{array}{lrcrcrcrcr} E_1: & x_1 & + & x_2 & & & + & 3\,x_4 & = & 4\,, \\ E_2: & & - & x_2 & - & x_3 & - & 5\,x_4 & = & -7\,, \\ E_3: & & & & + & 3\,x_3 & + & 13\,x_4 & = & 13\,, \\ E_4: & & & & & & - & 13\,x_4 & = & -13\,. \end{array} \qquad (XIII.4)$$

Este último sistema está ahora en **forma triangular** o reducida y puede resolverse fácilmente para encontrar las incógnitas por un proceso de **sustitución hacia atrás**. Notando que E_4 implica que $x_4 = 1$, E_3 puede resolverse para x_3:

$$x_3 = \frac{1}{3}\,(13 - 13\,x_4) = \frac{1}{3}\,(13 - 13) = 0\,.$$

Continuando, x_2 resulta ser:

$$x_2 = -(-7 + 5\,x_4 + x_3) = -(-7 + 5 + 0) = 2\,;$$

y x_1 es:

$$x_1 = 4 - 3\,x_4 - x_2 = 4 - 3 - 2 = -1\,.$$

Por lo tanto la solución a $(XIII.4)$ es $x_1 = -1$, $x_2 = 2$, $x_3 = 0$ y $x_4 = 1$. Se puede verificar fácilmente que estos valores son también solución de las ecuaciones $(XIII.2)$.

Cuando realizamos los cálculos del ejemplo, no necesitamos escribir las ecuaciones completas en cada paso, ya que la única variación de sistema a sistema ocurre en los coeficientes de las incógnitas y en los términos independientes de las ecuaciones. Por esta razón, un sistema lineal se reemplaza frecuentemente por una matriz, que contiene toda la información del sistema que es necesaria para determinar su solución, pero en forma compacta.

La notación para una matriz $n \times m$ será una letra mayúscula como A para la matriz y letras minúsculas con subíndices dobles como a_{ij}, para referirse a la componente en la intersección de la $i-$ésima fila y la $j-$ésima columna:

$$A = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}\,.$$

Para representar al sistema lineal $(XIII.1)$ puede usarse una matriz $n \times (n+1)$, construyendo primero

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{pmatrix}$$

y luego combinando estas matrices para formar la **matriz ampliada**

$$A_a = [A, \mathbf{b}] = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n \end{array}\right) ,$$

donde se usa la barra para separar los coeficientes de las incógnitas de los términos independientes de las ecuaciones.

Ejemplo. Repetiremos el ejemplo anterior en notación matricial. La matriz ampliada asociada con el sistema $(XIII.2)$ será:

$$\left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 2 & 1 & -1 & 1 & 1 \\ 3 & -1 & -1 & 2 & -3 \\ -1 & 2 & 3 & -1 & 4 \end{array}\right) . \qquad (XIII.5)$$

Las operaciones asociadas con $(E_2 - 2E_1) \to (E_2)$, $(E_3 - 3E_1) \to (E_3)$, y $(E_4 + E_1) \to (E_4)$ en el sistema $(XIII.2)$ se llevan a cabo manipulando las filas respectivas de la matriz ampliada $(XIII.5)$, la cual se transforma en la matriz correspondiente al sistema $(XIII.3)$:

$$\left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & -4 & -1 & -7 & -15 \\ 0 & 3 & 3 & 2 & 8 \end{array}\right) . \qquad (XIII.6)$$

Realizando las manipulaciones finales, $(E_3 - 4E_2) \to (E_3)$ y $(E_4 + 3E_2) \to (E_4)$, se obtiene la matriz ampliada correspondiente al sistema $(XIII.4)$:

$$\left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & 0 & 3 & 13 & 13 \\ 0 & 0 & 0 & -13 & -13 \end{array}\right) . \qquad (XIII.7)$$

Esta matriz puede transformarse ahora en su correspondiente sistema lineal $(XIII.4)$ y así obtener las soluciones x_1, x_2, x_3 y x_4.

El procedimiento descrito en este proceso se llama **eliminación Gaussiana con sustitución hacia atrás**. En un próximo capítulo consideraremos las condiciones bajo las cuales el método puede usarse con éxito para resolver el sistema lineal.

2. ALGEBRA LINEAL E INVERSION DE UNA MATRIZ

Esta sección se refiere al álgebra asociada con las matrices y la manera en que éstas pueden usarse para resolver problemas que involucran sistemas lineales.

Definición. Se dice que dos matrices A y B son **iguales** si son del mismo tamaño, digamos $m \times n$ y si $a_{ij} = b_{ij}$ para cada $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, n$.

Definición. Si A y B son matrices ambas $m \times n$, entonces la **suma** de A y B, denotada por $A + B$, es la matriz $m \times n$ cuyos elementos son $a_{ij} + b_{ij}$, para cada $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, n$.

Definición. Si A es una matriz $m \times n$ y λ es un número real, entonces el **producto escalar** de λ y A, denotado λA, es la matriz $m \times n$ cuyos elementos son λa_{ij}, para cada $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, n$.

Denotando la matriz que tiene todos sus elementos iguales a cero simplemente como O y como $-A$ la matriz cuyos elementos son $-a_{ij}$, podemos enumerar las siguientes propiedades generales de la adición y de la multiplicación escalar matricial. Estas propiedades son suficientes para clasificar el conjunto de todas las matrices $m \times n$ con elementos reales como un **espacio vectorial** sobre el campo de los números reales.

Teorema XIII.1

Sean A, B y C matrices $m \times n$ y λ y μ números reales. Se satisfacen las siguientes propiedades de la adición y multiplicación escalar:

a) $A + B = B + A$,
b) $(A + B) + C = A + (B + C)$,
c) $A + O = O + A = A$,
d) $A + (-A) = -A + A = O$,
e) $\lambda(A + B) = \lambda A + \lambda B$,
f) $(\lambda + \mu)\ A = \lambda A + \mu A$,
g) $\lambda(\mu A) = (\lambda\mu)A$,
h) $1A = A$.

Definición. Sean A una matriz $m \times n$ y B una matriz $n \times p$. El **producto matricial** de A y B, denotado por $A\ B$, es una matriz $m \times p$, cuyos elementos c_{ij} están dados por

$$c_{ij} = \sum_{k=1}^{n} a_{ik}\ b_{kj} = a_{i1}\ b_{1j} + a_{i2}\ b_{2j} + \ldots + a_{in}\ b_{nj}$$

para cada $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, p$.

Definición. Una matriz **diagonal** de orden n es una matriz $D = (d_{ij})$, $n \times n$, con la propiedad de que $d_{ij} = 0$ siempre que $i \neq j$. La **matriz identidad de orden n**, $I_n = (\delta_{ij})$, es la matriz diagonal con elementos

$$\delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} ;$$

Normalmente esta matriz se escribe simplemente como I.

Es bien conocido que la matriz identidad conmuta con una matriz A de orden n, es decir, el orden de la multiplicación no importa. Por otra parte, la propiedad conmutativa, $A\ B = B\ A$, no es generalmente cierta para la multiplicación matricial. Algunas de las propiedades relacionadas con la multiplicación de matrices, que sí se satisfacen, se presentan en el Teorema siguiente:

Teorema XIII.2

Sean A una matriz $n \times m$, B una matriz $m \times k$, C una matriz $k \times p$, D una matriz $m \times k$ y λ un número real. Se satisfacen las siguientes propiedades:

a) $A(B\ C) = (A\ B)C$,

b) $A(B + D) = A\ B + A\ D$,

c) $I_m B = B$, $B\ I_k = B$,

d) $\lambda(A\ B) = (\lambda A)B = A(\lambda B)$.

Un concepto fundamental del álgebra lineal que es muy útil para determinar la existencia y unicidad de soluciones de sistemas lineales es el **determinante** de una matriz $n \times n$. El único enfoque que se dará aquí para calcular el determinante será la definición recursiva. El determinante de una matriz A de denotará por "det A". Una **submatriz** de una matriz A es una matriz "extraída" de A suprimiendo algunas filas y/o columnas de A.

Definición.

a) Si $A = (a)$ es una matriz 1×1, entonces det $A = a$.

b) El **menor** M_{ij} es el determinante de la submatriz $(n-1) \times (n-1)$ de una matriz $n \times n$ de A obtenido suprimiendo la $i-$ésima fila y la $j-$ésima columna.

c) El **cofactor** A_{ij} asociado con M_{ij} se define como $A_{ij} = (-1)^{i+j}\ M_{ij}$.

d) El **determinante** de una matriz A, $n \times n$, donde $n > 1$ está dado ya sea por

$$\det A = \sum_{j=1}^{n} a_{ij} A_{ij} \qquad \text{para cualquier } i = 1, 2, \ldots, n \ , \tag{$XIII.8$}$$

o

$$\det A = \sum_{i=1}^{n} a_{ij} A_{ij} \qquad \text{para cualquier } j = 1, 2, \ldots, n \ . \tag{$XIII.9$}$$

Usando inducción matemática, se puede demostrar que, si $n > 1$, el uso de las definiciones dadas para calcular el determinante de una matriz, en general $n \times n$, requiere $n!$ multiplicaciones / divisiones y de $(n! - 1)$ sumas / restas. Incluso para valores relativamente pequeños de n, el número de cálculos puede llegar a ser inmanejable.

Teorema XIII.3

Sea A una matriz $n \times n$:

a) Si cualquier fila o columna de A tiene sólo componentes cero, entonces det $A = 0$.

b) Si $\tilde{A}$ se obtiene de A por medio de la operación $(E_i) \leftrightarrow (E_j)$, con $i \neq j$, entonces det $\tilde{A} = -$det A.

c) Si A tiene dos filas iguales, entonces $\det A = 0$.

d) Si $\tilde{A}$ se obtiene de A por medio de la operación $\lambda(E_i) \to (E_i)$, entonces $\det \tilde{A} = \lambda \det A$.

e) Si $\tilde{A}$ se obtiene de A por medio de la operación $(E_i + \lambda E_j) \to (E_j)$, con $i \neq j$, entonces $\det \tilde{A} = \det A$.

f) Si B es también una matriz $n \times n$ entonces $\det A\ B = \det A \det B$.

Definición. Se dice que una matriz A $n \times n$ es **no singular** si existe una matriz A^{-1}, $n \times n$, tal que $A\ A^{-1} = A^{-1}\ A = I$. La matriz A^{-1} se llama la **inversa** de A. Una matriz que no tiene inversa se llama **singular**.

Para encontrar un método para calcular A^{-1}, suponiendo su existencia, consideramos nuevamente la multiplicación matricial. Sea B_j la $j-$ésima columna de la matriz B $n \times n$. Realizaremos el producto

$$A\ B_j = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} b_{1j} \\ b_{2j} \\ \dots \\ b_{nj} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{n} a_{1k}\ b_{kj} \\ \sum_{k=1}^{n} a_{2k}\ b_{kj} \\ \dots \\ \sum_{k=1}^{n} a_{nk}\ b_{kj} \end{pmatrix} .$$

Si $A\ B = C$, entonces la $j-$ésima columna de C está dada por

$$C_j = \begin{pmatrix} c_{1j} \\ c_{2j} \\ \dots \\ c_{nj} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{n} a_{1k}\ b_{kj} \\ \sum_{k=1}^{n} a_{2k}\ b_{kj} \\ \dots \\ \sum_{k=1}^{n} a_{nk}\ b_{kj} \end{pmatrix} .$$

Por lo tanto, la $j-$ésima columna del producto $A\ B$ es el producto de A con la $j-$ésima columna de B. Supongamos que A^{-1} existe y que $A^{-1} = B = (b_{ij})$; entonces $A\ B = I$ y

$$A\ B_j = \begin{pmatrix} 0 \\ \dots \\ 0 \\ 1 \\ 0 \\ \dots \\ 0 \end{pmatrix} ,$$

donde el valor 1 aparece en la $j-$ésima fila. Para encontrar B debemos resolver n sistemas lineales en los cuales la $j-$ésima columna de la matriz inversa es la solución del sistema lineal con término independiente igual a la $j-$ésima columna de I.

Otra manera de calcular A^{-1} es relacionarla con el determinante de la matriz y con su adjunto.

Definición. Se define el **adjunto** de una matriz A, $n \times n$, como la matriz

$$A^{+} = \begin{pmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \dots & \dots & \dots & \dots \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{pmatrix} ,$$

donde A_{ij} son los cofactores (menores con signo) de los elementos correspondientes a_{ij} $(i, j = 1, 2, \dots, n)$. [Nótese que los adjuntos de los elementos de las filas de una matriz caen en las columnas correspondientes al adjunto, es decir, se verifica la operación de transposición].

Para encontrar la inversa de la matriz A, se dividen todos los elementos de la matriz adjunta A^{+} por el valor del determinante de A:

$$A^{-1} = \frac{1}{\det\, A}\; A^{+} \; .$$

Presentaremos ahora el resultado clave que relaciona a la no-singularidad, la eliminación Gaussiana, los sistemas lineales y los determinantes.

Teorema XIII.4

Para una matriz A $n \times n$ las siguientes afirmaciones son equivalentes:

a) La ecuación $A\ \mathbf{x} = \mathbf{0}$ tiene la única solución $\mathbf{x} = \mathbf{0}$.
b) El sistema lineal $A\ \mathbf{x} = \mathbf{b}$ tiene una solución única para cualquier vector columna $\mathbf{b}$ $n-$dimensional.
c) La matriz A es no singular, es decir, A^{-1} existe.
d) $\det A \neq 0$.
e) El algoritmo de la eliminación Gaussiana con intercambio de filas (que veremos más adelante) se puede aplicar al sistema lineal $A\ \mathbf{x} = \mathbf{b}$ para cualquier vector columna $\mathbf{b}$ $n-$dimensional.

Por medio de la definición de la multiplicación de matrices se puede discutir la relación entre los sistemas lineales y el álgebra lineal. El sistema lineal

$$\begin{aligned} E_1: &\quad a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n = b_1 \ , \\ E_2: &\quad a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n = b_2 \ , \\ &\quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \\ E_n: &\quad a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n = b_n \ , \end{aligned} \qquad (XIII.1)$$

puede verse como la ecuación matricial

$$A\ \mathbf{x} = \mathbf{b} \ , \qquad (XIII.10)$$

donde

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} , \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \qquad \text{y} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{pmatrix} \; .$$

El concepto de la **matriz inversa** de una matriz está también relacionado con los sistemas lineales, dado que multiplicando a la izquierda ambos miembros de $(XIII.10)$ por la matriz inversa A^{-1}, obtenemos

$$A^{-1}\ A\ \mathbf{x} = A^{-1}\ \mathbf{b}\ , \quad \text{o} \quad \mathbf{x} = A^{-1}\ \mathbf{b}\ , \tag{$XIII.11$}$$

que nos da la solución única del sistema $(XIII.1)$. Ese método es conocido como **regla de Cramer**. Dado que

$$A^{-1} = \frac{A^+}{\det\ A}\ ,$$

donde A^+ es el adjunto de A, se tiene que

$$\mathbf{x} = \frac{A^+}{\det\ A}\ \mathbf{b}\ , \quad \text{o} \quad \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} = \frac{1}{\det\ A}\ \begin{pmatrix} \Delta_1 \\ \Delta_2 \\ \dots \\ \Delta_n \end{pmatrix}\ , \tag{$XIII.12$}$$

donde

$$\Delta_i = \sum_{j=1}^{n} A_{ji}\ b_j = \det \begin{pmatrix} a_{11} & \dots & a_{1,i-1} & b_1 & a_{1,i+1} & \dots & a_{1n} \\ a_{21} & \dots & a_{2,i-1} & b_2 & a_{2,i+1} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{n,i-1} & b_n & a_{n,i+1} & \dots & a_{nn} \end{pmatrix}$$

son los determinantes obtenidos del determinante $\det\ A$ sustituyendo su $i-$ésima columna por la columna de términos constantes del sistema $(XIII.1)$. De la ecuación $(XIII.12)$ tenemos las **fórmulas de Cramer**:

$$x_1 = \frac{\Delta_1}{\det\ A}\ , \quad x_2 = \frac{\Delta_2}{\det\ A}\ , \quad \dots, \quad x_n = \frac{\Delta_n}{\det\ A}\ . \tag{$XIII.13$}$$

De este modo, si el determinante del sistema $(XIII.1)$ es distinto de cero, entonces el sistema tiene una solución única $\mathbf{x}$ definida por la fórmula matricial $(XIII.11)$ o por las fórmulas escalares $(XIII.13)$ equivalentes. Además, la solución de un sistema lineal como $(XIII.1)$ con n incógnitas se reduce a evaluar al $(n+1)-$ésimo determinante de orden n. Si n es grande, el cálculo de los determinantes es laborioso. Por esta razón, se han elaborado técnicas directas para hallar las raíces de un sistema lineal de ecuaciones.

3. TIPOS ESPECIALES DE MATRICES

Presentamos ahora material adicional sobre matrices. El primer tipo de matrices que consideraremos es el producido cuando se aplica eliminación Gaussiana a un sistema lineal.

Definición. Una matriz **triangular superior** U $n \times n$ tiene para cada j, los elementos $u_{ij} = 0$ para cada $i = j+1, j+2, \dots, n$; una matriz **triangular inferior** L $n \times n$ tiene para cada j, los elementos $l_{ij} = 0$ para cada $i = 1, 2, \dots, j-1$. (Una matriz **diagonal** es a la vez triangular superior e inferior). Es decir,

$$L = \begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{pmatrix}\ , \quad U = \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix}\ ,$$

y

$$D = \begin{pmatrix} d_{11} & 0 & 0 & 0 \\ 0 & d_{22} & 0 & 0 \\ 0 & 0 & d_{33} & 0 \\ 0 & 0 & 0 & d_{44} \end{pmatrix} .$$

El cálculo del determinante de una matriz arbitraria puede requerir un gran número de manipulaciones. Sin embargo, una matriz en forma triangular tiene un determinante fácil de calcular.

Teorema XIII.5

Si $A = (a_{ij})$ es una matriz $n \times n$ triangular superior (o triangular inferior o diagonal), entonces $\det A = \prod_{i=1}^{n} a_{ii}$.

Ejemplo. Reconsidereremos los ejemplos anteriores, en los cuales el sistema lineal

$$\begin{array}{lrcrcrcrcr} E_1: & x_1 & + & x_2 & & & + & 3\,x_4 & = & 4\,, \\ E_2: & 2\,x_1 & + & x_2 & - & x_3 & + & x_4 & = & 1\,, \\ E_3: & 3\,x_1 & - & x_2 & - & x_3 & + & 2\,x_4 & = & -3\,, \\ E_4: & -\,x_1 & + & 2\,x_2 & + & 3\,x_3 & - & x_4 & = & 4\,, \end{array}$$

fue reducido al sistema equivalente

$$\left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & 0 & 3 & 13 & 13 \\ 0 & 0 & 0 & -13 & -13 \end{array}\right) .$$

Sea U la matriz triangular superior de 4×4

$$U = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix} ,$$

la cual es el resultado de efectuar la eliminación Gaussiana a A. Para $i = 1, 2, 3$, definimos m_{ji} para cada $j = i+1, i+2, \ldots, 4$ como el número usado en el paso de eliminación $(E_j - m_{ji}E_i) \to E_j$; es decir $m_{21} = 2$, $m_{31} = 3$, $m_{41} = -1$, $m_{32} = 4$, $m_{42} = -3$ y $m_{43} = 0$. Si L se define como la matriz triangular inferior de 4×4 con elementos l_{ji} dados por

$$l_{ji} = \begin{cases} 0\,, & \text{cuando } i = 1, 2, \ldots, j-1\,, \\ 1\,, & \text{cuando } i = j\,, \\ m_{ji}\,, & \text{cuando } i = j+1, j+2, \ldots, n\,, \end{cases}$$

entonces

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix} ,$$

y es fácil verificar que

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{pmatrix} = A \ .$$

Los resultados de este ejemplo son ciertos en general y están dados en el Teorema siguiente.

Teorema XIII.6

Si el procedimiento de eliminación Gaussiana puede aplicarse al sistema $A\mathbf{x} = \mathbf{b}$ sin intercambio de fila, entonces la matriz A puede factorizarse como el producto de una matriz triangular inferior L con una matriz triangular superior U:

$$A = L\ U \ ,$$

donde $U = (u_{ij})$ y $L = (l_{ij})$ están definidas para cada j por:

$$u_{ij} = \begin{cases} a_{ij}^{(i)} \ , & \text{cuando } i = 1, 2, \ldots, j \ , \\ 0 \ , & \text{cuando } i = j+1, j+2, \ldots, n \ , \end{cases}$$

y

$$l_{ij} = \begin{cases} 0 \ , & \text{cuando } i = 1, 2, \ldots, j-1 \ , \\ 1 \ , & \text{cuando } i = j \ , \\ m_{ij} \ , & \text{cuando } i = j+1, j+2, \ldots, n \ , \end{cases}$$

donde $a_{ij}^{(i)}$ es el elemento i, j de la matriz final obtenida por el método de eliminación Gaussiana y m_{ij} es el multiplicador.

Si se tienen que efectuar intercambios de filas para que el procedimiento funcione, entonces A se puede factorizar como $L\ U$, donde U es la misma que en el Teorema XIII.6, pero en general, L no será triangular inferior.

El problema de calcular el determinante de una matriz se puede simplificar reduciendo primero la matriz a forma triangular y después usando el Teorema XIII.5 para encontrar el determinante de una matriz triangular.

Definición. La **traspuesta** de una matriz A $m \times n$, denotada por A^t, es una matriz $n \times m$ cuyos elementos son $(A^t)_{ij} = (A)_{ji}$. Una matriz cuya traspuesta es ella misma se llama **simétrica**.

Teorema XIII.7

Las siguientes operaciones que involucran a la traspuesta de una matriz se satisfacen siempre que la operación sea posible:

1. $(A^t)^t = A$,
2. $(A+B)^t = A^t + B^t$,
3. $(A\ B)^t = B^t\ A^t$,
4. si A^{-1} existe, $(A^{-1})^t = (A^t)^{-1}$,
5. $\det A^t = \det A$.

Definición. Una matriz $n \times n$ se llama una **matriz banda** si existen enteros p y q, $1 < p,\ q < n$, con la propiedad de que $a_{ij} = 0$ siempre que $i+p \leq j$ ó $j+q \leq i$. El **ancho de banda** para una matriz de este tipo se define como $w = p + q - 1$.

La definición de matriz de banda fuerza a estas matrices a concentrar todos sus elementos no cero alrededor de la diagonal. Dos casos especiales de matrices de banda que ocurren frecuentemente en la práctica son $p = q = 2$ y $p = q = 4$. Las matrices con ancho de banda 3 (que se presenta cuando $p = q = 2$) se llaman generalmente **tridiagonales** ya que tienen la forma

$$A = \begin{pmatrix} a_{11} & a_{12} & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ a_{21} & a_{22} & a_{23} & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & a_{32} & a_{33} & a_{34} & 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & a_{i,i-1} & a_{ii} & a_{i,i+1} & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & a_{n-2,n-1} & a_{n-1,n-1} & a_{n-1,n} \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & a_{n,n-1} & a_{nn} \end{pmatrix}$$

Definición. Se dice que la matriz A de orden n es **estrictamente dominante diagonalmente** en el caso de que satisfaga

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \ ,$$

para cada $i = 1, 2, \dots, n$.

Teorema XIII.8

Si A es una matriz $n \times n$ estrictamente dominante diagonalmente, entonces A es no singular. Además, se puede efectuar eliminación Gaussiana en cualquier sistema lineal de la forma $A\ \mathbf{x} = \mathbf{b}$ para obtener su solución única sin intercambios de filas o columnas, y los cálculos son estables con respecto al crecimiento de los errores de redondeo.

La última clase especial de matrices que se discutirá en esta sección se llama positiva definida.

Definición. Una matriz simétrica A $n \times n$ se llama **positiva definida** si $\mathbf{x}^t\ A\ \mathbf{x} > 0$ para todo vector columna $n-$dimensional $\mathbf{x} \neq \mathbf{0}$,

$$\mathbf{x}^t\ A\ \mathbf{x} = (x_1, x_2, \dots, x_n) \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} =$$

$$= (x_1, x_2, \ldots, x_n) \begin{pmatrix} \sum_{j=1}^{n} a_{1j}\ x_j \\ \sum_{j=1}^{n} a_{2j}\ x_j \\ \ldots \\ \sum_{j=1}^{n} a_{nj}\ x_j \end{pmatrix} = (\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\ x_i\ x_j)\ .$$

Teorema XIII.9

Si A es una matriz $n \times n$ positiva definida, entonces A es no singular. Además, se puede efectuar eliminación Gaussiana en cualquier sistema lineal de la forma $A\ \mathbf{x} = \mathbf{b}$ para obtener su solución única sin intercambios de filas o columnas, y los cálculos son estables con respecto al crecimiento de los errores de redondeo.

4. NORMAS DE VECTORES Y MATRICES

Sea $\mathcal{R}^n$ el conjunto de todos los vectores columna con componentes reales. Para definir una distancia en $\mathcal{R}^n$, usaremos la idea de la **norma** de un vector.

Definición. Una **norma vectorial** en $\mathcal{R}^n$ es una función $||\ \cdot\ ||$, de $\mathcal{R}^n$ en $\mathcal{R}$ con las siguientes propiedades:

- $||\mathbf{x}|| \geq 0$ para todo $\mathbf{x} \in \mathcal{R}^n$;
- $||\mathbf{x}|| = 0$ si y sólo si $\mathbf{x} = (0, 0, \ldots, 0)^t \equiv \mathbf{0}$;
- $||\alpha\ \mathbf{x}|| = |\alpha|\ ||\mathbf{x}||$ para todo $\alpha \in \mathcal{R}$ y $\mathbf{x} \in \mathcal{R}^n$;
- $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ para todo $\mathbf{x}, \mathbf{y} \in \mathcal{R}^n$.

Para nuestros propósitos sólo necesitaremos tres normas específicas en $\mathcal{R}^n$.

Definición. Las normas l_1, l_2 y l_∞ para el vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ se definen como

$$||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|\ , \qquad ||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \quad \text{y} \quad ||\mathbf{x}||_\infty = \max_{1 \leq i \leq n} |x_i|\ .$$

La norma l_2 se denomina frecuentemente **norma Euclideana** del vector $\mathbf{x}$ ya que representa la noción usual de distancia al origen en el caso en el que $\mathbf{x}$ esté en $\mathcal{R}$, $\mathcal{R}^2$ o $\mathcal{R}^3$.

Ya que la norma de un vector da una medida de la distancia entre el vector y el origen, la distancia entre dos vectores se puede definir como la norma de la diferencia de los dos vectores.

Definición. Si $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ e $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$ son vectores en $\mathcal{R}^n$, las distancias l_1, l_2 y l_∞ entre $\mathbf{x}$ e $\mathbf{y}$ se definen como:

$$||\mathbf{x} - \mathbf{y}||_1 = \sum_{i=1}^{n} |x_i - y_i|\ ,$$

$$||\mathbf{x} - \mathbf{y}||_2 = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \quad \text{y} \quad ||\mathbf{x} - \mathbf{y}||_\infty = \max_{1 \leq i \leq n} |x_i - y_i|\ .$$

El concepto de distancia en $\mathcal{R}^n$ puede usarse también para definir el límite de una sucesión de vectores en este espacio.

Definición. Se dice que una sucesión $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ de vectores en $\mathcal{R}^n$ **converge** a $\mathbf{x}$ con respecto a la norma $||\ \cdot\ ||$ si, dado cualquier $\varepsilon > 0$, existe un entero $N(\varepsilon)$ tal que

$$||\mathbf{x}^{(k)} - \mathbf{x}|| < \varepsilon \qquad \text{para toda} \quad k \geq N(\varepsilon)\ .$$

Teorema XIII.10

La sucesión de vectores $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ converge a $\mathbf{x}$ en $\mathcal{R}^n$ con respecto a $||\ \cdot\ ||_\infty$ si y sólo si $\lim\limits_{k\to\infty} x_i^{(k)} = x_i$ para cada $i = 1, 2, \ldots, n$.

Teorema XIII.11

Para cada $\mathbf{x} \in \mathcal{R}^n$,

$$||\mathbf{x}||_\infty \leq ||\mathbf{x}||_2 \leq \sqrt{n}\ ||\mathbf{x}||_\infty\ .$$

Demostración: sea x_j una coordenada de $\mathbf{x}$ tal que $||\mathbf{x}||_\infty = \max\limits_{1\leq i\leq n} |x_i| = |x_j|$. Entonces

$$||\mathbf{x}||_\infty^2 = |x_j|^2 = x_j^2 \leq \sum_{i=1}^{n} x_i^2 \leq \sum_{i=1}^{n} x_j^2 = nx_j^2 = n||\mathbf{x}||_\infty^2\ .$$

Por lo tanto

$$||\mathbf{x}||_\infty \leq \Big[\sum_{i=1}^{n} x_i^2\Big]^{1/2} = ||\mathbf{x}||_2 \leq \sqrt{n}||\mathbf{x}||_\infty\ .$$

c.q.d.

Se puede demostrar que todas las normas en $\mathcal{R}^n$ son equivalentes con respecto a la convergencia; es decir, si $||\ \cdot\ ||$ y $||\ \cdot\ ||'$ son dos normas cualesquiera en $\mathcal{R}^n$ y $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ tiene el límite $\mathbf{x}$ con respecto a $||\ \cdot\ ||$, entonces $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ tiene el límite $\mathbf{x}$ con respecto a $||\ \cdot\ ||'$

Es necesario también tener un método para medir distancias entre dos matrices $n\times n$, lo cual nuevamente requiere el uso del concepto de norma.

Definición. Una **norma matricial** en el conjunto de todas las matrices reales $n \times n$ es una función de valores reales $||\ \cdot\ ||$, definida en este conjunto que satisface, para todas las matrices A y B $n \times n$ y todo número real α:

- $||A|| \geq 0$;
- $||A|| = 0$ si y sólo si $A = O$;
- $||\alpha\ A|| = |\alpha|\ ||A||$;
- $||A + B|| \leq ||A|| + ||B||$;
- $||A \cdot B|| \leq ||A|| \cdot ||B||$.

Una **distancia entre las matrices** A y B $n \times n$ se puede definir de la manera usual como $||A - B||$. Aún cuando las normas de las matrices pueden obtenerse de varias

maneras, las únicas normas que consideraremos son aquellas que son una consecuencia natural de las normas vectoriales l_1, l_2 y l_∞.

Teorema XIII.12

Si $||\ \cdot\ ||$ es cualquier norma vectorial en $\mathcal{R}^n$, entonces

$$||A|| = \max_{||\mathbf{x}||=1} ||A\mathbf{x}||$$

define una norma matricial en el conjunto de las matrices reales $n \times n$, que se llama la **norma natural**.

Consecuentemente, las normas matriciales que consideraremos tienen las formas

$$||A||_1 = \max_{||\mathbf{x}||_1=1} ||A\mathbf{x}||_1 \ , \quad \text{norma } \ l_1 \ ,$$

$$||A||_2 = \max_{||\mathbf{x}||_2=1} ||A\mathbf{x}||_2 \ , \quad \text{norma } \ l_2 \ ,$$

y

$$||A||_\infty = \max_{||\mathbf{x}||_\infty=1} ||A\mathbf{x}||_\infty \ , \quad \text{norma } \ l_\infty \ .$$

Teorema XIII.13

Si $A = (a_{ij})$ es una matriz $n \times n$, entonces

$$a) \qquad ||A||_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ ,$$

$$b) \qquad ||A||_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}| \ .$$

Demostración: a) sea $\mathbf{x}$ un vector columna $n-$dimensional tal que su norma l_∞ sea uno; es decir, $||\mathbf{x}||_\infty = \max\limits_{1\le i\le n} |x_i| = 1$. Como $A\mathbf{x}$ es también un vector columna $n-$dimensional,

$$||A\mathbf{x}||_\infty = \max_{1\le i\le n} |(A\ x)_i| = \max_{1\le i\le n} |\sum_{j=1}^{n} a_{ij}\ x_j| \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ \max_{1\le j\le n} |x_j|$$

$$= \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ ||\mathbf{x}||_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ .$$

Así que $||A\mathbf{x}||_\infty \le \max\limits_{1\le i\le n} \sum\limits_{j=1}^{n} |a_{ij}|$ para toda $\mathbf{x}$ con $||\mathbf{x}||_\infty = 1$. Consecuentemente,

$$||A||_\infty = \max_{||\mathbf{x}||_\infty=1} ||A\mathbf{x}||_\infty \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ .$$

Por otro lado, si p es el entero $1 \le p \le n$, con

$$\sum_{j=1}^{n} |a_{pj}| = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \ ,$$

y $\mathbf{x}$ se escoge de tal manera que

$$x_j = \begin{cases} 1\ , & \text{si } a_{pj} \geq 0, \\ -1\ , & \text{si } a_{pj} < 0, \end{cases}$$

entonces $||\mathbf{x}||_\infty = 1$ y $|a_{pj}\ x_j| = |a_{pj}|$ para toda $j = 1, 2, \ldots, n$. Además,

$$||A\mathbf{x}||_\infty = \max_{1\leq i\leq n} |\sum_{j=1}^{n} a_{ij}\ x_j| \geq |\sum_{j=1}^{n} a_{pj}\ x_j| = \sum_{j=1}^{n} |a_{pj}| = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|\ .$$

Esto implica que

$$||A||_\infty = \max_{||\mathbf{x}||_\infty = 1} ||A\mathbf{x}||_\infty \geq \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|\ .$$

Entonces,

$$||A||_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|\ .$$

Demostremos ahora la parte b); sea $\mathbf{x}$ un vector columna $n-$dimensional tal que su norma l_1 sea uno; es decir, $||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i| = 1$. Como $A\mathbf{x}$ es también un vector columna $n-$dimensional,

$$\begin{aligned} ||A\mathbf{x}||_1 = &\sum_{i=1}^{n} |(A\ x)_i| = \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{ij}\ x_j| = \sum_{j=1}^{n} |(\sum_{i=1}^{n} a_{ij})\ x_j| \leq \\ \leq &\sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}| = \sum_{i=1}^{n} |a_{ij}|\ ||\mathbf{x}||_1 = \sum_{i=1}^{n} |a_{ij}|\ . \end{aligned}$$

Así que $||A\mathbf{x}||_1 \leq \sum_{i=1}^{n} |a_{ij}|$ para toda $\mathbf{x}$ con $||\mathbf{x}||_1 = 1$. Consecuentemente,

$$||A||_1 = \max_{||\mathbf{x}||_1 = 1} ||A\mathbf{x}||_1 \leq \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|\ .$$

Por otro lado, si p es el entero $1 \leq p \leq n$, con

$$\sum_{i=1}^{n} |a_{ip}| = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|\ ,$$

y $\mathbf{x}$ se escoge de tal manera que

$$x_j = \begin{cases} 1\ , & \text{si } j = p, \\ 0\ , & \text{en el resto de los casos}, \end{cases}$$

entonces $||\mathbf{x}||_1 = \sum_{j=1}^{n} |x_j| = 1$. Además,

$$||A\mathbf{x}||_1 = \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{ij}\ x_j| \geq \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{ip}\ x_j| = \sum_{i=1}^{n} |a_{ip}| \sum_{j=1}^{n} |x_j| = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|\ .$$

Esto implica que

$$||A\mathbf{x}||_1 = \max_{||\mathbf{x}||_1=1} ||A\mathbf{x}||_1 \geq \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \ .$$

Entonces,

$$||A\mathbf{x}||_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \ .$$

c.q.d.

Para investigar la norma l_2, es necesario discutir algunos conceptos adicionales del álgebra lineal.

Definición. Si A es una matriz real $n \times n$, el polinomio definido por

$$p(\lambda) = \det(A - \lambda \ I)$$

se llama **polinomio característico** de A.

Es fácil demostrar que $p(\lambda)$ es un polinomio de grado n con coeficientes reales y consecuentemente, tiene a lo más n ceros distintos, algunos de los cuales pueden ser complejos. Si λ es un cero de $p(\lambda)$, entonces debido a que $\det(A - \lambda \ I) = 0$, el Teorema XIII.4 implica que el sistema lineal definido por $(A - \lambda \ I) \ \mathbf{x} = \mathbf{0}$ tiene una solución diferente de la solución idénticamente cero (ó solución trivial). Deseamos estudiar los ceros de $p(\lambda)$ y las soluciones no triviales correspondientes de estos sistemas.

Definición. Si $p(\lambda)$ es el polinomio característico de la matriz A los ceros de $p(\lambda)$ se llaman **autovalores** (también llamados valores propios o valores característicos) de la matriz A. Si λ es un valor característico de A y $\mathbf{x} \neq \mathbf{0}$ tiene la propiedad de que $(A - \lambda \ I) \ \mathbf{x} = \mathbf{0}$, entonces $\mathbf{x}$ es el **autovector** (también llamado vector propio o vector característico) de A correspondiente al autovalor λ.

Definición. El **radio espectral** $\rho(A)$ de una matriz A se define como

$$\rho(A) = \max |\lambda|$$

donde λ es un valor característico de A.

El radio espectral está relacionado con la norma de una matriz, como muestra el siguiente Teorema.

Teorema XIII.14

Si $A = (a_{ij})$ es una matriz real $n \times n$, entonces

i) $[\rho(A^t \ A)]^{1/2} = ||A||_2$;

ii) $\rho(A) \leq ||A||$ para cualquier norma natural $|| \ \cdot \ ||$.

Un resultado útil e interesante es que para cualquier matriz A y cualquier $\varepsilon > 0$, existe una norma $|| \ \cdot \ ||$ con la propiedad de que $||A|| < \rho(A) + \varepsilon$. Consecuentemente, $\rho(A)$ es la máxima cota inferior para las normas de A.

En el estudio de las técnicas iterativas de matrices, es de particular importancia saber cuándo las potencias de una matriz se hacen pequeñas, es decir, cuándo todas las componentes tienden a cero. Las matrices de este tipo se denominan **convergentes**.

Definición. Llamamos a A $n \times n$ una matriz **convergente** si

$$\lim_{k\to\infty} (A^k)_{ij} = 0$$

para cada $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, n$.

Existe una conexión importante entre el radio espectral de una matriz y su convergencia.

Teorema XIII.15

Las siguientes afirmaciones son equivalentes:

1. A es una matriz convergente;
2. $\lim_{n\to\infty} ||A^n|| = 0$, para alguna norma natural $||\ \cdot\ ||$;
3. $\rho(A) \leq 1$;
4. $||A|| \leq 1$;
5. $\lim_{n\to\infty} A^n\ \mathbf{x} = \mathbf{0}$, para toda $\mathbf{x}$.

CAPITULO XIV. ELIMINACION GAUSSIANA Y SUSTITUCION HACIA ATRAS

1. INTRODUCCION Y METODO

El procedimiento general de eliminación Gaussiana aplicado al sistema

$$\begin{aligned} E_1: &\quad a_{11}\,x_1 + a_{12}\,x_2 + \ldots + a_{1n}\,x_n = b_1\ , \\ E_2: &\quad a_{21}\,x_1 + a_{22}\,x_2 + \ldots + a_{2n}\,x_n = b_2\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad a_{n1}\,x_1 + a_{n2}\,x_2 + \ldots + a_{nn}\,x_n = b_n\ , \end{aligned} \tag{XIV.1}$$

se maneja de una manera similar al procedimiento seguido en el ejemplo del Capítulo XIII. Formamos la matriz ampliada A_a:

$$A_a = [A, \mathbf{b}] = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & a_{1,n+1} \\ a_{21} & a_{22} & \ldots & a_{2n} & a_{2,n+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & a_{n,n+1} \end{array}\right)\ , \tag{XIV.2}$$

donde A denota la matriz formada por los coeficientes y los elementos en la $(n+1)$-ésima columna son los valores de $\mathbf{b}$, es decir, $a_{i,n+1} = b_i$ para cada $i = 1, 2, \ldots, n$. Siempre y cuando $a_{11} \neq 0$, se efectúan las operaciones correspondientes a $(E_j - (a_{j1}/a_{11})E_1) \rightarrow (E_j)$ para cada $j = 2, 3, \ldots, n$ para eliminar el coeficiente de x_1 en cada una de estas filas. Aún cuando se espera que los elementos de las filas $2, 3, \ldots, n$ cambien, para facilitar la notación, denotaremos nuevamente el elemento en la $i-$ésima fila y en la $j-$ésima columna por a_{ij}. Teniendo en cuenta esto, seguiremos un procedimiento secuencial para $i = 2, 3, \ldots, n-1$ y realizamos la operación $(E_j - (a_{ji}/a_{ii})E_i) \rightarrow (E_j)$ para cada $j = i+1, i+2, \ldots, n$, siempre que $a_{ii} \neq 0$. Esto eliminará x_i en cada fila debajo de la $i-$ésima para todos los valores de $i = 1, 2, \ldots, n-1$. La matriz resultante tendrá la forma:

$$A_a^{(f)} = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & a_{1,n+1} \\ 0 & a_{22} & \ldots & a_{2n} & a_{2,n+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & a_{nn} & a_{n,n+1} \end{array}\right)\ .$$

Esta matriz representa un sistema lineal con el mismo conjunto de soluciones que el sistema (XIV.1). Como el sistema lineal equivalente es triangular:

$$\begin{array}{ccccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & \ldots & \ldots & + & a_{1n}\,x_n & = & a_{1,n+1}\ , \\ & & a_{22}\,x_2 & + & \ldots & \ldots & + & a_{2n}\,x_n & = & a_{2,n+1}\ , \\ & & & & \ldots & \ldots & \ldots & \ldots & = & \ldots \\ & & & & a_{n-1,n-1}\,x_{n-1} & & + & a_{n-1,n}\,x_n & = & a_{n-1,n+1}\ , \\ & & & & & & & a_{nn}\,x_n & = & a_{n,n+1}\ , \end{array}$$

se puede realizar la sustitución hacia atrás. Resolviendo la $n-$ésima ecuación para x_n se obtiene:

$$x_n = \frac{a_{n,n+1}}{a_{nn}}\ .$$

Resolviendo la ecuación $(n-1)$-ésima para x_{n-1} y usando x_n obtenemos:

$$x_{n-1} = \frac{(a_{n-1,n+1} - a_{n-1,n}\ x_n)}{a_{n-1,n-1}}\ .$$

Y continuando con este proceso, llegamos a que

$$x_i = \frac{(a_{i,n+1} - a_{in}\ x_n - a_{i,n-1}\ x_{n-1} - \ldots - a_{i,i+1}\ x_{i+1})}{a_{ii}} =$$

$$= \frac{(a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}\ x_j)}{a_{ii}}\ ,$$

para cada $i = n-1, n-2, \ldots, 2, 1$.

El procedimiento de eliminación Gaussiana se puede mostrar más detalladamente, aunque de forma más complicada, formando una secuencia de matrices ampliadas $A_a^{(1)}$, $A_a^{(2)}$, ..., $A_a^{(n)}$, donde $A_a^{(1)}$ es la matriz A_a dada en la ecuación $(XIV.2)$ y $A_a^{(k)}$ con $k = 2, 3, \ldots, n$ tiene los elementos $a_{ij}^{(k)}$ de la forma:

$$a_{ij}^{(k)} = \begin{cases} a_{ij}^{(k-1)} & \text{cuando } i = 1, 2, \ldots, k-1 \\ & \text{y } j = 1, 2, \ldots, n+1\ , \\ 0 & \text{cuando } i = k, k+1, \ldots, n \\ & \text{y } j = 1, 2, \ldots, k-1\ , \\ a_{ij}^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}}\ a_{k-1,j}^{(k-1)} & \text{cuando } i = k, k+1, \ldots, n \\ & \text{y } j = k, k+1, \ldots, n+1\ . \end{cases}$$

$$A_a^{(k)} = \left(\begin{array}{ccccccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \ldots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & \ldots & a_{1n}^{(1)} & | & a_{1,n+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & \ldots & a_{2n}^{(2)} & | & a_{2,n+1}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \ldots & a_{3,k-1}^{(3)} & a_{3k}^{(3)} & \ldots & a_{3n}^{(3)} & | & a_{3,n+1}^{(3)} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & | & \ldots \\ 0 & \ldots & \ldots & 0 & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \ldots & a_{k-1,n}^{(k-1)} & | & a_{k-1,n+1}^{(k-1)} \\ 0 & \ldots & \ldots & \ldots & 0 & a_{kk}^{(k)} & \ldots & a_{kn}^{(k)} & | & a_{k,n+1}^{(k)} \\ 0 & \ldots & \ldots & \ldots & 0 & \ldots & \ldots & \ldots & | & \ldots \\ 0 & \ldots & \ldots & \ldots & 0 & a_{nk}^{(k)} & \ldots & a_{nn}^{(k)} & | & a_{n,n+1}^{(k)} \end{array}\right)\ ,$$

es la matriz que representa el sistema lineal equivalente para el cual la variable x_{k-1} acaba de ser eliminada de las ecuaciones $E_k, E_{k+1}, \ldots, E_n$.

El procedimiento no funcionará si alguno de los elementos $a_{11}^{(1)}$, $a_{22}^{(2)}$, ..., $a_{n-1,n-1}^{(n-1)}$, $a_{nn}^{(n)}$ es cero, ya que en este caso el paso $(E_i - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\ E_k) \rightarrow E_i$ no se puede realizar (esto ocurre si una de las $a_{11}^{(1)}, a_{22}^{(2)}, a_{33}^{(3)}, \ldots, a_{n-1,n-1}^{(n-1)}$ es cero), o la sustitución hacia atrás no se puede llevar a cabo (en el caso $a_{nn}^{(n)}$). Esto no significa que el sistema lineal no sea resoluble, sino que la técnica de resolución debe alterarse. Cuando $a_{kk}^{(k)} = 0$ para algún $k = 1, 2, \ldots, n-1$, se busca en la $k-$ésima columna de $A_a^{(k-1)}$ desde la fila k hasta la n para encontrar el primer elemento diferente de cero. Si $a_{pk}^{(k)} \neq 0$ para algún p,

$k+1 \leq p \leq n$, entonces se efectúa la operación $(E_k) \leftrightarrow (E_p)$ para obtener $A_a^{(k-1)}$. El procedimiento puede continuar entonces para formar $A_a^{(k)}$, y así proseguir. Si $a_{pk}^{(k)} = 0$ para $p = k, k+1, \ldots, n$, se puede demostrar (Teorema XIII.4) que el sistema lineal no tiene una solución única y el procedimiento se para. Finalmente, si $a_{nn}^{(n)} = 0$ el sistema lineal no tiene una solución única y el procedimiento se para.

El ejemplo siguiente ilustra el funcionamiento de este método:

Ejemplo. Resolver el sistema de ecuaciones:

$$\begin{array}{lrcrcrcrcr} E_1: & x_1 & - & x_2 & + & 2\,x_3 & - & x_4 & = & -8\,, \\ E_2: & 2\,x_1 & - & 2\,x_2 & + & 3\,x_3 & - & 3\,x_4 & = & -20\,, \\ E_3: & x_1 & + & x_2 & + & x_3 & & & = & -2\,, \\ E_4: & x_1 & - & x_2 & + & 4\,x_3 & + & 3\,x_4 & = & 4\,. \end{array}$$

La matriz ampliada es

$$A_a = A_a^{(1)} = \left(\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 2 & -2 & 3 & -3 & -20 \\ 1 & 1 & 1 & 0 & -2 \\ 1 & -1 & 4 & 3 & 4 \end{array}\right)\,,$$

y efectuando las operaciones $(E_2 - 2E_1) \to (E_2)$, $(E_3 - E_1) \to (E_3)$ y $(E_4 - E_1) \to (E_4)$ llegamos a:

$$A_a^{(2)} = \left(\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right)\,.$$

Como el elemento $a_{22}^{(2)}$, llamado **elemento de pivote**, es cero, el procedimiento no puede continuar de la misma forma, pero la operación $(E_i) \leftrightarrow (E_j)$ está permitida, así que se hace una búsqueda de los elementos $a_{32}^{(2)}$ y $a_{42}^{(2)}$ para encontrar el primer elemento no cero. Ya que $a_{32}^{(2)} \neq 0$, se efectúa la operación $(E_2) \leftrightarrow (E_3)$ para obtener una nueva matriz

$$A_a^{(2)'} = \left(\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right)\,.$$

Como x_2 está ya eliminada de E_3 y E_4, $A_a^{(3)}$ será $A_a^{(2)'}$ y los cálculos pueden continuar con la operación $(E_4 + 2E_3) \to (E_4)$, dando

$$A_a^{(4)} = \left(\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 0 & 0 & 2 & 4 \end{array}\right)\,.$$

Finalmente, se puede aplicar la sustitución hacia atrás:

$$x_4 = \frac{4}{2} = 2\,, \qquad x_3 = \frac{[-4 - (-1)\,x_4]}{-1} = 2\,,$$
$$x_2 = \frac{[6 - x_4 - (-1)\,x_3]}{2} = 3\,, \qquad x_1 = \frac{[-8 - (-1)\,x_4 - 2\,x_3 - (-1)\,x_2]}{1} = -7\,.$$

2. ALGORITMO Y EJEMPLOS

Para resumir el método de eliminación Gaussiana completo con sustitución hacia atrás, se presenta el siguiente algoritmo.

Algoritmo de eliminación Gaussiana con sustitución hacia atrás.

==========

Para resolver el sistema lineal de $n \times n$:

$$\begin{aligned} E_1: &\quad a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n = a_{1,n+1} \\ E_2: &\quad a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n = a_{2,n+1} \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n = a_{n,n+1} \end{aligned}$$

Entrada: número de incógnitas y de ecuaciones n; matriz ampliada $A_a = (a_{ij})$ donde $1 \le i \le n$ y $1 \le j \le n+1$.

Salida: solución $x_1, x_2, \ldots, x_n$ ó mensaje de que el sistema lineal no tiene solución única.

Paso 1: Para $i = 1, 2, \ldots, n-1$ seguir los pasos 2–4 *(proceso de eliminación)*.

Paso 2: Sea p el menor entero con $i \le p \le n$ y $a_{pi} \neq 0$. Si p no puede encontrarse entonces SALIDA; *(no existe solución única)* PARAR.

Paso 3: Si $p \neq i$ entonces efectuar $(E_p) \leftrightarrow (E_i)$.

Paso 4: Para $j = i+1, i+2, \ldots, n$ seguir los pasos 5 y 6.

Paso 5: Tomar $m_{ji} = \frac{a_{ji}}{a_{ii}}$.

Paso 6: Efectuar $(E_j - m_{ji}\ E_i) \to (E_j)$.

Paso 7: Si $a_{nn} = 0$ entonces SALIDA; *(no existe solución única)* PARAR.

Paso 8: *(Empieza la sustitución hacia atrás)*; tomar

$$x_n = \frac{a_{n,n+1}}{a_{nn}}\ .$$

Paso 9: Para $i = n-1, n-2, \ldots, 1$ tomar

$$x_i = \frac{a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}\ x_j}{a_{ii}}\ .$$

Paso 10: SALIDA $(x_1, x_2, \ldots, x_n)$;
(procedimiento completado satisfactoriamente) PARAR.

==========

Ejemplo. Resolver los dos sistemas lineales:

$$\begin{array}{lrcrcrcrcr} E_{1,(1)}: & x_1 & + & x_2 & + & x_3 & + & x_4 & = & 7\ , \\ E_{2,(1)}: & x_1 & + & x_2 & & & + & 2\,x_4 & = & 8\ , \\ E_{3,(1)}: & 2\,x_1 & + & 2\,x_2 & + & 3\,x_3 & & & = & 10\ , \\ E_{4,(1)}: & -\,x_1 & - & x_2 & - & 2\,x_3 & + & 2\,x_4 & = & 0\ , \end{array}$$

y

$$\begin{array}{lrcrcrcrcr} E_{1,(2)}: & x_1 & + & x_2 & + & x_3 & + & x_4 & = & 7\,, \\ E_{2,(2)}: & x_1 & + & x_2 & & & + & 2\,x_4 & = & 5\,, \\ E_{3,(2)}: & 2\,x_1 & + & 2\,x_2 & + & 3\,x_3 & & & = & 10\,, \\ E_{4,(2)}: & -\,x_1 & - & x_2 & - & 2\,x_3 & + & 2\,x_4 & = & 0\,. \end{array}$$

Estos sistemas dan lugar a las matrices

$$A^{(1)}_{a(1)} = \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 1 & 1 & 0 & 2 & 8 \\ 2 & 2 & 3 & 0 & 10 \\ -1 & -1 & -2 & 2 & 0 \end{array}\right) \quad \text{y} \quad A^{(1)}_{a(2)} = \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 1 & 1 & 0 & 2 & 5 \\ 2 & 2 & 3 & 0 & 10 \\ -1 & -1 & -2 & 2 & 0 \end{array}\right)\,.$$

Ya que $a_{11} = 1 \neq 0$, los pasos para eliminar x_1 de E_2, E_3 y E_4 dan, para $i = 1$

$$m_{ji} = m_{j1} = \frac{a_{j1}}{a_{11}} = \frac{a_{j1}}{1} = a_{j1}\,.$$

Entonces:

$$j = 2,\;\; m_{21} = 1\,; \qquad j = 3,\;\; m_{31} = 2\,; \qquad j = 4,\;\; m_{41} = -1\,;$$

y las operaciones a efectuar son:

$$(E_2 - E_1) \to (E_2)\,; \qquad (E_3 - 2E_1) \to (E_3)\,; \qquad (E_4 + E_1) \to (E_4)\,.$$

Las matrices se trasforman en:

$$A^{(2)}_{a(1)} = \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 1 & -2 & -4 \\ 0 & 0 & -1 & 3 & 7 \end{array}\right) \quad \text{y} \quad A^{(2)}_{a(2)} = \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 0 & 0 & -1 & 1 & -2 \\ 0 & 0 & 1 & -2 & -4 \\ 0 & 0 & -1 & 3 & 7 \end{array}\right)\,.$$

Aquí $a_{22} = a_{32} = a_{42} = 0$ y el algoritmo requiere que el procedimiento se detenga y no se obtiene una solución para ninguno de los sistemas.

Para examinar más de cerca la razón de la dificultad, efectuamos $(E_4 + E_3) \to (E_4)$ para obtener $A^{(3)}_{a(1)} = A^{(4)}_{a(1)}$ y $A^{(3)}_{a(2)} = A^{(4)}_{a(2)}$

$$A^{(4)}_{a(1)} = \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 1 & -2 & -4 \\ 0 & 0 & 0 & 1 & 3 \end{array}\right) \quad \text{y} \quad A^{(4)}_{a(2)} = \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 7 \\ 0 & 0 & -1 & 1 & -2 \\ 0 & 0 & 1 & -2 & -4 \\ 0 & 0 & 0 & 1 & 3 \end{array}\right)\,.$$

Escribiendo las ecuaciones para cada sistema se obtiene:

$$\begin{array}{rcrcrcrcr} x_1 & + & x_2 & + & x_3 & + & x_4 & = & 7\,, \\ & & & - & x_3 & + & x_4 & = & 1\,, \\ & & & & x_3 & - & 2\,x_4 & = & -4\,, \\ & & & & & & x_4 & = & 3\,, \end{array}$$

y

$$\begin{array}{rcrcrcrcr} x_1 & + & x_2 & + & x_3 & + & x_4 & = & 7\,, \\ & & & - & x_3 & + & x_4 & = & -2\,, \\ & & & & x_3 & - & 2\,x_4 & = & -4\,, \\ & & & & & & x_4 & = & 3\,. \end{array}$$

Efectuando sustitución hacia atrás en cada sistema nos lleva a:

$$x_4 = 3 \quad \text{y} \quad x_3 = -4 + 2\,x_4 = 2\,,$$

en ambos sistemas. Si se continúa la sustitución hacia atrás hasta la segunda ecuación en cada caso, la diferencia entre los dos sistemas se hace aparente ya que en el primer sistema

$$-x_3 + x_4 = 1 \quad \text{implica que} \quad 1 = 1\,,$$

mientras que en el segundo sistema

$$-x_3 + x_4 = -2 \quad \text{implica que} \quad 1 = -2\,.$$

El primer sistema lineal tiene un número infinito de soluciones $x_4 = 3$, $x_3 = 2$, x_2 arbitraria y $x_1 = 2 - x_2$, mientras que el segundo nos lleva a una contadicción y no existe solución. En ambos casos, sin embargo, no hay una solución única como concluímos a partir del algoritmo de eliminación Gaussiana con sustitución hacia atrás.

Cuando se comparan las técnicas para resolver sistemas lineales, se necesita considerar otros conceptos además de la cantidad de lugar requerido para almacenamiento. Uno de éstos conceptos es el efecto del error de redondeo y otro es la cantidad de tiempo requerido para completar los cálculos. Ambos dependen del número de operaciones aritméticas que se necesitan efectuar para resolver un problema. En general, el tiempo requerido para realizar una multiplicación o división es considerablemente mayor que el requerido para realizar una suma o una resta. Para mostrar el procedimiento que se emplea para contar las operaciones en un método dado, contaremos las operaciones necesarias para resolver un sistema lineal típico de n ecuaciones con n incógnitas usando el algoritmo de la eliminación Gaussiana con sustitución hacia atrás.

Hasta los pasos 5 y 6 del algoritmo no se efectúan operaciones aritméticas. El paso 5 requiere que se realicen $(n-i)$ divisiones. El reemplazar la ecuación E_j por $(E_j - m_{ji}E_i)$ en el paso 6 requiere que m_{ji} se multiplique por cada término en E_i resultando un total de $(n-i)(n-i+2)$ multiplicaciones. Después de completar esto, cada término de la ecuación resultante se resta del término correspondiente en E_j. Esto requiere $(n-i)(n-i+2)$ restas. Para cada $i = 1, 2, \ldots, n-1$, las operaciones requeridas en los pasos 5 y 6 son

Multiplicaciones/Divisiones

$$(n-i) + (n-i)\,(n-i+2) = (n-i)\,(n-i+3)\,,$$

Sumas/Restas

$$(n-i)\,(n-i+2)\,.$$

El número total de operaciones requeridas en estos pasos se obtiene sumando las cuentas de las operaciones para cada i. Recordando que

$$\sum_{j=1}^{m} 1 = m \ , \quad \sum_{j=1}^{m} j = \frac{m(m+1)}{2} \ , \quad \sum_{j=1}^{m} j^2 = \frac{m(m+1)(2m+1)}{6} \ ,$$

obtenemos

Multiplicaciones/Divisiones

$$\sum_{i=1}^{n-1}(n-i)(n-i+3) = (n^2+3n)\sum_{i=1}^{n-1} 1 - (2n+3)\sum_{i=1}^{n-1} i + \sum_{i=1}^{n-1} i^2 =$$

$$= (n^2+3n)(n-1) - (2n+3)\frac{(n-1)n}{2} + \frac{(n-1)n(2n-1)}{6} = \frac{n^3+3n^2-4n}{3} \ ,$$

Sumas/Restas

$$\sum_{i=1}^{n-1}(n-i)(n-i+2) = (n^2+2n)\sum_{i=1}^{n-1} 1 - 2(n+1)\sum_{i=1}^{n-1} i + \sum_{i=1}^{n-1} i^2 =$$

$$= (n^2+2n)(n-1) - 2(n+1)\frac{(n-1)n}{2} + \frac{(n-1)n(2n-1)}{6} = \frac{2n^3+3n^2-5n}{6} \ .$$

Los otros pasos del algoritmo de la eliminación Gaussiana con sustitución hacia atrás que requieren de operaciones aritméticas son los pasos 8 y 9. El n^0 8 requiere de una división. El n^0 9 requiere de $(n-i)$ multiplicaciones y $(n-i-1)$ sumas para cada término con sumatorio y luego una resta y una división. El número total de operaciones en los pasos 8 y 9 es

Multiplicaciones/Divisiones

$$1 + \sum_{i=1}^{n-1}[(n-i)+1] = \frac{n^2+n}{2} \ ,$$

Sumas/Restas

$$\sum_{i=1}^{n-1}[(n-i-1)+1] = \frac{n^2-n}{2} \ .$$

El total de operaciones aritméticas en el algoritmo de la eliminación Gaussiana con sustitución hacia atrás es por lo tanto

Multiplicaciones/Divisiones

$$\frac{n^3+3n^2-4n}{3} + \frac{n^2+n}{2} = \frac{2n^3+9n^2-5n}{6} \ ,$$

Sumas/Restas

$$\frac{2n^3+3n^2-5n}{6} + \frac{n^2-n}{2} = \frac{n^3+3n^2-4n}{3} \ .$$

Como el número total de multiplicaciones y de divisiones es aproxidamente $n^3/3$, y similar para sumas y restas, la cantidad de cómputo y el tiempo requerido se incrementarán con n proporcionalmente a n^3.

CAPITULO XV. ESTRATEGIAS DE PIVOTEO

1. INTRODUCCION Y METODO

Durante la derivación del algoritmo de la eliminación Gaussiana con sustitución hacia atrás, se encontró que para obtener un cero para el elemento pivote $a_{kk}^{(k)}$ era necesario un intercambio de filas de la forma $(E_k) \leftrightarrow (E_p)$ donde $k+1 \leq p \leq n$ era el entero más pequeño con $a_{pk}^{(k)} \neq 0$. En la práctica frecuentemente es deseable realizar intercambios de las filas que contienen a los elementos pivote, aun cuando éstos no sean cero. Cuando los cálculos se realizan usando aritmética de dígitos finitos, como sería el caso de las soluciones generadas con calculadora u ordenador, un elemento pivote que sea pequeño comparado con los elementos de debajo de él en la misma columna puede llevar a un error de redondeo sustancial. En el ejemplo siguiente se da una ilustración de esta dificultad.

Ejemplo. El sistema lineal

$$\begin{array}{rrcrcl} E_1: & 0.003\,x_1 & + & 59.14\,x_2 & = & 59.17\,, \\ E_2: & 5.291\,x_1 & - & 6.130\,x_2 & = & 46.78\,, \end{array}$$

tiene la solución exacta $x_1 = 10.00$ y $x_2 = 1.000$.

Para ilustrar las dificultades del error de redondeo, se aplicará eliminación Gaussiana a este sistema usando aritmética de cuatro dígitos con redondeo.

El primer elemento pivote es $a_{11}^{(1)} = 0.003$ y su multiplicador asociado es

$$m_{21} = \frac{5.291}{0.003} = 1763.\overline{6}\,,$$

el cual se redondea a 1764. Efectuando la operación $(E_2 - m_{21}E_1) \rightarrow (E_2)$ y el redondeo apropiado ($1764 \cdot 59.14 = 104322 = 104300$ y $1764 \cdot 59.17 = 104375 = 104400$),

$$\begin{array}{rcrcr} 0.003\,x_1 & - & 59.14\,x_2 & = & 59.17\,, \\ & - & 104300\,x_2 & = & -104400\,. \end{array}$$

La sustitución hacia atrás implica que

$$x_2 = 1.001\,, \qquad x_1 = \frac{59.17 - 59.14 \cdot 1.001}{0.003} = \frac{59.17 - 59.20}{0.003} = -\frac{0.030}{0.003} = -10.00\,.$$

El error absoluto tan grande en la solución numérica de x_1 resulta del error pequeño de 0.001 al resolver para x_2. Este error absoluto fue amplificado por un factor de 20000 en la solución de x_1 debido al orden en el que fueron realizados los cálculos.

El ejemplo anterior ilustra las dificultades que pueden surgir en algunos casos cuando el elemento pivote $a_{kk}^{(k)}$ es pequeño en relación a los elementos $a_{ij}^{(k)}$ para $k \leq i \leq n$ y $k \leq j \leq n$. Las estrategias de pivoteo se llevan a cabo en general seleccionando un nuevo elemento como pivote $a_{pq}^{(k)}$ intercambiando las filas k y p, e intercambiando las columnas k y q, si es necesario. La estrategia más simple consiste en seleccionar el elemento en la misma columna que está debajo de la diagonal y que tiene el mayor valor absoluto; es decir, se determina p tal que

$$|a_{pk}^{(k)}| = \max_{k \leq i \leq n} |a_{ik}^{(k)}|\,,$$

y se efectúa $(E_k) \leftrightarrow (E_p)$. En este caso no se considera un intercambio de columnas.

Ejemplo. Reconsideremos el sistema lineal del ejemplo anterior:

$$\begin{array}{lrcrcl} E_1: & 0.003\ x_1 & + & 59.14\ x_2 & = & 59.17\ , \\ E_2: & 5.291\ x_1 & - & 6.130\ x_2 & = & 46.78\ . \end{array}$$

Usando el procedimiento de pivoteo descrito arriba resulta que primero se encuentra

$$\max\{|a_{11}^{(1)}|, |a_{21}^{(1)}|\} = \max\{|0.003|, |5.291|\} = |5.291| = |a_{21}^{(1)}|\ .$$

Así, se realiza la operación $(E_2) \leftrightarrow (E_1)$ la cual da el sistema

$$\begin{array}{lrcrcl} E_1: & 5.291\ x_1 & - & 6.130\ x_2 & = & 46.78\ , \\ E_2: & 0.003\ x_1 & + & 59.14\ x_2 & = & 59.17\ . \end{array}$$

El multiplicador para este sistema es

$$m_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{0.003}{5.291} = 0.000567\ ,$$

y la operación $(E_2 - m_{21}E_1) \to (E_2)$ con el redondeo apropiado ($0.000567 \cdot 6.13 = 0.003476$ y $0.000567 \cdot 46.78 = 0.02652$) reduce el sistema a

$$\begin{array}{rcrcl} 5.291\ x_1 & - & 6.130\ x_2 & = & 46.78\ , \\ & & 59.14\ x_2 & = & 59.14\ . \end{array}$$

Las respuestas con cuatro dígitos que resultan de la sustitución hacia atrás son los valores correctos $x_1 = 10.00$ y $x_2 = 1.000$.

Esta técnica se conoce como **pivoteo máximo de columna** o **pivoteo parcial**.

2. ALGORITMOS DE ELIMINACION GAUSSIANA CON PIVOTEO

A continuación se presenta el algoritmo de eliminación Gaussiana con pivoteo parcial (pivoteo máximo de columna). Los procedimientos detallados en este algoritmo son suficientes para garantizar que cada multiplicador m_{ij} tiene una magnitud que no excede a uno.

Algoritmo de eliminación Gaussiana con pivoteo máximo de columna.

==========

Para resolver el sistema lineal de $n \times n$:

$$\begin{array}{ll} E_1: & a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n = a_{1,n+1} \\ E_2: & a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n = a_{2,n+1} \\ & \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: & a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n = a_{n,n+1} \end{array}$$

Entrada: número de incógnitas y de ecuaciones n; matriz ampliada $A_a = (a_{ij}) = (a(i,j))$ donde $1 \le i \le n$ y $1 \le j \le n+1$.

Salida: solución $x_1, x_2, \ldots, x_n$ ó mensaje de que el sistema lineal no tiene solución única.

Paso 1: Para $i = 1, 2, \ldots, n$ tomar $F(i) = i$;
(inicializar el indicador de la fila).

Paso 2: Para $i = 1, 2, \ldots, n-1$ seguir los pasos 3–6 *(proceso de eliminación).*

Paso 3: Sea p el menor entero con $i \leq p \leq n$ y

$$|a(F(p), i)| = \max_{i \leq j \leq n} |a(F(j), i)| \ .$$

Paso 4: Si $a(F(p), i) = 0$ entonces SALIDA;
(no existe solución única) PARAR.

Paso 5: Si $F(i) \neq F(p)$ entonces tomar $AUX = F(i)$, $F(i) = F(p)$, $F(p) = AUX$; *(intercambio de filas simulado).*

Paso 6: Para $j = i+1, i+2, \ldots, n$ seguir los pasos 7 y 8.

Paso 7: Tomar $m(F(j), i) = \frac{a(F(j),i)}{a(F(i),i)}$.

Paso 8: Efectuar $(E_{F(j)} - m(F(j), i)\ E_{F(i)}) \rightarrow (E_{F(j)})$.

Paso 9: Si $a(F(n), n) = 0$ entonces SALIDA; *(no existe solución única)* PARAR.

Paso 10: *(Empieza la sustitución hacia atrás)*; tomar

$$x_n = \frac{a(F(n), n+1)}{a(F(n), n)} \ .$$

Paso 11: Para $i = n-1, n-2, \ldots, 1$ tomar

$$x_i = \frac{a(F(i), n+1) - \sum_{j=i+1}^{n} a(F(i), j)\ x_j}{a(F(i), i)} \ .$$

Paso 12: SALIDA $(x_1, x_2, \ldots, x_n)$;
(procedimiento completado satisfactoriamente) PARAR.

==========

Aún cuando la estrategia del pivoteo máximo de columna es suficiente para la mayoría de los sistemas lineales, se presentan a veces situaciones en las que esta estrategia resulta inadecuada.

Ejemplo. El sistema lineal:

$$\begin{array}{rrcrcr} E_1: & 30.00\ x_1 & + & 591400\ x_2 & = & 591700\ , \\ E_2: & 5.291\ x_1 & - & 6.130\ x_2 & = & 46.78\ , \end{array}$$

es el mismo sistema que el presentado en los ejemplos previos excepto que todos los coeficientes en la primera ecuación están multiplicados por 10^4. El procedimiento descrito en el algoritmo de eliminación Gaussiana con pivoteo máximo de columna con aritmética de 4 dígitos lleva a los mismos resultados que se obtuvieron en el primer ejemplo.

El máximo valor en la primera columna es 30.00 y el multiplicador

$$m_{21} = \frac{5.291}{30.00} = 0.1764$$

y la operación $(E_2 - m_{21}E_1) \to (E_2)$ con el redondeo apropiado ($0.1764 \cdot 591400 = 104322 = 104300$ y $0.1764 \cdot 591700 = 104375 = 104400$) transformaría el sistema en

$$\begin{array}{rcrcr} 30.00\, x_1 & + & 591400\, x_2 & = & 591700\ , \\ & - & 104300\, x_2 & = & -104400\ , \end{array}$$

el cual tiene soluciones $x_2 = 1.001$ y $x_1 = -10.00$.

Para el sistema del último ejemplo es apropiada una técnica conocida como **pivoteo escalado de columna**. El primer paso en este procedimiento consiste en definir un factor de escala s_l para cada fila $l = 1, \ldots, n$

$$s_l = \max_{1 \le j \le n} |a_{lj}|\ .$$

Si $s_l = 0$ para algún l, los Teoremas XIII.3 y XIII.4 implican que no existe solución única y el procedimiento se detiene. El intercambio apropiado de filas para luego obtener ceros en la primera columna queda determinado escogiendo el primer entero $1 \le k \le n$ con

$$\frac{|a_{k1}|}{s_k} = \max_{1 \le j \le n} \frac{|a_{j1}|}{s_j}\ ,$$

y realizando $(E_1) \leftrightarrow (E_k)$. Igualmente, al paso generico i, el intercambio apropiado para llevar el elemento pivote a_{ii} en su posición, queda determinado escogiendo el menor entero k, $i \le k \le n$, con

$$\frac{|a_{ki}|}{s_k} = \max_{i \le j \le n} \frac{|a_{ji}|}{s_j}\ ,$$

y realizando $(E_i) \leftrightarrow (E_k)$. Si al efectuar este intercambio no se variían los factores de escala, diremos que estamos aplicando una estrategia de **pivoteo escalado de columna con factores de escalas fijos**. Por otra parte, otra estategia es efectuar también el intercambio $(s_i) \leftrightarrow (s_k)$ si se está haciendo el intercambio de filas $(E_i) \leftrightarrow (E_k)$ $(1 \le i \le n,$ $i \le k \le n)$. En este caso diremos que se aplica la estrategia de pivoteo escalado de columna con intercambio completo o simplemente **pivoteo escalado de columna**.

Una modificación de esta técnica de pivoteo escalado de columna, que llameremos **pivoteo escalado de columna modificado**, consiste en redefinir los factores de escala a cada paso, es decir, al paso i-ésimo de nuestro algoritmo $(1 \le i \le n)$ se definen los factores de escala s_l para cada fila $l = i, \ldots, n$

$$s_l = \max_{i \le j \le n} |a_{lj}|\ .$$

Entonces, el intercambio apropiado de filas para llevar el elemento pivote a_{ii} en su posición queda determinado escogiendo el primer entero k, $i \le k \le n$, con

$$\frac{|a_{ki}|}{s_k} = \max_{i \le j \le n} \frac{|a_{j1}|}{s_j}\ ,$$

y realizando luego $(E_i) \leftrightarrow (E_k)$.

El efecto de escalar consiste en asegurar que el elemento mayor de cada fila tenga una magnitud relativa de uno antes de que se empiece la comparación para el intercambio de filas. El escalamiento se hace solamente con propósitos de comparación, así que la división entre los factores de escala no produce un error de redondeo en el sistema.

Aplicando la técnica de pivoteo escalado de columna al último ejemplo se obtiene

$$s_1 = \max\{|30.00|, |591400|\} = 591400 \ ,$$

$$s_2 = \max\{|5.291|, |-6.130|\} = 6.130 \ .$$

Consecuentemente,

$$\frac{|a_{11}|}{s_1} = \frac{30.00}{591400} = 0.5073 \times 10^{-4} \quad \text{y} \quad \frac{|a_{21}|}{s_2} = \frac{5.291}{6.130} = 0.8631 \ ,$$

y por lo cual se hace el intercambio $(E_1) \leftrightarrow (E_2)$.

Aplicando eliminación Gaussiana, el nuevo sistema

$$\begin{array}{rcrcr} 5.291\ x_1 & - & 6.130\ x_2 & = & 46.78 \ , \\ 30.00\ x_1 & + & 591400\ x_2 & = & 591700 \ , \end{array}$$

producirá los resultados correctos $x_1 = 10.00$ y $x_2 = 1.000$. De hecho, el multiplicador es

$$m_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{30.00}{5.291} = 5.67 \ ,$$

y la operación $(E_2 - m_{21}E_1) \rightarrow (E_2)$ (con $5.67 \cdot 6.13 = 34.76$ y $5.67 \cdot 46.78 = 256.2$) reduce el sistema a

$$\begin{array}{rcrcr} 5.291\ x_1 & - & 6.130\ x_2 & = & 46.78 \ , \\ & & 591400\ x_2 & = & 591400 \ . \end{array}$$

Las respuestas con cuatro dígitos que resultan de la sustitución hacia atrás son los valores correctos $x_1 = 10.00$ y $x_2 = 1.000$.

Algoritmo de eliminación Gaussiana con pivoteo escalado de columna.

==========

Para resolver el sistema lineal de $n \times n$:

$$\begin{array}{rl} E_1: & a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n = a_{1,n+1} \\ E_2: & a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n = a_{2,n+1} \\ & \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: & a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n = a_{n,n+1} \end{array}$$

Entrada: número de incógnitas y de ecuaciones n; matriz ampliada $A_a = (a_{ij}) = (a(i,j))$ donde $1 \le i \le n$ y $1 \le j \le n+1$.

Salida: solución $x_1, x_2, \ldots, x_n$ ó mensaje de que el sistema lineal no tiene solución única.

Paso 1: Para $i = 1, 2, \ldots, n$ tomar

$$s_i = s(i) = \max_{1 \le j \le n} |a(i,j)| \ ;$$

si $s_i = 0$ entonces SALIDA; *(no existe solución única)* PARAR. Tomar $F(i) = i$; *(inicializar el indicador de la fila).*

Paso 2: Para $i = 1, 2, \ldots, n-1$ seguir los pasos 3–6 *(proceso de eliminación).*

Paso 3: Sea p el menor entero con $i \leq p \leq n$ y

$$\frac{|a(F(p), i)|}{s(F(p))} = \max_{i \leq j \leq n} \frac{|a(F(j), i)|}{s(F(j))} \ .$$

Paso 4: Si $a(F(p), i) = 0$ entonces SALIDA;
(no existe solución única) PARAR.

Paso 5: Si $F(i) \neq F(p)$ entonces tomar $AUX = F(i)$, $F(i) = F(p)$, $F(p) = AUX$; *(intercambio de filas simulado).*

Paso 6: Para $j = i+1, i+2, \ldots, n$ seguir los pasos 7 y 8.

Paso 7: Tomar $m(F(j), i) = \frac{a(F(j),i)}{a(F(i),i)}$.

Paso 8: Efectuar $(E_{F(j)} - m(F(j), i)\ E_{F(i)}) \rightarrow (E_{F(j)})$.

Paso 9: Si $a(F(n), n) = 0$ entonces SALIDA; *(no existe solución única)* PARAR.

Paso 10: *(Empieza la sustitución hacia atrás)*; tomar

$$x_n = \frac{a(F(n), n+1)}{a(F(n), n)} \ .$$

Paso 11: Para $i = n-1, n-2, \ldots, 1$ tomar

$$x_i = \frac{a(F(i), n+1) - \sum_{j=i+1}^{n} a(F(i), j)\ x_j}{a(F(i), i)} \ .$$

Paso 12: SALIDA $(x_1, x_2, \ldots, x_n)$;
(procedimiento completado satisfactoriamente) PARAR.

==========

Los cálculos adicionales requeridos para el pivoteo escalado de columna resultan primero de la determinación de los factores de escala, es decir $(n-1)$ comparaciones para cada uno de las n filas, que da un total de

$$n(n-1) \quad \text{comparaciones} \ .$$

Para determinar el primer intercambio correcto, se realizan n divisiones y se hacen $(n-1)$ comparaciones. La determinación del primer intercambio entonces, añade un total de

$$\textit{comparaciones} \quad n(n-1) + (n-1) \quad \text{y} \quad \textit{divisiones} \quad n \ .$$

Como los factores de escala se calculan sólo una vez, el segundo paso requiere solamente

$$\textit{comparaciones} \quad (n-2) \quad \text{y} \quad \textit{divisiones} \quad (n-1) \ .$$

Procediendo de manera similar, el procedimiento de pivoteo escalado de columna agrega un total de

$$\textit{comparaciones} \quad (n-1) + \sum_{k=2}^{n}(k-1) = \frac{3}{2}n(n-1)$$

y

$$\textit{divisiones} \quad \sum_{k=2}^{n} k = \frac{n(n+1)}{2} - 1 \ ,$$

al procedimiento de eliminación Gaussiana. El tiempo requerido para realizar una comparación es comparable, aunque un poco mayor, al de suma/resta. Entonces la técnica de escalamiento no incrementa significativamente el tiempo de cómputo requerido para resolver un sistema para valores grandes de n.

Si un sistema garantiza el tipo de pivoteo que da un pivoteo escalado de columna modificado, entonces se debe usar **pivoteo máximo** o **total**. Es decir, este pivoteo máximo en el $k-$ésimo paso busca todos los elementos

$$a_{ij} \quad \text{para} \quad i = k, k+1, \ldots, n, \quad \text{y} \quad j = k, k+1, \ldots, n,$$

para encontrar el elemento que tiene la magnitud más grande. Se realizan intercambios de filas y de columnas para traer este elemento a la posición pivote.

El primer paso de pivoteo total requiere que se realicen $(n^2 - 1)$ comparaciones, el segundo paso requiere $[(n-1)^2 - 1]$ comparaciones, y así sucesivamente. El tiempo total adicional requerido para incorporar el pivoteo total en la eliminación Gaussiana es consecuentemente

$$\textit{comparaciones} \quad \sum_{k=2}^{n}(k^2 - 1) = \frac{n(n-1)(2n+5)}{6} \ .$$

Este número es comparable con el número requerido por una técnica de pivoteo de columna modificada, pero no es necesaria ninguna división. El pivoteo total es consecuentemente la estrategia recomendada para la mayoría de los sistemas complicados para los cuales se puede justificar la cantidad de tiempo de ejecución tan intensa.

3. EJEMPLO DE ALGORITMO FORTRAN

En esta sección vamos a presentar una versión FORTRAN muy sencilla del algoritmo de eliminación Gaussiana con pivoteo máximo de columna. En el esquema de la programación estructurada FORTRAN, el problema de la búsqueda de solución de un sistema de ecuaciones lineales será desarrollado dividiéndolo en un programa principal y en varios subprogramas, donde cada uno de ellos resuelve una tarea particular. En nuestro caso, el problema será resuelto usando un programa principal que llama a la subrutina MATRIZA, para la lectura de los elemento de la matriz ampliada A_a, correspondiente al sistema dado $A\ \mathbf{x} = \mathbf{b}$ y a las subrutinas GAUSELI, GAUSMAX o GAUSESC, dependiendo de qué método se quiere usar, para el desarrollo del algoritmo de eliminación Gaussiana sin pivoteo, la primera, con pivoteo máximo de columna, la segunda, y con pivoteo escalado

de columna, la tercera. Aquí se dará solamente la versión FORTRAN de la subrutina GAUSMAX (las otras se pueden obtener de ésta con sencillas modificaciones).

```
C       PROGRAMA PRINCIPAL
        PROGRAM SISLIN
        PARAMETER (M = 20, MM = 21)
        REAL XX(M), AA(M, MM)
        INTEGER N, I, J, INDEX
        EXTERNAL MATRIZA, GAUSELI, GAUSMAX, GAUSESC
C
        PRINT*, 'NUMERO DE DIMENSION MAXIMA', M
        PRINT*, 'DAR LA DIMENSION DEL PROBLEMA'
        READ*, N
        PRINT*, 'ESCOGER EL METODO A USAR'
        PRINT*, 'INDEX = 0, ELIMINACION GAUSSIANA CON'
        PRINT*, '  SUSTITUCION HACIA ATRAS SIN PIVOTEO'
        PRINT*, 'INDEX = 1, ELIMINACION GAUSSIANA CON'
        PRINT*, '  PIVOTEO MAXIMO DE COLUMNA'
        PRINT*, 'INDEX = 2, ELIMINACION GAUSSIANA CON'
        PRINT*, '  PIVOTEO ESCALADO DE COLUMNA'
        READ*, INDEX
        IF(INDEX.EQ.0) PRINT*, ' NO USARE PIVOTEO'
        IF(INDEX.EQ.1) PRINT*, ' USARE PIVOTEO MAXIMO'
        IF(INDEX.EQ.2) PRINT*, ' USARE PIVOTEO ESCALADO'
C
        CALL MATRIZA(N, AA, M)
        IF(INDEX.EQ.0) CALL GAUSELI (N, AA, M, XX)
        IF(INDEX.EQ.1) CALL GAUSMAX (N, AA, M, XX)
        IF(INDEX.EQ.2) CALL GAUSESC (N, AA, M, XX)
C
        PRINT*, ' LA APROXIMACION A LA SOLUCION ES'
        DO 10 I = 1, N
10          PRINT*, XX(I)
        STOP
        END
*******************************************************
        SUBROUTINE GAUSMAX (N, A, M, XX)
        PARAMETER (MM = 20)
        INTEGER I, J, K, N, IFIL(MM)
        REAL AA(M,*), XX(M), CHECK, CHECK1, MUL(MM,MM)
C
        DO 10 K = 1, N
10          IFIL(K) = K
C
        DO 99 I = 1, N - 1
          PRINT*, '    *** PASO NUMERO ***  I : ', I
          CHECK = ABS(A(IFIL(I), I))
          IP = I
         DO 20 J = I + 1, N
           CHECK1 = ABS(A(IFIL(J), I))
           IF(CHECK1.GT.CHECK) THEN
            CHECK = CHECK1
            IP = J
```

```
            PRINT*, ' HAY  INTERCAMBIO  DE  I : ', I
            PRINT*, '            CON  IP : ', IP
          ENDIF
20       CONTINUE
          IF(A(IFIL(IP),I).EQ.0.0)  THEN
           PRINT*, ' NO  EXISTE  SOLUCION  UNICA '
           GOTO  999
          ENDIF
          IF(IFIL(I).NE.IFIL(IP))  THEN
           AUX = IFIL(I)
           IFIL(I) = IFIL(IP)
           IFIL(IP) = AUX
          ENDIF
         DO  77  J = I + 1,  N
           MUL(IFIL(J),I) = A(IFIL(J),I)/A(IFIL(I),I)
           PRINT*, ' MULTIPLICADOR '
           PRINT*,  I,  J,  MUL(IFIL(J),I)
          DO  88  K = 1,  N + 1
88          A(IFIL(J),K) = A(IFIL(J),K) - MUL(IFIL(J),I) * A(IFIL(I),K)
77       CONTINUE
          PRINT*,  ((A(K,J) ,   J = 1,  N + 1) ,   K = 1,  N)
99      CONTINUE
C
        IF(A(IFIL(N),N).EQ.0.0)  THEN
          PRINT*, ' NO  EXISTE  SOLUCION  UNICA '
          GOTO  999
        ENDIF
C
           XX(N) = A(IFIL(N),N + 1)/A(IFIL(N),N)
        DO  55  I = N - 1,  1,  -1
           SUMA = 0.0
          DO  44  J = I + 1,  N
44          SUMA = SUMA + A(IFIL(I),J) * XX(J)
           XX(I) = (A(IFIL(I),N + 1) - SUMA)/A(IFIL(I),I)
55      CONTINUE
        PRINT*, ' EL  PROCEDIMIENTO  HA  SIDO '
        PRINT*, ' COMPLETADO  SATISFACTORIAMENTE '
999     CONTINUE
        RETURN
        END
    ************************************************************
        SUBROUTINE  MATRIZA (N,  AA,  M)
        INTEGER  N,  I,  J,  M
        REAL  AA(M,*)
        OPEN (UNIT = 13,FILE =' IN.DAT')
C
        DO  10  I = 1,  N
          DO  10  J = 1,  N + 1
10          READ(13,*) AA(I,  J)
        CLOSE(13)
        RETURN
        END
```

4. EL ALGORITMO DE GAUSS-JORDAN

Como hemos visto en el método conocido como regla de Cramer (ver Capítulo XXII), para resolver el sistema lineal $A\ \mathbf{x} = \mathbf{b}$ se puede necesitar la matriz inversa A^{-1} para obtener $\mathbf{x} = A^{-1}\ \mathbf{b}$ como única solución del sistema. Sin embargo la inversa A^{-1} de una matriz $n \times n$ no singular A no se necesita a menudo, dado que existen otros métodos para resolver los sistemas lineales. De cualquier manera, el **algoritmo de Gauss-Jordan** nos da un método para invertir la aplicación $\mathbf{x} \longrightarrow A\ \mathbf{x} = \mathbf{y}$, $\mathbf{x} \in \mathcal{R}^n$, $\mathbf{y} \in \mathcal{R}^n$, de una manera sistemática.

Consideremos el sistema $A\ \mathbf{x} = \mathbf{y}$:

$$\begin{aligned} E_1: &\quad a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n = y_1\ , \\ E_2: &\quad a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n = y_2\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n = y_n\ . \end{aligned} \tag{XV.1}$$

En el primer paso del método de Gauss-Jordan, la variable x_1 se cambia por una de las variables y_r. Para hacer esto, se busca un coeficiente $a_{r1} \neq 0$, por ejemplo con el pivoteo máximo de columna:

$$|a_{r1}| = \max_{1 \leq i \leq n} |a_{i1}|$$

y las ecuaciones E_1 y E_r vienen intercambiadas, es decir, se hace un intercambio de filas $(E_1) \leftrightarrow (E_r)$. De esta manera se obtiene un sistema:

$$\begin{aligned} E_1: &\quad \overline{a}_{11}\ x_1 + \overline{a}_{12}\ x_2 + \ldots + \overline{a}_{1n}\ x_n = \overline{y}_1\ , \\ E_2: &\quad \overline{a}_{21}\ x_1 + \overline{a}_{22}\ x_2 + \ldots + \overline{a}_{2n}\ x_n = \overline{y}_2\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad \overline{a}_{n1}\ x_1 + \overline{a}_{n2}\ x_2 + \ldots + \overline{a}_{nn}\ x_n = \overline{y}_n\ , \end{aligned} \tag{XV.2}$$

en el cual las variables $\overline{y}_1$, ..., $\overline{y}_n$ son permutaciones de y_1, ..., y_n, y además $\overline{a}_{11} = a_{r1}$, $\overline{y}_1 = y_r$. Ahora $\overline{a}_{11} \neq 0$, porque si no fuese así,tendríamos $a_{i1} = 0$ para todo i, con lo que A sería singular. Resolvamos la primera ecuación de $(XV.2)$ para x_1, y sustituyamos el resultado en todas las demás ecuaciones del sistema. Entonces se obtiene el sistema:

$$\begin{aligned} E_1: &\quad a'_{11}\ \overline{y}_1 + a'_{12}\ x_2 + \ldots + a'_{1n}\ x_n = x_1\ , \\ E_2: &\quad a'_{21}\ \overline{y}_1 + a'_{22}\ x_2 + \ldots + a'_{2n}\ x_n = \overline{y}_2\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad a'_{n1}\ \overline{y}_1 + a'_{n2}\ x_2 + \ldots + a'_{nn}\ x_n = \overline{y}_n\ , \end{aligned} \tag{XV.3}$$

donde, para todo $i, k = 2, 3, \ldots, n$,

$$a'_{11} = \frac{1}{\overline{a}_{11}}, \qquad a'_{1k} = -\frac{\overline{a}_{1k}}{\overline{a}_{11}},$$

$$a'_{i1} = \frac{\overline{a}_{i1}}{\overline{a}_{11}}, \qquad a'_{ik} = \overline{a}_{ik} - \overline{a}_{i1}\ \frac{\overline{a}_{1k}}{\overline{a}_{11}}.$$

En el paso siguiente, la variable x_2 se cambia con una de las variables $\overline{y}_2$, ..., $\overline{y}_n$; es decir, se busca $a_{r2} \neq 0$, tal que $|a_{r2}| = \max\limits_{2\leq i\leq n} |a_{i2}|$ y se hace un intercambio de filas $(E_2) \leftrightarrow (E_r)$; luego se resuelve la segunda ecuación para x_2, y se sustituye en todas las demás ecuaciones del sistema. Después se repite para las variables x_3 y para todas las demás. Si representamos los sistemas con sus matrices, partiendo de $A = A^{(0)}$, se obtiene una sucesión $A^{(0)} \to A^{(1)} \to \ldots \to A^{(n)}$. La matriz genérica $A^{(j)} = a_{ik}^{(j)}$ representa el *sistema mixto de ecuaciones* de la forma

$$\begin{aligned}
E_1: &\quad a_{11}^{(j)}\ \tilde{y}_1 + \ldots + a_{1j}^{(j)}\ \tilde{y}_j + a_{1,j+1}^{(j)}\ x_{j+1} + \ldots + a_{1n}^{(j)}\ x_n = x_1\ ,\\
&\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\\
E_j: &\quad a_{21}^{(j)}\ \tilde{y}_1 + \ldots + a_{jj}^{(j)}\ \tilde{y}_j + a_{j,j+1}^{(j)}\ x_{j+1} + \ldots + a_{jn}^{(j)}\ x_n = x_j\ ,\\
E_{j+1}: &\quad a_{21}^{(j)}\ \tilde{y}_1 + \ldots + a_{j+1,j}^{(j)}\ \tilde{y}_j + a_{j+1,j+1}^{(j)}\ x_{j+1} + \ldots + a_{2n}^{(j)}\ x_n = \tilde{y}_{j+1}\ ,\\
&\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\\
E_n: &\quad a_{n1}^{(j)}\ \tilde{y}_1 + \ldots + a_{nj}^{(j)}\ \tilde{y}_j + a_{n,j+1}^{(j)}\ x_{j+1} + \ldots + a_{nn}^{(j)}\ x_n = \tilde{y}_n\ .
\end{aligned} \qquad (XV.4)$$

En este sistema ($\tilde{y}_1$, ..., $\tilde{y}_n$) indica una permutación de las variables originarias (y_1, ..., y_n). En el paso $A^{(j-1)} \to A^{(j)}$ la variable x_j se intercambia por $\tilde{y}_j$. Entonces, se obtiene $A^{(j)}$ de $A^{(j-1)}$ según las reglas dadas abajo. Por simplicidad, los elementos de $A^{(j-1)}$ se indican con a_{ik}, y los elementos de $A^{(j)}$ con a'_{ik}.

Reglas para el algoritmo de Gauss-Jordan con pivoteo máximo de columna.

a) Determinar r como el menor entero $j \leq r \leq n$ tal que

$$|a_{rj}| = \max_{j\leq i\leq n} |a_{ij}|\ .$$

Si $a_{rj} = 0$, la matriz es singular y no hay solución.

b) Intercambiar las filas r y j de la matriz $A^{(j-1)}$ y llamar al resultado $\overline{A} = \overline{a}_{ik}$.

c) Calcular $A^{(j)} = a'_{ik}$, para $i, k \neq j$, según las fórmulas

$$\begin{aligned}
a'_{jj} &= \frac{1}{\overline{a}_{jj}}, & a'_{jk} &= -\frac{\overline{a}_{jk}}{\overline{a}_{jj}},\\
a'_{ij} &= \frac{\overline{a}_{ij}}{\overline{a}_{jj}}, & a'_{ik} &= \overline{a}_{ik} - \overline{a}_{ij}\ \frac{\overline{a}_{jk}}{\overline{a}_{jj}}.
\end{aligned}$$

El sistema ($XV.4$) implica que

$$A^{(n)}\ \hat{\mathbf{y}} = \mathbf{x}\ , \qquad \hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^t$$

donde $(\hat{y}_1, \ldots, \hat{y}_n)$ es una permutación de las variables originales $(y_1, \ldots, y_n)$; es decir, $\hat{\mathbf{y}} = P\ \mathbf{y}$ que corresponde a los intercambios de filas hechos en el paso b) del algoritmo de Gauss-Jordan, y puede ser fácilmente determinado. Entonces, $A^{(n)}\ \hat{\mathbf{y}} = A^{(n)}\ P\ \mathbf{y} = \mathbf{x}$ además de $A\ \mathbf{x} = \mathbf{y}$, lo que implica

$$A^{-1} = A^{(n)}\ P\ .$$

En la práctica, cuando se hacen los cálculos para resolver *a mano* un sistema de ecuaciones lineal, no se construyen las matrices $A^{(k)}$, si no que se trabaja directamente sobre el sistema. Mostraremos esta manera de proceder en el ejemplo siguiente.

Ejemplo. Resolvemos el sistema lineal

$$\begin{array}{lrcrcrcl} E_1: & x_1 & + & 2\,x_2 & - & x_3 & = & 2\,, \\ E_2: & 2\,x_1 & + & x_2 & & & = & 3\,, \\ E_3: & -\,x_1 & + & x_2 & + & 2\,x_3 & = & 4\,, \end{array}$$

con el método de Gauss-Jordan con pivoteo escalado de columna y aritmética de tres dígitos.

En el primer paso del método de Gauss-Jordan, la variable x_1 se cambia por una de las variables y_r. Para hacer esto, se busca un coeficiente $a_{r1} \neq 0$, con el pivoteo escalado de columna:

$$s_i = \max_{1\leq j\leq 3} |a_{ij}|\,, \qquad \frac{|a_{r1}|}{s_r} = \max_{1\leq i\leq 3} \frac{|a_{i1}|}{s_i}\,,$$

y las ecuaciones E_1 y E_r vienen intercambiadas, $(E_1) \leftrightarrow (E_r)$. En nuestro caso

$$s_1 = s_2 = s_3 = 2\,,$$

$$\frac{|a_{11}|}{s_1} = \frac{1.0}{2.0} = 0.5\,, \qquad \frac{|a_{21}|}{s_2} = \frac{2.0}{2.0} = 1.0\,, \qquad \frac{|a_{31}|}{s_3} = \frac{1.0}{2.0} = 0.5\,,$$

y así tenemos que intercambiar las primera y la segunda ecuación, y también tenemos que intercambiar los factores de escala, aunque en este caso quedan iguales $s_1 = s_2 = s_3 = 2$. De esta manera se obtiene el sistema:

$$\begin{array}{rcrcrcl} 2\,x_1 & + & x_2 & & & = & 3\,, \\ x_1 & + & 2\,x_2 & - & x_3 & = & 2\,, \\ -\,x_1 & + & x_2 & + & 2\,x_3 & = & 4\,. \end{array}$$

Ahora, resolvemos la primera ecuación por x_1, y sustituyemos el resultado en todas las demás ecuaciones:

$$\begin{array}{rcrcrcrcl} 1.5 & - & 0.5\,x_2 & & & & & = & x_1\,, \\ (1.5 & - & 0.5\,x_2) & + & 2\,x_2 & - & x_3 & = & 2\,, \\ (-1.5 & + & 0.5\,x_2) & + & x_2 & + & 2\,x_3 & = & 4\,. \end{array}$$

Entonces,

$$\begin{array}{rcrcrcl} 1.5 & - & 0.5\,x_2 & & & = & x_1\,, \\ 1.5 & + & 1.5\,x_2 & - & x_3 & = & 2\,, \\ -1.5 & + & 1.5\,x_2 & + & 2\,x_3 & = & 4\,. \end{array}$$

Ahora aplicamos otra vez el pivoteo escalado de columna:

$$\frac{|a_{22}|}{s_2} = \frac{1.5}{2.0} = 0.75\,, \qquad \frac{|a_{32}|}{s_3} = \frac{1.5}{2.0} = 0.75\,,$$

y así no hay que intercambiar ecuaciones. Entonces, podemos resolver la segunda ecuación por x_2, y sustituir el resultado en las demás:

$$\begin{array}{rcrcrcl} 1.5 & - & 0.5\,(0.333 + 0.667\,x_3) & & & = & x_1\ , \\ 0.333 & & & + & 0.667\,x_3 & = & x_2\ , \\ -1.5 & + & 1.5\,(0.333 + 0.667\,x_3) & + & 2\,x_3 & = & 4\ , \end{array}$$

que nos da

$$\begin{array}{rcrcl} 1.33 & - & 0.334\,x_3 & = & x_1\ , \\ 0.333 & + & 0.667\,x_3 & = & x_2\ , \\ -1.00 & + & 3\,x_3 & = & 4\ . \end{array}$$

Finalmente, resolvemos la tercera ecuación por la variable x_3, y sustituyamos el resultado en las demás,

$$\begin{array}{rcrcll} 3\,x_3 & = & 5\ , & \Rightarrow & x_3 = \frac{5}{3} = 1.67 \\ 1.33 & - & 0.334\,(1.67) & = & x_1\ , \\ 0.333 & + & 0.667\,(1.67) & = & x_2\ , \end{array}$$

para obtener la solución

$$x_1 = 0.772\ , \qquad x_2 = 1.44\ , \qquad x_3 = 1.67\ ,$$

que es una buena aproximación de la solución exacta

$$x_1 = \frac{7}{9}\ , \qquad x_2 = \frac{13}{9}\ , \qquad x_3 = \frac{15}{9}\ .$$

CAPITULO XVI. FACTORIZACION DIRECTA DE MATRICES

1. INTRODUCCION Y METODO

La discusión centrada alrededor del Teorema XIII.6 se refirió a la factorización de una matriz A en términos de una matriz triangular inferior L y de una matriz triangular superior U. Esta factorización existe cuando se puede resolver de manera única el sistema lineal $A\,\mathbf{x} = \mathbf{b}$ por eliminación Gaussiana sin intercambios de filas o columnas. El sistema $L\,U\,\mathbf{x} = A\,\mathbf{x} = \mathbf{b}$ puede transformarse entonces en el sistema $U\,\mathbf{x} = L^{-1}\,\mathbf{b}$ y como U es triangular superior, se puede aplicar una sustitución hacia atrás. Aún cuando las formas específicas de L y U se pueden obtener del proceso de eliminación Gaussiana, es deseable encontrar un método más directo para su determinación, para que, si fuera necesaria la solución de varios sistemas usando A, sólo se necesitaría realizar una sustitución hacia adelante y otra hacia atrás. Para ilustrar un procedimiento para calcular los elementos de estas matrices, consideremos un ejemplo.

Ejemplo. Considere la matriz estrictamente dominante diagonalmente de 4×4:

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

Los Teoremas XIII.6 y XIII.8 garantizan que A se puede factorizar en la forma $A = L\,U$, donde:

$$L = \begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix} .$$

Los 16 elementos conocidos de A se pueden usar para determinar parcialmente los diez elementos desconocidos de L y el mismo número de U. Sin embargo si el procedimiento nos debe llevar a una solución única, se necesitan cuatro condiciones adicionales para los elementos de L y de U. El método a usar en este ejemplo consiste en requerir arbitrariamente que $l_{11} = l_{22} = l_{33} = l_{44} = 1$, y se conoce como el **método de Doolittle**. Más adelante en este capítulo, se considerarán métodos que requieren que todos los elementos de la diagonal de U sean uno (**método de Crout**) y que $l_{ii} = u_{ii}$ para cada valor de i (**método de Choleski**).

La parte de la multiplicación de L con U,

$$L\,U = \begin{pmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix} =$$

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

que determina la primera fila de A, da lugar a las cuatro ecuaciones

$$u_{11} = 6 \ , \quad u_{12} = 2 \ , \quad u_{13} = 1 \ , \quad u_{14} = -1 \ .$$

La parte de la multiplicación de L con U que determina los elementos restantes de la primera columna de A da las ecuaciones

$$l_{21}\ u_{11} = 2 \ , \quad l_{31}\ u_{11} = 1 \ , \quad l_{41}\ u_{11} = -1 \ ,$$

y entonces

$$l_{21} = \frac{1}{3} \ , \quad l_{31} = \frac{1}{6} \ , \quad l_{41} = -\frac{1}{6} \ .$$

Hasta aquí las matrices L y U asumen la forma:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/3 & 1 & 0 & 0 \\ 1/6 & l_{32} & 1 & 0 \\ -1/6 & l_{42} & l_{43} & 1 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 6 & 2 & 1 & -1 \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix} \ .$$

La parte de la multiplicación que determina los elementos restantes en la segunda fila de A lleva a las ecuaciones

$$\begin{aligned} l_{21}\ u_{12} + u_{22} &= \frac{2}{3} + u_{22} = 4 \ , \\ l_{21}\ u_{13} + u_{23} &= \frac{1}{3} + u_{23} = 1 \ , \\ l_{21}\ u_{14} + u_{24} &= -\frac{1}{3} + u_{24} = 0 \ , \end{aligned}$$

así que

$$u_{22} = \frac{10}{3} \ , \quad u_{23} = \frac{2}{3} \ , \quad u_{24} = \frac{1}{3} \ ;$$

y la que determina los elementos restantes de la segunda columna de A da

$$\begin{aligned} l_{31}\ u_{12} + l_{32}\ u_{22} &= \frac{2}{6} + \frac{10}{3}\ l_{32} = 1 \ , \\ l_{41}\ u_{12} + l_{32}\ u_{22} &= -\frac{2}{6} + \frac{10}{3}\ l_{42} = 0 \ , \end{aligned}$$

así que

$$l_{32} = \frac{1}{5} \ , \quad l_{42} = \frac{1}{10} \ .$$

Ahora las matrices L y U tienen la forma:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/3 & 1 & 0 & 0 \\ 1/6 & 1/5 & 1 & 0 \\ -1/6 & 1/10 & l_{43} & 1 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 6 & 2 & 1 & -1 \\ 0 & 10/3 & 2/3 & 1/3 \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix} \ .$$

La parte de la multiplicación que determina los elementos restantes en la tercera fila de A lleva a las ecuaciones

$$l_{31}\ u_{13} + l_{32}\ u_{23} + u_{33} = \frac{1}{6} + \frac{1}{5}\cdot\frac{2}{3} + u_{33} = 4\ ,$$
$$l_{31}\ u_{14} + l_{32}\ u_{24} + u_{34} = -\frac{1}{6} + \frac{1}{5}\cdot\frac{1}{3} + u_{34} = -1\ ,$$

así que

$$u_{33} = \frac{37}{10} \quad \text{y} \quad u_{34} = -\frac{9}{10}\ ;$$

y la que determina los elementos restantes de la tercera columna de A da

$$l_{41}\ u_{13} + l_{42}\ u_{23} + l_{43}\ u_{33} = -\frac{1}{6} + \frac{1}{10}\ \frac{2}{3} + \frac{37}{10}\ l_{43} = -1\ ,$$

así que

$$l_{43} = -\frac{9}{37}\ .$$

Y finalmente, la última ecuación es:

$$l_{41}\ u_{14} + l_{42}\ u_{24} + l_{43}\ u_{34} + u_{44} = -\frac{1}{6}(-1) + \frac{1}{10}\cdot\frac{1}{3} - \frac{9}{37}\ (-\frac{9}{10}) + u_{44} = 3\ ,$$

así que

$$u_{4u} = \frac{191}{74}\ ;$$

para obtener finalmente:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{3} & 1 & 0 & 0 \\ \frac{1}{6} & \frac{1}{5} & 1 & 0 \\ -\frac{1}{6} & \frac{1}{10} & -\frac{9}{37} & 1 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 6 & 2 & 1 & -1 \\ 0 & \frac{10}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{37}{10} & -\frac{9}{10} \\ 0 & 0 & 0 & \frac{191}{74} \end{pmatrix}\ .$$

2. LOS ALGORITMOS DE DOOLITTLE Y DE CROUT

En el siguiente algoritmo de factorización directa está contenido un procedimiento general para **factorizar matrices** en un producto de **matrices triangulares**. Aunque se construyen nuevas matrices L y U, los valores generados pueden reemplazar a los elementos correspondientes de A que no son ya necesarios. Por lo tanto, la nueva matriz tiene elementos $a_{ij} = l_{ij}$ para cada $i = 2, 3, \ldots, n$ y $j = 1, 2, 3, \ldots, i-1$; y $a_{ij} = u_{ij}$ para cada $i = 1, 2, 3, \ldots, n$ y $j = i, i+1, \ldots, n$.

Algoritmo de factorización directa de Doolittle o de Crout.

==========

Para factorizar una matriz $A = (a_{ij})$ de $n \times n$ en el producto de la matriz triangular inferior $L = (l_{ij})$ con la matriz triangular superior $U = (u_{ij})$; esto es, $A = L\ U$, donde está dada la diagonal principal de L ó U.

Entrada: dimensión n; los elementos $a_{ij}, 1 \leq i, j \leq n$ de A; la diagonal $l_{11}, l_{22}, \ldots, l_{nn}$ de L (método de Doolittle) ó $u_{11}, u_{22}, \ldots, u_{nn}$ de U (método de Crout).

Salida: los elementos l_{ij}, $1 \leq j \leq i$, $1 \leq i \leq n$ de L y los elementos u_{ij}, $1 \leq i \leq n$, $i \leq j \leq n$, de U.

Paso 1: Seleccionar l_{11} y u_{11} satisfaciendo $l_{11}\ u_{11} = a_{11}$.
Si $l_{11}\ u_{11} = 0$ entonces SALIDA; *(factorización imposible)* PARAR.

Paso 2: Para $j = 2, 3, \ldots, n$ tomar
$u_{1j} = \frac{a_{1j}}{l_{11}}$; *(primera fila de U)*;
$l_{j1} = \frac{a_{j1}}{u_{11}}$; *(primera columna de L).*

Paso 3: Para $i = 2, 3, \ldots, n-1$ seguir los pasos 4 y 5.

Paso 4: Seleccionar l_{ii} y u_{ii} satisfaciendo $l_{ii}\ u_{ii} = a_{ii} - \sum_{k=1}^{i-1} l_{ik}\ u_{ki}$.
Si $l_{ii}\ u_{ii} = 0$ entonces SALIDA; *(factorización imposible)*
PARAR.

Paso 5: Para $j = i+1, i+2, \ldots, n$ tomar
$u_{ij} = \frac{1}{l_{ii}}\ [a_{ij} - \sum_{k=1}^{i-1} l_{ik}\ u_{kj}]$; *(i−ésima fila de U)*;
$l_{ji} = \frac{1}{u_{ii}}\ [a_{ji} - \sum_{k=1}^{i-1} l_{jk}\ u_{ki}]$; *(i−ésima columna de L).*

Paso 6: Seleccionar l_{nn} y u_{nn} satisfaciendo $l_{nn}\ u_{nn} = a_{nn} - \sum_{k=1}^{n-1} l_{nk}\ u_{kn}$.
Si $l_{nn}\ u_{nn} = 0$ entonces $A = L\ U$ pero A es singular.

Paso 7: SALIDA (l_{ij} y u_{ij} para $j = 1, \ldots, n$ e $i = 1, \ldots, n$); PARAR.

==========

Una dificultad que puede surgir cuando se usa este algoritmo para obtener la factorización de la matriz de coeficientes de un sistema lineal de ecuaciones es la causada por el hecho de que no se usa pivoteo para reducir el efecto del error de redondeo. Se ha visto en cálculos anteriores que el error de redondeo puede ser muy significativo cuando se usa aritmética de dígitos finitos y que cualquier algoritmo eficiente debe de tomar esto en consideración.

Aún cuando el intercambio de columnas es difícil de incorporar en el algoritmo de factorización, el algoritmo puede alterarse fácilmente para incluir una técnica de intercambio de filas equivalente al procedimiento de pivoteo máximo de columna descrito en el capítulo XV. Este intercambio resulta suficiente en la mayoría de los casos.

El siguiente algoritmo incorpora el procedimiento de factorización del algoritmo de factorización directa junto con el pivoteo máximo de columna y la sustitución hacia adelante y hacia atrás para obtener una solución a un sistema lineal de ecuaciones. El proceso requiere que el sistema lineal $A\ \mathbf{x} = \mathbf{b}$ se escriba como $L\ U\ \mathbf{x} = \mathbf{b}$. La sustitución hacia adelante resuelve el sistema $L\ \mathbf{z} = \mathbf{b}$ y la sustitución hacia atrás resuelve al sistema $U\ \mathbf{x} = L^{-1}\ \mathbf{b} = \mathbf{z}$. Se debe hacer notar que los elementos diferentes de cero de L y U se pueden guardar en los elementos correspondientes de A excepto los de la diagonal de L ó U, la cual debe darse en entrada.

Algoritmo de factorización directa con pivoteo máximo de columna.

==========

Para resolver el sistema lineal $n \times n$ $A\ \mathbf{x} = \mathbf{b}$ en la forma:

$$\begin{aligned} E_1: &\quad a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n = a_{1,n+1} \\ E_2: &\quad a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n = a_{2,n+1} \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n = a_{n,n+1} \end{aligned}$$

factorizando A en $L\ U$ y resolviendo $L\ \mathbf{z} = \mathbf{b}$ y $U\ \mathbf{x} = \mathbf{z}$ donde se da la diagonal principal de L o U.

Entrada: dimensión n; los elementos a_{ij}, $1 \le i \le n$, $1 \le j \le n+1$ de la matriz ampliada de A; la diagonal $l_{11}, l_{22}, \ldots, l_{nn}$ de L (método de Doolittle) o la diagonal $u_{11}, u_{22}, \ldots, u_{nn}$ de U (método de Crout).

Salida: solución $x_1, x_2, \ldots, x_n$ ó mensaje de que el sistema lineal no tiene solución única.

Paso 1: Sea p el menor entero tal que $1 \le p \le n$ y $|a_{p1}| = \max\limits_{1 \le j \le n} |a_{j1}|$; *(encontrar el primer elemento pivote).*
Si $|a_{p1}| = 0$ SALIDA; *(no existe solución única)* PARAR.

Paso 2: Si $p \neq 1$ entonces intercambiar las filas p y 1 en A.

Paso 3: Seleccionar l_{11} y u_{11} satisfaciendo $l_{11}\ u_{11} = a_{11}$.

Paso 4: Para $j = 2, 3, \ldots, n$ tomar
$u_{1j} = \frac{a_{1j}}{l_{11}}$; *(primera fila de U)*;
$l_{j1} = \frac{a_{j1}}{u_{11}}$; *(primera columna de L).*

Paso 5: Para $i = 2, 3, \ldots, n-1$ seguir los pasos 6–9.

Paso 6: Sea p el menor entero tal que $i \le p \le n$ y

$$\left| a_{pi} - \sum_{k=1}^{i-1} l_{pk}\ u_{ki} \right| = \max_{i \le j \le n} \left| a_{ji} - \sum_{k=1}^{i-1} l_{jk}\ u_{ki} \right| ;$$

(encontrar el $i-$ésimo elemento pivote).
Si el máximo es cero entonces SALIDA;
(no existe solución única) PARAR.

Paso 7: Si $p \neq i$ entonces intercambiar las filas p e i en la matriz A e intercambiar los elementos de las filas p e i de las primeras $(i-1)$ columnas de L.

Paso 8: Seleccionar l_{ii} y u_{ii} satisfaciendo $l_{ii}\ u_{ii} = a_{ii} - \sum\limits_{k=1}^{i-1} l_{ik}\ u_{ki}$.

Paso 9: Para $j = i+1, i+2, \ldots, n$ tomar

$u_{ij} = \frac{1}{l_{ii}} \left[a_{ij} - \sum\limits_{k=1}^{i-1} l_{ik}\ u_{kj} \right]$; *($i-$ésima fila de U)*;

$l_{ji} = \frac{1}{u_{ii}} \left[a_{ji} - \sum\limits_{k=1}^{i-1} l_{jk}\ u_{ki} \right]$; *($i-$ésima columna de L).*

Paso 10: Tomar $AUX = a_{nn} - \sum\limits_{k=1}^{n-1} l_{nk}\ u_{kn}$.

Si $AUX = 0$ entonces SALIDA; *(no existe solución única)* PARAR.

Seleccionar l_{nn} y u_{nn} que satisfagan $l_{nn}\ u_{nn} = a_{nn} - \sum\limits_{k=1}^{n-1} l_{nk}\ u_{kn}$.

(Los pasos 11 y 12 resuelven el sistema triangular inferior $L\ \mathbf{z} = \mathbf{b}$.)

Paso 11: Tomar $z_1 = \frac{a_{1,n+1}}{l_{11}}$.

Paso 12: Para $i = 2, 3, \ldots, n$ tomar
$z_i = \frac{1}{l_{ii}}\ [a_{i,n+1} - \sum_{j=1}^{i-1} l_{ij}\ z_j]$.

(Los pasos 13 y 14 resuelven el sistema triangular superior $U\ \mathbf{x} = \mathbf{z}$.)

Paso 13: Tomar $x_n = \frac{z_n}{u_{nn}}$.

Paso 14: Para $i = n-1, n-2, \ldots, 1$ tomar
$x_i = \frac{1}{u_{ii}}\ [z_i - \sum_{j=i+1}^{n} u_{ij}\ x_j]$.

Paso 15: SALIDA $(x_1, x_2, \ldots, x_n)$;
(procedimiento completado satisfactoriamente) PARAR.

==========

Ejemplo. Para ilustrar el procedimiento seguido en el algoritmo de factorización directa con pivoteo máximo de columna, consideremos el sistema lineal

$$\begin{array}{rcrcrcrcl} 1.00\ x_1 & + & 0.333\ x_2 & + & 1.50\ x_3 & - & 0.333\ x_4 & = & 3.00\ , \\ -2.01\ x_1 & + & 1.45\ x_2 & + & 0.50\ x_3 & + & 2.95\ x_4 & = & 5.40\ , \\ 4.32\ x_1 & - & 1.95\ x_2 & & & + & 2.08\ x_4 & = & 0.13\ , \\ 5.11\ x_1 & - & 4.00\ x_2 & + & 3.33\ x_3 & - & 1.11\ x_4 & = & 3.77\ . \end{array}$$

Seguiremos los pasos del algoritmo de factorización directa con pivoteo máximo de columna con $l_{11} = l_{22} = l_{33} = l_{44} = 1$, usando aritmética de redondeo a tres dígitos. En primer lugar escribimos la matriz ampliada:

$$A_a = [A, \mathbf{b}] = \left(\begin{array}{rrrr|r} 1.00 & 0.333 & 1.50 & -0.333 & 3.00 \\ -2.01 & 1.45 & 0.50 & 2.95 & 5.40 \\ 4.32 & -1.95 & 0.00 & 2.08 & 0.13 \\ 5.11 & -4.00 & 3.33 & -1.11 & 3.77 \end{array}\right) .$$

Además, las matrices triangular inferior L y triangular superior U son:

$$L = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ l_{21} & 1.00 & 0 & 0 \\ l_{31} & l_{32} & 1.00 & 0 \\ l_{41} & l_{42} & l_{43} & 1.00 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix} .$$

Paso 1: Tenemos que encontrar el primer elemento pivote, es decir, el menor entero p tal que $1 \le p \le n$ y $|a_{p1}| = \max_{1 \le j \le n} |a_{j1}|$. En nuestro caso

$$p = 4\ .$$

Paso 2: Dado que $p \neq 1$, entonces tenemos que intercambiar las filas $p = 4$ y 1 en A. La matriz ampliada se transforma en

$$[A, \mathbf{b}] = \left(\begin{array}{rrrr|r} 5.11 & -4.00 & 3.33 & -1.11 & 3.77 \\ -2.01 & 1.45 & 0.50 & 2.95 & 5.40 \\ 4.32 & -1.95 & 0.00 & 2.08 & 0.13 \\ 1.00 & 0.333 & 1.50 & -0.333 & 3.00 \end{array}\right) .$$

Paso 3: Se necesita seleccionar l_{11} y u_{11} satisfaciendo $l_{11}\ u_{11} = a_{11} = 5.11$. Y como $l_{11} = 1.00$,

$$u_{11} = 5.11$$

Paso 4: Para $j = 2, 3, 4$ debemos tomar $u_{1j} = \frac{a_{1j}}{l_{11}}$ y $l_{j1} = \frac{a_{j1}}{u_{11}}$.
Es decir,

$$u_{12} = \frac{a_{12}}{l_{11}} = -4.00\ , \quad u_{13} = \frac{a_{13}}{l_{11}} = 3.33\ , \quad u_{14} = \frac{a_{14}}{l_{11}} = -1.11\ ,$$

y

$$l_{21} = \frac{a_{21}}{u_{11}} = \frac{-2.01}{5.11} = -0.393\ ,$$
$$l_{31} = \frac{a_{31}}{u_{11}} = \frac{4.32}{5.11} = 0.845\ ,$$
$$l_{41} = \frac{a_{41}}{u_{11}} = \frac{1.00}{5.11} = 0.196\ .$$

Entonces, las matrices L y U asumen la forma

$$L = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ -0.393 & 1.00 & 0 & 0 \\ 0.845 & l_{32} & 1.00 & 0 \\ 0.196 & l_{42} & l_{43} & 1.00 \end{pmatrix} \text{ y } U = \begin{pmatrix} 5.11 & -4.00 & 3.33 & -1.11 \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix}\ .$$

Paso 5: Para $i = 2$ seguir los pasos 6–9.

Paso 6: Ahora tenemos que encontrar el segundo elemento pivote, es decir, encontrar el menor entero p tal que $2 \le p \le 4$ y

$$|a_{p2} - l_{p1}\ u_{12}| = \max_{2 \le j \le 4} |a_{j2} - l_{j1}\ u_{12}|\ .$$

En nuestro caso,

$$|a_{22} - l_{21}\ u_{12}| = |1.45 - (-0.393)(-4.00)| = |-0.12| = 0.12\ ,$$
$$|a_{32} - l_{31}\ u_{12}| = |-1.95 - (0.845)(-4.00)| = |1.43| = 1.43\ ,$$
$$|a_{42} - l_{41}\ u_{12}| = |0.333 - (0.196)(-4.00)| = |1.12| = 1.12\ .$$

Así, $p = 3$.

Paso 7: Dado que $p = 3 \neq 2 = i$, tenemos que intercambiar las filas $p = 3$ e $i = 2$ en la matriz A e intercambiar los elementos de las filas $p = 3$ e $i = 2$ de la primera columna de L. Entonces,

$$[A, \mathbf{b}] = \left(\begin{array}{cccc|c} 5.11 & -4.00 & 3.33 & -1.11 & 3.77 \\ 4.32 & -1.95 & 0.00 & 2.08 & 0.13 \\ -2.01 & 1.45 & 0.50 & 2.95 & 5.40 \\ 1.00 & 0.333 & 1.50 & -0.333 & 3.00 \end{array}\right)\ ,$$

$$L = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ 0.845 & 1.00 & 0 & 0 \\ -0.393 & l_{32} & 1.00 & 0 \\ 0.196 & l_{42} & l_{43} & 1.00 \end{pmatrix}\ .$$

Paso 8: Tenemos que seleccionar l_{22} y u_{22} satisfaciendo

$$l_{22}\ u_{22} = a_{22} - l_{21}\ u_{12}\ .$$

Dado que $l_{22} = 1.00$, entonces

$$u_{22} = a_{22} - l_{21}\ u_{12} = -1.95 - (0.845)(-4.00) = 1.43\ .$$

Paso 9: Para $j = 3, 4$ tenemos que tomar

$$u_{2j} = \frac{1}{l_{22}}\ [a_{2j} - l_{2k}\ u_{kj}]$$

$$l_{j2} = \frac{1}{u_{22}}\ [a_{j2} - l_{jk}\ u_{k2}]$$

En nuestro caso,

$$\begin{aligned}
u_{23} &= \frac{1}{l_{22}}\ [a_{23} - l_{21}\ u_{13}] = [0.00 - (0.845)(3.33)] = -2.81\ , \\
u_{24} &= \frac{1}{l_{22}}\ [a_{24} - l_{21}\ u_{14}] = [2.08 - (0.845)(-1.11)] = 3.01\ , \\
l_{32} &= \frac{1}{u_{22}}\ [a_{32} - l_{31}\ u_{12}] = \frac{1}{1.43}\ [1.45 - (-0.393)(-4.00)] = -0.0839\ , \\
l_{42} &= \frac{1}{u_{22}}\ [a_{42} - l_{41}\ u_{12}] = \frac{1}{1.43}\ [0.333 - (0.196)(-4.00)] = 0.783\ .
\end{aligned}$$

Entonces, las matrices L y U asumen la forma

$$L = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ 0.845 & 1.00 & 0 & 0 \\ -0.393 & -0.0839 & 1.00 & 0 \\ 0.196 & 0.783 & l_{43} & 1.00 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 5.11 & -4.00 & 3.33 & -1.11 \\ 0 & 1.43 & -2.81 & 3.02 \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{pmatrix}\ .$$

Paso 5: Para $i = 3$ seguir los pasos 6–9.

Paso 6: Ahora tenemos que encontrar el tercer elemento pivote, es decir, encontrar el menor entero p tal que $3 \leq p \leq 4$ y

$$|a_{p3} - \sum_{k=1}^{2} l_{pk}\ u_{k3}| = \max_{3 \leq j \leq 4} |a_{j3} - \sum_{k=1}^{2} l_{jk}\ u_{k3}|\ .$$

En nuestro caso,

$$\begin{aligned}
|a_{33} - (l_{31}\ u_{13} + l_{32}\ u_{23})| &= |0.5 - ((-0.393)(3.33) + (-0.0839)(-2.81))| = 1.57\ , \\
|a_{43} - (l_{41}\ u_{13} + l_{42}\ u_{23})| &= |1.5 - ((0.196)(3.33) + (0.783)(-2.81)))| = 3.05\ .
\end{aligned}$$

Así, $p = 4$.

Paso 7: Dado que $p = 4 \neq 3 = i$, tenemos que intercambiar las filas $p = 5$ y $i = 3$ en la matriz A e intercambiar los elementos de las filas $p = 5$ e $i = 3$ de la primera y segunda columnas de L. Entonces,

$$[A, \mathbf{b}] = \left(\begin{array}{cccc|c} 5.11 & -4.00 & 3.33 & -1.11 & 3.77 \\ 4.32 & -1.95 & 0.00 & 2.08 & 0.13 \\ 1.00 & 0.333 & 1.50 & -0.333 & 3.00 \\ -2.01 & 1.45 & 0.50 & 2.95 & 5.40 \end{array} \right) ,$$

$$L = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ 0.845 & 1.00 & 0 & 0 \\ 0.196 & 0.783 & 1.00 & 0 \\ -0.393 & -0.0839 & l_{43} & 1.00 \end{pmatrix} .$$

Paso 8: Tenemos que seleccionar l_{33} y u_{33} satisfaciendo

$$l_{33}\ u_{33} = a_{33} - (l_{31}\ u_{13} + l_{32}\ u_{23}) .$$

Dado que $l_{33} = 1.00$, entonces

$$u_{33} = a_{33} - (l_{31}\ u_{13} + l_{32}\ u_{23}) = 1.50 - (0.196)(3.33) + (-0.0839)(-2.81) = 3.05 .$$

Paso 9: Para $j = 4$ tenemos que tomar

$$u_{3j} = \frac{1}{l_{33}}\ [a_{3j} - (l_{31}\ u_{1j} + l_{32}\ u_{2j})]$$

$$l_{j3} = \frac{1}{u_{33}}\ [a_{j3} - (l_{j1}\ u_{13} + l_{j2}\ u_{23})] .$$

En nuestro caso,

$$\begin{aligned} u_{34} &= \frac{1}{l_{33}}\ [a_{34} - (l_{31}\ u_{14} + l_{32}\ u_{24})] \\ &= [-0.333 - ((0.196)(-1.11) + (0.783)(3.02))] = -2.47 , \\ l_{43} &= \frac{1}{u_{33}}\ [a_{43} - (l_{41}\ u_{13} + l_{42}\ u_{23})] \\ &= \frac{1}{3.05}\ [0.5 - ((-0.393)(3.33) + (-0.0839)(-2.81))] = 0.515 . \end{aligned}$$

Entonces, las matrices L y U asumen la forma

$$L = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ 0.845 & 1.00 & 0 & 0 \\ 0.196 & 0.783 & 1.00 & 0 \\ -0.393 & -0.0839 & 0.515 & 1.00 \end{pmatrix} \text{ y } U = \begin{pmatrix} 5.11 & -4.00 & 3.33 & -1.11 \\ 0 & 1.43 & -2.81 & 3.02 \\ 0 & 0 & 3.05 & -2.47 \\ 0 & 0 & 0 & u_{44} \end{pmatrix} .$$

Paso 10: Finalmente, tenemos que seleccionar l_{44} y u_{44} que satisfagan

$$l_{44}\ u_{44} = a_{44} - \sum_{k=1}^{3} l_{4k}\ u_{k4} .$$

Dado que $l_{44} = 1.00$, entonces

$$\begin{aligned} u_{44} &= a_{44} - (l_{41}\ u_{14} + l_{42}\ u_{24} + l_{43}\ u_{23}) \\ &= 2.95 - ((-0.393)(-1.11) + (-0.0839)(3.02) + (0.515)(-2.47)) = 4.04\ . \end{aligned}$$

La factorización está completa:

$$A = \begin{pmatrix} 5.11 & -4.00 & 3.33 & -1.11 \\ 4.32 & -1.95 & 0.00 & 2.08 \\ 1.00 & 0.333 & 1.50 & -0.333 \\ -2.01 & 1.45 & 0.50 & 2.95 \end{pmatrix} =$$

$$= \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ 0.845 & 1.00 & 0 & 0 \\ 0.196 & 0.783 & 1.00 & 0 \\ -0.393 & -0.0839 & 0.515 & 1.00 \end{pmatrix} \begin{pmatrix} 5.11 & -4.00 & 3.33 & -1.11 \\ 0 & 1.43 & -2.81 & 3.02 \\ 0 & 0 & 3.05 & -2.47 \\ 0 & 0 & 0 & 4.04 \end{pmatrix}\ .$$

(Los pasos 11 y 12 resuelven el sistema triangular inferior $L\ \mathbf{z} = \mathbf{b}$.)

Paso 11: Tomar $z_1 = \frac{a_{1,5}}{l_{11}} = \frac{3.77}{1.00} = 3.77$.

Paso 12: Para $i = 2, 3, 4$ tomar

$$z_i = \frac{1}{l_{ii}}\ [a_{i,n+1} - \sum_{j=1}^{i-1} l_{ij}\ z_j]\ .$$

En nuestro caso:

$$\begin{aligned} z_2 &= \frac{1}{l_{22}}\ [a_{25} - l_{21}\ z_1] \\ &= 0.13 - (0.845)(3.77) = -3.06 \\ z_3 &= \frac{1}{l_{33}}\ [a_{35} - (l_{31}\ z_1 + l_{32}\ z_2)] \\ &= 3.00 - ((0.196)(3.77) + (0.783)(-3.06)) = 4.66 \\ z_4 &= \frac{1}{l_{44}}\ [a_{45} - (l_{41}\ z_1 + l_{42}\ z_2 + l_{43}\ z_3)] \\ &= 5.40 - ((-0.393)(3.77) + (-0.0839)(-3.06) + (0.515)(4.66)) = 4.22\ . \end{aligned}$$

(Los pasos 13 y 14 resuelven el sistema triangular superior $U\ \mathbf{x} = \mathbf{z}$.)

Paso 13: Tomar $x_4 = \frac{z_4}{u_{44}} = \frac{4.22}{4.04} = 1.04$.

Paso 14: Para $i = 3, 2, 1$ tomar

$$x_i = \frac{1}{u_{ii}}\ [z_i - \sum_{j=i+1}^{n} u_{ij}\ x_j]\ .$$

En nuestro caso:

$$
\begin{aligned}
x_3 &= \frac{1}{u_{33}}\,[z_3 - u_{34}\,x_4] \\
&= \frac{1}{3.05}\,[4.66 - (-2.47)(1.04)] = 2.37 \\
x_2 &= \frac{1}{u_{22}}\,[z_2 - (u_{23}\,x_3 + u_{24}\,x_4)] \\
&= \frac{1}{1.43}\,[-3.06 - ((-2.81)(2.37) + (3.02)(1.04))] = 0.322 \\
x_1 &= \frac{1}{u_{33}}\,[z_3 - (u_{12}\,x_2 + u_{13}\,x_3 + u_{14}\,x_4)] \\
&= \frac{1}{5.11}\,[3.77 - ((-4.00)(0.322) + (3.33)(2.37) + (-1.11)(1.04))] = -0.329\ .
\end{aligned}
$$

Paso 15: SALIDA. La solución es

$$x_1 = -0.329\ ,\quad x_2 = 0.322\ ,\quad x_3 = 2.37\ ,\quad x_4 = 1.04\ .$$

(procedimiento completado satisfactoriamente) PARAR.

Una aplicación del algoritmo de factorización directa da lugar a la factorización

$$
A = \begin{pmatrix} 1.00 & 0.333 & 1.50 & -0.333 \\ -2.01 & 1.45 & 0.50 & 2.95 \\ 4.32 & -1.95 & 0.00 & 2.08 \\ 5.11 & -4.00 & 3.33 & -1.11 \end{pmatrix} =
$$

$$
= \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ -2.01 & 1.00 & 0 & 0 \\ 4.32 & -1.60 & 1.00 & 0 \\ 5.11 & -2.69 & -6.04 & 1.00 \end{pmatrix} \begin{pmatrix} 1.00 & 0.333 & 1.50 & -0.333 \\ 0 & 2.12 & 3.52 & 2.28 \\ 0 & 0 & -0.85 & 7.17 \\ 0 & 0 & 0 & 50.0 \end{pmatrix}\ .
$$

Aplicando entonces los pasos 11 hasta el 15 del algoritmo de factorización directa con pivoteo máximo de columna se obtiene la solución

$$x_1 = -0.370\ ,\quad x_2 = 0.236\ ,\quad x_3 = 2.42\ ,\quad x_4 = 1.03\ .$$

La siguiente tabla compara los resultados del algoritmo de factorización directa con pivoteo máximo de columna, del algoritmo de factorización directa y de la respuesta real a tres dígitos. Nótese la mejoría en la precisión cuando se incluyen intercambios de filas.

Tabla 1

	x_1	x_2	x_3	x_4
Alg. fact. pivoteo	−0.329	0.322	2.37	1.04
Alg. fact. directa	−0.370	0.236	2.42	1.03
Real	−0.324	0.321	2.37	1.04

3. EL ALGORITMO DE CHOLESKY

Cuando se sabe que la matriz real es simétrica y positiva definida, se puede mejorar significativamente la técnica de factorización de una matriz con respecto al número de operaciones aritméticas requeridas.

Teorema XVI.1

Si A es una matriz real de $n \times n$ simétrica y positiva definida, entonces A tiene una factorización de la forma $A = L\ L^t$, donde L es una matriz triangular inferior. La factorización se puede lograr aplicando el algoritmo de factorización directa con $l_{ii} = u_{ii}$ para cada $i = 1, 2, \ldots, n$.

Para una matriz simétrica y positiva definida, este Teorema se puede usar para simplificar el algoritmo de factorización directa. Además, si se tiene que resolver un sistema lineal representado por una matriz positiva definida, los pasos 1–6 del siguiente algoritmo (**algoritmo de Choleski**) pueden sustituirse por los pasos 1–10 del algoritmo de factorización directa con pivoteo máximo de columna para aprovechar la simplificación que resulta, siempre y cuando u_{ij} sea reemplazado por l_{ij} en los pasos 13 y 14. El procedimiento de factorización se describe en el siguiente algoritmo.

Algoritmo de Choleski.

==========

Para factorizar una matriz $n \times n$ simétrica y positiva definida $A = (a_{ij})$ como $A = L\ L^t$, donde L es triangular inferior.

Entrada: dimensión n; los elementos $a_{ij}, 1 \leq i, j \leq n$ de A.

Salida: los elementos l_{ij}, $1 \leq j \leq i$, $1 \leq i \leq n$ de L; (los elementos de $U = L^t$ son $u_{ij} = l_{ji}$, $i \leq j \leq n$, $1 \leq i \leq n$).

Paso 1: Tomar

$$l_{11} = \sqrt{a_{11}}\ .$$

Paso 2: Para $j = 2, 3, \ldots, n$ tomar

$$l_{j1} = \frac{a_{j1}}{l_{11}}\ .$$

Paso 3: Para $i = 2, 3, \ldots, n-1$ seguir los pasos 4 y 5.

Paso 4: Tomar

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}\ .$$

Paso 5: Para $j = i+1, i+2, \ldots, n$ tomar

$$l_{ji} = \frac{1}{l_{ii}}\ [a_{ji} - \sum_{k=1}^{i-1} l_{jk}\ l_{ik}]\ .$$

Paso 6: Tomar

$$l_{nn} = \sqrt{a_{nn} - \sum_{k=1}^{n-1} l_{nk}^2}\ .$$

Paso 7: SALIDA (l_{ij} para $j = 1, \ldots, i$ e $i = 1, \ldots, n$); PARAR.

==========

La solución de un sistema lineal típico representado por una matriz positiva definida usando el algoritmo de Cholesky requiere de

raíces cuadradas

$$n$$

multiplicaciones/divisiones

$$\frac{n^3 + 9\ n^2 + 2\ n}{6}$$

sumas/restas

$$\frac{n^3 + 6\ n^2 - 7\ n}{6}\ .$$

Estas son alrededor de la mitad de las operaciones aritméticas requeridas en el algoritmo de eliminación Gaussiana. La vantaja computacional del método de Cholesky depende del número de operaciones que se requieran para determinar los valores de las n raíces cuadradas, el cual, debido a que es un factor lineal con n, decrecerá significativamente conforme n crezca.

4. EL ALGORITMO DE CROUT PARA SISTEMAS TRIDIAGONALES

Los algoritmos de factorización se pueden simplificar considerablemente en el caso de matrices de banda debido al gran número de ceros que aparecen en patrones regulares en estas matrices. Es particularmente interesante observar la forma que los métodos de Crout o Doolittle toman en este caso. Para ilustrar esta situación, supongamos que una matriz tridiagonal

$$A = \begin{pmatrix} a_{11} & a_{12} & 0 & \ldots & \ldots & 0 \\ a_{21} & a_{22} & a_{23} & 0 & \ldots & 0 \\ 0 & a_{32} & a_{33} & a_{34} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ 0 & \ldots & \ldots & 0 & a_{n,n-1} & a_{nn} \end{pmatrix},$$

pueda factorizarse en las matrices triangulares L y U.

Como A tiene solamente $(3\ n - 2)$ elementos distintos de cero, habrá sólo $(3\ n - 2)$ condiciones para determinar a los elementos de L y U siempre y cuando se obtengan también los elementos cero de A. Supongamos que realmente es posible encontrar las matrices en la forma

$$L = \begin{pmatrix} l_{11} & 0 & \ldots & \ldots & \ldots & 0 \\ l_{21} & l_{22} & 0 & \ldots & \ldots & 0 \\ 0 & l_{32} & l_{33} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & l_{n-1,n-2} & l_{n-1,n-1} & 0 \\ 0 & \ldots & \ldots & 0 & l_{n,n-1} & l_{nn} \end{pmatrix},$$

y

$$U = \begin{pmatrix} 1 & u_{12} & 0 & \dots & \dots & 0 \\ 0 & 1 & u_{23} & 0 & \dots & 0 \\ 0 & 0 & 1 & u_{34} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & 1 & u_{n-1,n} \\ 0 & \dots & \dots & 0 & 0 & 1 \end{pmatrix} .$$

De esta forma hay $(2\ n-1)$ elementos indeterminados de L y $(n-1)$ elementos indeterminados de U, que en total son iguales, en número, a las condiciones mencionadas arriba y además, los elementos cero de A se obtienen automáticamente.

La multiplicación $A = L\ U$ da, sin contar los elementos cero, las ecuaciones:

$$\begin{aligned} a_{11} &= l_{11} \ , \\ a_{i,i-1} &= l_{i,i-1} \ , \quad \text{para cada } \ i = 2, 3, \dots, n \ , \\ a_{ii} &= l_{i,i-1}\ u_{i-1,i} + l_{ii} \ , \quad \text{para cada } \ i = 2, 3, \dots, n \ , \\ a_{i,i+1} &= l_{ii}\ u_{i,i+1} \ , \quad \text{para cada } \ i = 1, 2, \dots, n-1 \ . \end{aligned}$$

Una solución a este sistema de ecuaciones puede encontarse obteniendo primero todos los términos no cero fuera de la diagonal de L, usando la segunda ecuación y luego usando la cuarta y la tercera para obtener alternadamente el resto de los elementos de U y L, los cuales se pueden ir guardando en los elementos correspondientes de A.

A continuación se da un algoritmo completo para resolver un sistema de ecuaciones lineales de $n \times n$ cuya matriz de coeficientes es tridiagonal.

Algoritmo de reducción de Crout para sistemas lineales tridiagonales.

==========

Para resolver el sistema lineal tridiagonal de $n \times n$

$$\begin{aligned} E_1 &: a_{11}\ x_1 + a_{12}\ x_2 = a_{1,n+1} \ , \\ E_2 &: a_{21}\ x_1 + a_{22}\ x_2 + a_{23}\ x_3 = a_{2,n+1} \ , \\ &\dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \\ E_{n-1} &: a_{n-1,n-2}\ x_{n-2} + a_{n-1,n-1}\ x_{n-1} + a_{n-1,n}\ x_n = a_{n-1,n+1} \ , \\ E_n &: a_{n,n-1}\ x_{n-1} + a_{nn}\ x_n = a_{n,n+1} \ . \end{aligned}$$

el cual se supone tiene solución única.

Entrada: dimensión n; los elementos a_{ij}, $1 \le i \le n$ y $1 \le j \le n+1$ de A_a.

Salida: solución $x_1, x_2, \dots, x_n$.

Paso 1: Tomar

$$l_{11} = a_{11} \quad \text{y} \quad u_{12} = \frac{a_{12}}{l_{11}} \ .$$

Paso 2: Para $i = 2, 3, \dots, n-1$ tomar
$l_{i,i-1} = a_{i,i-1}$; ($i-$*ésima fila de* L).
$l_{ii} = a_{ii} - l_{i,i-1}\ u_{i-1,i}$.
$u_{i,i+1} = \frac{a_{i,i+1}}{l_{ii}}$; ($(i+1)-$*ésima columna de* U).

Paso 3: Tomar $l_{n,n-1} = a_{n,n-1}$; (*$n-$ésima fila de* L).
$l_{nn} = a_{nn} - l_{n,n-1}\ u_{n-1,n}$.
(Los pasos 4 y 5 resuelven $L\ \mathbf{z} = \mathbf{b}$).

Paso 4: Tomar

$$z_1 = \frac{a_{1,n+1}}{l_{11}}\ .$$

Paso 5: Para $i = 2, 3, \ldots, n$ tomar

$$z_i = \frac{1}{l_{ii}}\ [a_{i,n+1} - l_{i,i-1}\ z_{i-1}]\ .$$

(Los pasos 6 y 7 resuelven $U\ \mathbf{x} = \mathbf{z}$).

Paso 6: Tomar

$$x_n = z_n\ .$$

Paso 7: Para $i = n-1, n-2, \ldots, 1$ tomar

$$x_i = z_i - u_{i,i+1}\ x_{i+1}\ .$$

Paso 8: SALIDA $(x_1, x_2, \ldots, x_n\ )$; PARAR.

==========

Este algoritmo requiere sólo de $(5\ n-4)$ multiplicaciones/divisiones y de $(3\ n-3)$ sumas/restas, y consecuentemente tiene una ventaja computacional considerable sobre los métodos que no consideran la triadiagonalidad de la matriz, especialmente para valores grandes de n.

El algoritmo de reducción de Crout para sistemas lineales tridiagonales puede aplicarse cuando $l_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$. Dos condiciones, cualquiera de las cuales asegurará que esto es cierto, son que la matriz de coeficientes del sistema sea positiva definida o que sea estrictamente dominante diagonalmente. Una condición adicional que garantiza que este algoritmo se puede aplicar está dada en el siguiente Teorema.

Teorema XVI.2

Supóngase que $A = (a_{ij})$ es tridiagonal con $a_{i,i-1} \cdot a_{i,i+1} \neq 0$ para cada $i = 2, 3, \ldots, n-1$. Si $|a_{11}| > |a_{12}|$, $|a_{ii}| > |a_{i,i-1}| + |a_{i,i+1}|$ para cada $i = 2, 3, \ldots, n-1$, y $|a_{nn}| > |a_{n,n-1}|$, entonces A es no singular y los valores de l_{ii} descritos en el algoritmo de reducción de Crout son diferentes de cero para cada $i = 1, 2, \ldots, n$.

CAPITULO XVII. TECNICAS ITERATIVAS PARA RESOLVER SISTEMAS LINEALES

1. INTRODUCCION Y METODO

Una técnica iterativa para resolver un sistema lineal $A\,\mathbf{x} = \mathbf{b}$ de $n \times n$ empieza con una aproximación inicial $\mathbf{x}^{(0)}$ a la solución $\mathbf{x}$, y genera una sucesión de vectores $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ que converge a $\mathbf{x}$. La mayoría de estas técnicas iterativas involucran un proceso que convierte el sistema $A\,\mathbf{x} = \mathbf{b}$ en un sistema equivalente de la forma $\mathbf{x} = T\,\mathbf{x} + \mathbf{c}$ para alguna matriz T de $n \times n$ y un vector $\mathbf{c}$. Ya seleccionado el vector inicial $\mathbf{x}^{(0)}$ la sucesión de vectores de solución aproximada se genera calculando

$$\mathbf{x}^{(k)} = T\,\mathbf{x}^{(k-1)} + \mathbf{c} \qquad (XVII.1)$$

para cada $k = 1, 2, 3, \ldots$. Este tipo de procedimiento nos recuerda a la iteración del punto fijo estudiada en la tercera parte.

Las técnicas iterativas se emplean raras veces para resolver sistemas lineales de dimensión pequeña ya que el tiempo requerido para lograr una precisión suficiente excede al de las técnicas directas como el método de eliminación Gaussiana. Sin embargo, para sistemas grandes con un gran porcentaje de ceros, estas técnicas son eficientes en términos de almacenamiento en la computadora y del tiempo requerido. Los sistemas de este tipo surgen frecuentemente en la solución numérica de problemas de valores en la frontera y de ecuaciones diferenciales parciales.

Ejemplo. El sistema lineal $A\,\mathbf{x} = \mathbf{b}$ dado por

$$\begin{array}{lrcrcrcrcr}
E_1: & 10\,x_1 & - & x_2 & + & 2\,x_3 & & & = & 6\,, \\
E_2: & -\,x_1 & + & 11\,x_2 & - & x_3 & + & 3\,x_4 & = & 25\,, \\
E_3: & 2\,x_1 & - & x_2 & + & 10\,x_3 & - & x_4 & = & -11\,, \\
E_4: & & & 3\,x_2 & - & x_3 & + & 8\,x_4 & = & 15\,,
\end{array}$$

tiene por solución a $\mathbf{x} = (1, 2, -1, 1)^t$. Para convertir $A\,\mathbf{x} = \mathbf{b}$ a la forma $\mathbf{x} = T\,\mathbf{x} + \mathbf{c}$, resolvemos la ecuación E_i para cada $i = 1, 2, 3, 4$, obteniendo:

$$\begin{array}{lrcrcrcrcr}
x_1 = & & & \frac{1}{10}x_2 & - & \frac{1}{5}x_3 & & & + & \frac{3}{5}\,, \\
x_2 = & \frac{1}{11}x_1 & & & + & \frac{1}{11}x_3 & - & \frac{3}{11}x_4 & + & \frac{25}{11}\,, \\
x_3 = & -\frac{1}{5}x_1 & + & \frac{1}{10}x_2 & & & + & \frac{1}{10}x_4 & - & \frac{11}{10}\,, \\
x_4 = & & - & \frac{3}{8}x_2 & + & \frac{1}{8}x_3 & & & + & \frac{15}{8}\,.
\end{array}$$

En este ejemplo,

$$T = \begin{pmatrix} 0 & \frac{1}{10} & -\frac{1}{5} & 0 \\ \frac{1}{11} & 0 & \frac{1}{11} & -\frac{3}{11} \\ -\frac{1}{5} & \frac{1}{10} & 0 & \frac{1}{10} \\ 0 & -\frac{3}{8} & \frac{1}{8} & 0 \end{pmatrix} \quad \text{y} \quad \mathbf{c} = \begin{pmatrix} \frac{3}{5} \\ \frac{25}{11} \\ -\frac{11}{10} \\ \frac{15}{8} \end{pmatrix}.$$

Como una aproximación inicial tomemos a $\mathbf{x}^{(0)} = (0,0,0,0)^t$ y generemos $\mathbf{x}^{(1)}$ mediante:

$$\begin{aligned}
x_1^{(1)} &= \tfrac{1}{10}x_2^{(0)} - \tfrac{1}{5}x_3^{(0)} + \tfrac{3}{5} = 0.6000\ ,\\
x_2^{(1)} &= \tfrac{1}{11}x_1^{(0)} + \tfrac{1}{11}x_3^{(0)} - \tfrac{3}{11}x_4^{(0)} + \tfrac{25}{11} = 2.2727\ ,\\
x_3^{(1)} &= -\tfrac{1}{5}x_1^{(0)} + \tfrac{1}{10}x_2^{(0)} + \tfrac{1}{10}x_4^{(0)} - \tfrac{11}{10} = -1.1000\ ,\\
x_4^{(1)} &= -\tfrac{3}{8}x_2^{(0)} + \tfrac{1}{8}x_3^{(0)} + \tfrac{15}{8} = 1.8750\ .
\end{aligned}$$

Las iteraciones adicionales $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)})^t$, se generan de manera similar y se presentan en la tabla siguiente.

Tabla 1

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$x_4^{(k)}$
0	0.0000	0.0000	0.0000	0.0000
1	0.6000	2.2727	−1.1000	1.8750
2	1.0473	1.7159	−0.80523	0.88524
3	0.93264	2.0533	−1.0493	1.1309
4	1.0152	1.9537	−0.96811	0.97385
5	0.98899	2.0114	−1.0103	1.0213
6	1.0032	1.9923	−0.99453	0.99444
7	0.99814	2.0023	−1.0020	1.0036
8	1.0006	1.9987	−0.99904	0.99889
9	0.99968	2.0004	−1.0004	1.0006
10	1.0001	1.9998	−0.99984	0.99980

La decisión de parar después de diez iteraciones está basada en el hecho de que

$$\frac{||\mathbf{x}^{(10)} - \mathbf{x}^{(9)}||_\infty}{||\mathbf{x}^{(10)}||_\infty} = \frac{8.0 \times 10^{-4}}{1.9998} < 10^{-3}\ .$$

En realidad, $||\mathbf{x}^{(10)} - \mathbf{x}||_\infty = 0.0002$.

El método del ejemplo anterior se llama **método iterativo de Jacobi**. Este consiste en resolver la $i-$ésima ecuación de $A\ \mathbf{x} = \mathbf{b}$ para x_i para obtener, siempre y cuando $a_{ii} \neq 0$, que

$$x_i = \sum_{\substack{j=1\\ j\neq i}}^{n} \Big(-\frac{a_{ij}\ x_j}{a_{ii}} \Big) + \frac{b_i}{a_{ii}} \qquad \text{para}\ \ i = 1,2,\ldots,n \tag{$XVII.2$}$$

y generar cada $x_i^{(k)}$ de las componentes de $\mathbf{x}^{(k-1)}$ para $k \geq 1$ con

$$x_i^{(k)} = \frac{1}{a_{ii}}[\sum_{\substack{j=1\\ j\neq i}}^{n}(-a_{ij}\ x_j^{(k-1)}) + b_i] \qquad \text{para}\ \ i = 1,2,\ldots,n\ . \tag{$XVII.3$}$$

El método puede escribirse en la forma $\mathbf{x}^{(k)} = T\ \mathbf{x}^{(k-1)} + \mathbf{c}$ dividiendo a A en su parte diagonal y no-diagonal. Para ver esto, sean D la matriz diagonal cuya diagonal es la

misma que la diagonal de A, $-L$ la parte triangular estrictamente inferior de A, y $-U$ la parte triangular estrictamente superior de A. Con esta notación, se separa en

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix} +$$
$$- \begin{pmatrix} 0 & 0 & \dots & 0 \\ -a_{21} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ -a_{n1} & \dots & -a_{n,n-1} & 0 \end{pmatrix} - \begin{pmatrix} 0 & -a_{12} & \dots & -a_{1n} \\ 0 & 0 & \dots & \dots \\ \dots & \dots & \dots & -a_{n-1,n} \\ 0 & 0 & \dots & 0 \end{pmatrix} =$$
$$= \; D \; - \; L \; - \; U \; .$$

La ecuación $A\,\mathbf{x} = \mathbf{b}$ ó $(D - L - U)\,\mathbf{x} = \mathbf{b}$ se transforma entonces en $D\,\mathbf{x} = (L+U)\,\mathbf{x} + \mathbf{b}$, y finalmente

$$\mathbf{x} = D^{-1}\,(L+U)\,\mathbf{x} + D^{-1}\,\mathbf{b}\;. \qquad (XVII.4)$$

Esto da lugar a la forma matricial de la técnica iterativa de Jacobi:

$$\mathbf{x}^{(k)} = D^{-1}\,(L+U)\,\mathbf{x}^{(k-1)} + D^{-1}\,\mathbf{b}\;, \qquad k = 1, 2, \dots . \qquad (XVII.5)$$

En la práctica, la ecuación $(XVII.3)$ es la que se usa para los cálculos, reservando a la ecuación $(XVII.5)$ para propósitos teóricos.

2. LOS ALGORITMOS DE JACOBI Y DE GAUSS-SEIDEL

Para resumir el método iterativo de Jacobi, presentamos el siguiente algoritmo:

Algoritmo iterativo de Jacobi.

==========

Para resolver el sistema lineal $A\,\mathbf{x} = \mathbf{b}$ con una aproximación inicial dada $\mathbf{x}^{(0)}$.

Entrada: número de incógnitas y de ecuaciones n; las componentes de la matriz $A = (a_{ij})$ donde $1 \le i, j \le n$; las componentes b_i, con $1 \le i \le n$, del término no homogéneo $\mathbf{b}$; las componentes XO_i, con $1 \le i \le n$, de la aproximación inicial $\mathbf{XO} = \mathbf{x}^{(0)}$; la tolerancia TOL; el número máximo de iteraciones N_0.

Salida: solución aproximada $x_1, x_2, \dots, x_n$ ó mensaje de que el número de iteraciones fue excedido.

Paso 1: Tomar $k = 1$.

Paso 2: Mientras que $k \le N_0$ seguir los pasos 3–6.

Paso 3: Para $i = 1, 2, \dots, n$ tomar

$$x_i = \frac{1}{a_{ii}}[-\sum_{\substack{j=1 \\ j \neq i}}^{n} (a_{ij}\; XO_j) + b_i]\;.$$

Paso 4: Si $||\mathbf{x} - \mathbf{XO}|| < TOL$ entonces SALIDA $(x_1, x_2, \dots, x_n)$;
(procedimiento completado satisfactoriamente) PARAR.

Paso 5: Tomar $k = k + 1$.

Paso 6: Para $i = 1, 2, \ldots, n$ tomar $XO_i = x_i$.

Paso 7: SALIDA *(número máximo de iteraciones excedido)*; *(procedimiento completado sin éxito)* PARAR.

==========

El paso 3 del algoritmo requiere que $a_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$. Si éste no es el caso, se puede realizar un reordenamiento de las ecuaciones para que ningún $a_{ii} = 0$, a menos que el sistema sea singular. Se sugiere que las ecuaciones sean arregladas de tal manera que a_{ii} sea lo más grande posible para acelerar la convergencia.

En el paso 4, el criterio de paro ha sido $||\mathbf{x} - \mathbf{XO}|| < TOL$; otro criterio de paro es iterar hasta que

$$\frac{||\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}||}{||\mathbf{x}^{(k)}||}$$

sea menor que alguna tolerancia predeterminada $\varepsilon > 0$. Para este propósito, se puede usar cualquier norma conveniente; la que más se usa es la norma l_∞.

Un análisis de la ecuación $(XVII.3)$ sugiere una posible mejora en el algoritmo iterativo de Jacobi. Para calcular $x_i^{(k)}$, se usan las componentes de $\mathbf{x}^{(k-1)}$. Como para $i > 1$, $x_1^{(k)}$, $x_2^{(k)}$, ..., $x_{i-1}^{(k)}$ ya han sido calculadas y supuestamente son mejores aproximaciones a la solución real x_1, x_2, ..., x_{i-1} que $x_1^{(k)}$, $x_2^{(k)}$, ..., $x_{i-1}^{(k)}$, parece razonable calcular $x_i^{(k)}$ usando los valores calculados más recientemente; es decir,

$$x_i^{(k)} = \frac{1}{a_{ii}}[-\sum_{j=1}^{i-1}(a_{ij}\ x_j^{(k)}) - \sum_{j=i+1}^{n}(a_{ij}\ x_j^{(k-1)}) + b_i]\ , \qquad (XVII.6)$$

para cada $i = 1, 2, \ldots, n$ en vez de la ecuación $(XVII.3)$.

Ejemplo. El sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dado por

$$\begin{array}{lrcrcrcrcr}
E_1: & 10\,x_1 & - & x_2 & + & 2\,x_3 & & & = & 6\ , \\
E_2: & -\,x_1 & + & 11\,x_2 & - & x_3 & + & 3\,x_4 & = & 25\ , \\
E_3: & 2\,x_1 & - & x_2 & + & 10\,x_3 & - & x_4 & = & -11\ , \\
E_4: & & & 3\,x_2 & - & x_3 & + & 8\,x_4 & = & 15\ ,
\end{array}$$

fue resuelto en el ejemplo anterior con el método iterativo de Jacobi. Incorporando la ecuación $(XVII.6)$ en el algoritmo iterativo de Jacobi, se obtienen las ecuaciones que se usarán para cada $k = 1, 2, \ldots$:

$$\begin{array}{lrcrcrcrcr}
x_1^{(k)} = & & & \frac{1}{10}x_2^{(k-1)} & - & \frac{1}{5}x_3^{(k-1)} & & & + & \frac{3}{5}\ , \\
x_2^{(k)} = & \frac{1}{11}x_1^{(k)} & & & + & \frac{1}{11}x_3^{(k-1)} & - & \frac{3}{11}x_4^{(k-1)} & + & \frac{25}{11}\ , \\
x_3^{(k)} = & -\frac{1}{5}x_1^{(k)} & + & \frac{1}{10}x_2^{(k)} & & & + & \frac{1}{10}x_4^{(k-1)} & - & \frac{11}{10}\ , \\
x_4^{(k)} = & & - & \frac{3}{8}x_2^{(k)} & + & \frac{1}{8}x_3^{(k)} & & & + & \frac{15}{8}\ .
\end{array}$$

Tomando $\mathbf{x}^{(0)} = (0, 0, 0, 0)^t$, generamos los vectores iterados de la tabla 2.

Tabla 2

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$x_4^{(k)}$
0	0.0000	0.0000	0.0000	0.0000
1	0.6000	2.3273	−0.98727	0.87885
2	1.0302	2.0369	−1.0145	0.98435
3	1.0066	2.0035	−1.0025	0.99838
4	1.0009	2.0003	−1.0003	0.99985
5	1.0001	2.0000	−1.0000	1.0000

Ya que

$$\frac{||\mathbf{x}^{(5)} - \mathbf{x}^{(4)}||_\infty}{||\mathbf{x}^{(4)}||_\infty} = \frac{0.0008}{2.000} = 4 \times 10^{-4} \ ,$$

se acepta $\mathbf{x}^{(5)}$ como una aproximación razonable a la solución. Es interesante notar que el método de Jacobi en el ejemplo dado requiere el doble de iteraciones para la misma precisión.

La técnica presentada en el último ejemplo se llama **método iterativo de Gauss-Seidel**. Para escribir este método en la forma matricial $(XVII.1)$ se multiplican ambos lados de la ecuación $(XVII.6)$ por a_{ii} y se recolectan todos los $k-$ésimos términos iterados para dar

$$a_{i1}\ x_1^{(k)} + a_{i2}\ x_3^{(k)} + \ldots + a_{ii}\ x_i^{(k)} = -a_{i,i+1}\ x_{i+1}^{(k-1)} - \ldots - a_{in}\ x_n^{(k-1)} + b_i \ ,$$

para cada $i = 1, 2, \ldots, n$. Escribiendo las n ecuaciones tenemos:

$$a_{11}\ x_1^{(k)} = -a_{12}\ x_2^{(k-1)} - a_{13}\ x_3^{(k-1)} - \ldots - a_{1n}\ x_n^{(k-1)} + b_1 \ ,$$

$$a_{21}\ x_1^{(k)} + a_{22}\ x_2^{(k)} = -a_{23}\ x_3^{(k-1)} \ldots - a_{2n}\ x_n^{(k-1)} + b_2 \ ,$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$

$$a_{n1}\ x_1^{(k)} + a_{n2}\ x_2^{(k)} + \ldots + a_{nn}\ x_n^{(k)} = b_n \ ,$$

y se sigue que, en forma matricial, el método de Gauss-Seidel puede ser representado como $(D - L)\ \mathbf{x}^{(k)} = U\ \mathbf{x}^{(k-1)} + \mathbf{b}$, ó

$$\mathbf{x}^{(k)} = (D - L)^{-1}\ U\ \mathbf{x}^{(k-1)} + (D - L)^{-1}\ \mathbf{b} \ . \qquad (XVII.7)$$

Para que la matriz triangular inferior $(D - L)$ sea no singular, es necesario y suficiente que $a_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$.

Para resumir el método iterativo de Gauss-Seidel, presentamos el siguiente algoritmo:

Algoritmo iterativo de Gauss-Seidel.

==========

Para resolver el sistema lineal $A\ \mathbf{x} = \mathbf{b}$ con una aproximación inicial dada $\mathbf{x}^{(0)}$.

Entrada: número de incógnitas y de ecuaciones n; las componentes de la matriz $A = (a_{ij})$ donde $1 \leq i, j \leq n$; las componentes b_i, con $1 \leq i \leq n$, del término no homogéneo $\mathbf{b}$; las

componentes XO_i, con $1 \leq i \leq n$, de la aproximación inicial $\mathbf{XO} = \mathbf{x}^{(0)}$; la tolerancia TOL; el número máximo de iteraciones N_0.

Salida: solución aproximada $x_1, x_2, \ldots, x_n$ ó mensaje de que el número de iteraciones fue excedido.

Paso 1: Tomar $k = 1$.

Paso 2: Mientras que $k \leq N_0$ seguir los pasos 3–6.

Paso 3: Para $i = 1, 2, \ldots, n$ tomar

$$x_i = \frac{1}{a_{ii}}[-\sum_{j=1}^{i-1}(a_{ij}\ x_j) - \sum_{j=i+1}^{n}(a_{ij}\ XO_j) + b_i]\ .$$

Paso 4: Si $||\mathbf{x} - \mathbf{XO}|| < TOL$ entonces SALIDA $(x_1, x_2, \ldots, x_n)$; *(procedimiento completado satisfactoriamente)* PARAR.

Paso 5: Tomar $k = k + 1$.

Paso 6: Para $i = 1, 2, \ldots, n$ tomar $XO_i = x_i$.

Paso 7: SALIDA *(número máximo de iteraciones excedido)*; *(procedimiento completado sin éxito)* PARAR.

==========

Los resultados de los ejemplos parecen implicar que el método de Gauss-Seidel es superior al método de Jacobi. Este es generalmente cierto, pero no siempre. En realidad, hay sistemas lineales para los cuales el método de Jacobi converge y el método de Gauss-Seidel no, y viceversa.

3. CONVERGENCIA DE LOS PROCESOS ITERATIVOS

Para estudiar la convergencia de las técnicas generales de iteración, consideramos la fórmula $(XVII.1)$

$$\mathbf{x}^{(k)} = T\ \mathbf{x}^{(k-1)} + \mathbf{c}$$

para cada $k = 1, 2, \ldots$, donde $\mathbf{x}^{(0)}$ es arbitrario. Este estudio requerirá del siguiente lema:

Lema XVII.1

Si el radio espectral $\rho(T)$ satisface que $\rho(T) < 1$, ó si la norma de la matriz T satisface que $||T|| < 1$, entonces $(I - T)^{-1}$ existe y

$$(I - T)^{-1} = I + T + T^2 + \ldots\ .$$

Teorema XVII.2

Para cualquier $\mathbf{x}^{(0)} \in \mathcal{R}^n$, la sucesión $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ definida por $(XVII.1)$

$$\mathbf{x}^{(k)} = T\ \mathbf{x}^{(k-1)} + \mathbf{c}$$

para cada $k \geq 1$ y $\mathbf{c} \neq \mathbf{0}$, converge a la solución única de $\mathbf{x} = T\ \mathbf{x} + \mathbf{c}$ si y sólo si $\rho(T) < 1$.

Demostración: de la ecuación $(XVII.1)$, se tiene que

$$\begin{aligned}\mathbf{x}^{(k)} &= T\ \mathbf{x}^{(k-1)} + \mathbf{c} = \\ &= T\ (T\ \mathbf{x}^{(k-2)} + \mathbf{c}) + \mathbf{c} = \\ &= T^2\ \mathbf{x}^{(k-2)} + (T + I)\ \mathbf{c} = \\ &\quad \ldots \\ &= T^k\ \mathbf{x}^{(0)} + (T^{k-1} + \ldots + T + I)\ \mathbf{c}\ .\end{aligned}$$

Suponiendo que $\rho(T) < 1$, podemos usar el Teorema XIII.15 y el Lema XVII.1 para obtener

$$\begin{aligned}\lim_{k\to\infty} \mathbf{x}^{(k)} &= \lim_{k\to\infty} T^k\ \mathbf{x}^{(0)} + \lim_{k\to\infty} \big(\sum_{j=0}^{k-1} T^j\big)\mathbf{c} \\ &= 0 \cdot \mathbf{x}^{(0)} + (I - T)^{-1}\mathbf{c} = (I - T)^{-1}\mathbf{c}\ .\end{aligned}$$

De $(XVII.1)$ $\mathbf{x} = \lim_{k\to\infty} \mathbf{x}^{(k)} = (I - T)^{-1}\mathbf{c}$ será la solución única de $\mathbf{x} = T\ \mathbf{x} + \mathbf{c}$.

Para probar el recíproco, sea $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ convergente a $\mathbf{x}$ para cualquier $\mathbf{x}^{(0)}$. De la ecuación $(XVII.1)$ sigue que $\mathbf{x} = T\ \mathbf{x} + \mathbf{c}$, así que para cada k,

$$\mathbf{x} - \mathbf{x}^{(k)} = T\ (\mathbf{x} - \mathbf{x}^{(k-1)}) = \ldots = T^k\ (\mathbf{x} - \mathbf{x}^{(0)})\ .$$

Por lo tanto, para cualquier vector $\mathbf{x}^{(0)}$,

$$\lim_{k\to\infty} T^k\ (\mathbf{x} - \mathbf{x}^{(0)}) = \lim_{k\to\infty} \mathbf{x} - \mathbf{x}^{(k)} = \mathbf{0}\ .$$

Consecuentemente, si $\mathbf{z}$ es un vector arbitrario y $\mathbf{x}^{(0)} = \mathbf{x} - \mathbf{z}$, entonces

$$\lim_{k\to\infty} T^k\ \mathbf{z} = \lim_{k\to\infty} T^k\ [\mathbf{x} - (\mathbf{x} - \mathbf{z})] = \mathbf{0}\ ,$$

lo cual, por el Teorema XIII.15, implica que $\rho(T) < 1$. **c.q.d.**

Un Teorema parecido nos dará condiciones de suficiencia para la convergencia de los procesos de iteración usando las normas en lugar del radio espectral.

Teorema XVII.3

Si $||T|| < 1$, para cualquier norma matricial natural, entonces la sucesión definida en la ecuación $(XVII.1)$, $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$, converge para cualquier $\mathbf{x}^{(0)} \in \mathcal{R}^n$, a un vector $\mathbf{x} \in \mathcal{R}^n$, y se satisfacen las siguientes cotas de error:

$$||\mathbf{x} - \mathbf{x}^{(k)}|| \leq ||T||^k\ ||\mathbf{x}^{(0)} - \mathbf{x}||\ , \qquad (XVII.8)$$

y

$$||\mathbf{x} - \mathbf{x}^{(k)}|| \leq \frac{||T||^k}{1 - ||T||}\ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}||\ . \qquad (XVII.9)$$

Demostración: comenzando con un vector arbitrario $\mathbf{x}^{(0)}$, formaremos una secuencia de aproximaciones

$$\mathbf{x}^{(1)} = T\ \mathbf{x}^{(0)} + \mathbf{c}\ ,$$

$$\mathbf{x}^{(2)} = T\ \mathbf{x}^{(1)} + \mathbf{c}\ ,$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots$$

$$\mathbf{x}^{(k)} = T\ \mathbf{x}^{(k-1)} + \mathbf{c}\ ,$$

de donde

$$\mathbf{x}^{(k)} = T^k\ \mathbf{x}^{(0)} + (T^{k-1} + \ldots + T + I)\ \mathbf{c}\ .$$

Como para $||T|| < 1$ tenemos $||T^k|| \to 0$ cuando $k \to \infty$, se deduce que

$$\lim_{k\to\infty} T^k = 0 \quad \text{y} \quad \lim_{k\to\infty} (I + T + T^2 + \ldots + T^{k-1}) = \sum_{k=0}^{\infty} T^k = (I - T)^{-1}\ .$$

Y por tanto, pasando al límite cuando $k \to \infty$, tenemos

$$\mathbf{x} = \lim_{k\to\infty} \mathbf{x}^{(k)} = (I - T)^{-1}\mathbf{c}\ .$$

Esto prueba la convergencia del proceso iterativo. Además, tenemos $(I - T)\ \mathbf{x} = \mathbf{c}$ ó $\mathbf{x} = T\ \mathbf{x} + \mathbf{c}$, lo cual quiere decir que el vector $\mathbf{x}$ en el límite es una solución del sistema. Como la matriz $(I - T)$ no es singular, la solución $\mathbf{x}$ es única. Hemos así demostrado la primera parte del Teorema.

Demostramos ahora la cota de error $(XVII.8)$. Supongamos que $\mathbf{x}^{(k+p)}$ y $\mathbf{x}^{(k)}$ son dos aproximaciones de la solución del sistema lineal $\mathbf{x} = T\ \mathbf{x} + \mathbf{c}$; de la ecuación $(XVII.1)$, tenemos:

$$\begin{aligned} ||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)}|| &= ||T\ \mathbf{x}^{(k+p-1)} - T\ \mathbf{x}^{(k-1)}|| = ||T\ (\mathbf{x}^{(k+p-1)} - \mathbf{x}^{(k-1)})|| = \ldots \\ &= ||T^k\ (\mathbf{x}^{(p)} - \mathbf{x}^{(0)})|| \leq ||T||^k\ ||\mathbf{x}^{(p)} - \mathbf{x}^{(0)}||\ . \end{aligned}$$

Ahora pasando al límite cuando $p \to \infty$, obtenemos

$$\lim_{p\to\infty} ||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)}|| \leq \lim_{p\to\infty} ||T||^k\ ||\mathbf{x}^{(p)} - \mathbf{x}^{(0)}|| = ||T||^k \lim_{p\to\infty} ||\mathbf{x}^{(p)} - \mathbf{x}^{(0)}||$$

y entonces

$$||\mathbf{x} - \mathbf{x}^{(k)}|| \leq ||T||^k\ ||\mathbf{x} - \mathbf{x}^{(0)}||\ ,$$

que es la cota de error $(XVII.8)$

Finalmente demostramos la cota de error $(XVII.9)$. Como antes, supongamos que $\mathbf{x}^{(k+p)}$ y $\mathbf{x}^{(k)}$ son dos aproximaciones de la solución del sistema lineal $\mathbf{x} = T\ \mathbf{x} + \mathbf{c}$. Tenemos

$$||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)}|| \leq ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| + ||\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}|| + \ldots + ||\mathbf{x}^{(k+p)} - \mathbf{x}^{(k+p-1)}||\ .$$

Por lo visto antes:

$$||\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)}|| \leq ||T||\ ||\mathbf{x}^{(m)} - \mathbf{x}^{(m-1)}|| \leq ||T||^{m-k}\ ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| \ ,$$

para $m > k \geq 1$. Entonces tenemos:

$$\begin{aligned} ||\mathbf{x}^{(p+k)} - \mathbf{x}^{(k)}|| &\leq ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| + ||T||\ ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| + \ldots + \\ &+ ||T||^{p-1}\ ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| \leq \frac{1}{1-||T||}\ ||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| \leq \\ &\leq \frac{||T||}{1-||T||}\ ||\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}|| \leq \ldots \leq \frac{||T||^k}{1-||T||}\ ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}|| \ , \end{aligned}$$

de donde se deduce la cota de error $(XVII.9)$. **c.q.d.**

Notése que si en particular elegimos $\mathbf{x}^{(0)} = \mathbf{c}$, entonces $\mathbf{x}^{(1)} = T\ \mathbf{c} + \mathbf{c}$ y

$$||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}|| = ||T\ \mathbf{c}|| \leq ||T||\ ||\mathbf{c}|| \ ,$$

y la cota $(XVII.9)$ nos da:

$$||\mathbf{x} - \mathbf{x}^{(k)}|| \leq \frac{||T||^{k+1}}{1-||T||}\ ||\mathbf{c}|| \ . \qquad (XVII.9')$$

Ejemplo. Demostrar que el proceso de iteración de Jacobi es convergente para el sistema lineal siguiente:

$$\begin{array}{lrcrcrcrcr} E_1: & 10\ x_1 & - & x_2 & + & 2\ x_3 & - & 3\ x_4 & = & 0\ , \\ E_2: & x_1 & + & 10\ x_2 & - & x_3 & + & 2\ x_4 & = & 5\ , \\ E_3: & 2\ x_1 & + & 3\ x_2 & + & 20\ x_3 & - & x_4 & = & -10\ , \\ E_4: & 3\ x_1 & + & 2\ x_2 & + & x_3 & + & 20\ x_4 & = & 15\ . \end{array}$$

¿Cuántas iteraciones han de efectuarse para hallar las raíces del sistema con un error menor de 10^{-4}?

Reduciendo el sistema a la forma especial para la iteración de Jacobi, tenemos

$$\begin{array}{lrcrcrcrcr} x_1 = & & & 0.1\ x_2 & - & 0.2\ x_3 & + & 0.3\ x_4\ , & & \\ x_2 = & -0.1\ x_1 & & & + & 0.1\ x_3 & - & 0.2\ x_4 & + & 0.5\ , \\ x_3 = & -0.1\ x_1 & - & 0.15\ x_2 & & & + & 0.05\ x_4 & - & 0.5\ , \\ x_4 = & -0.15\ x_1 & - & 0.1\ x_2 & - & 0.05\ x_3 & & & + & 0.75\ . \end{array}$$

Entonces la matriz del sistema es:

$$T = \begin{pmatrix} 0 & 0.1 & -0.2 & 0.3 \\ -0.1 & 0 & 0.1 & -0.2 \\ -0.1 & -0.15 & 0 & 0.05 \\ -0.15 & -0.1 & -0.05 & 0 \end{pmatrix} \ .$$

Utilizando, por ejemplo, la norma l_1, tenemos:

$$||T||_1 = \max\{0.35, 0.35, 0.35, 0.55\} = 0.55 < 1 \ .$$

En consecuencia el proceso de iteración para el sistema dado es convergente. Si consideramos como aproximación inicial de la raíz $\mathbf{x}$ el vector

$$\mathbf{x}^{(0)} = \mathbf{c} = (0.0, 0.5, -0.5, 0.75)^t \ ,$$

entonces

$$||\mathbf{c}||_1 = 0.0 + 0.5 + 0.5 + 0.75 = 1.75 \ .$$

Sea ahora k el número de iteraciones requeridas para conseguir la exactitud especificada. Utilizando la fórmula $(XVII.9')$, tenemos:

$$||\mathbf{x} - \mathbf{x}^{(k)}||_1 \leq \frac{||T||_1^{k+1}}{1 - ||T||_1} \ ||\mathbf{c}||_1 = \frac{0.55^{k+1} \times 1.75}{0.45} < 10^{-4} \ .$$

De aquí,

$$0.55^{k+1} < \frac{45}{175} \ 10^{-4}$$

o sea

$$(k+1) \log_{10} 0.55 < \log_{10} 45 - \log_{10} 175 - 4$$

$$-(k+1) \ 0.25964 < 1.65321 - 2.24304 - 4 = -4.58983$$

y consecuentemente

$$k + 1 > \frac{4.58983}{0.25964} \approx 17.7 \qquad \Longrightarrow \qquad k > 16.7 \ .$$

Podemos tomar $k = 17$. Notése que la estimación teórica del número de iteraciones necesarias para asegurar la exactitud especificada es excesivamente alto. A menudo se obtiene la exactitud deseada en un número menor de iteraciones.

Para aplicar los resultados de arriba a las técnicas iterativas de Jacobi o Gauss-Seidel, necesitamos escribir las matrices de iteración del método de Jacobi, T_J, dadas en $(XVII.5)$ y del método de Gauss-Seidel, T_{GS}, dadas en $(XVII.7)$, como

$$T_J = D^{-1} \ (L + U) \qquad \text{y} \qquad T_{GS} = (D - L)^{-1} \ U \ .$$

De ser $\rho(T_J)$ ó $\rho(T_{GS})$ menores que uno, es claro que la sucesión $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ converge a la solución $\mathbf{x}$ de $A \ \mathbf{x} = \mathbf{b}$. Por ejemplo, el esquema de Jacobi (ver ecuación $(XVII.5)$) tiene:

$$\mathbf{x}^{(k)} = D^{-1} \ (L + U) \ \mathbf{x}^{(k-1)} + D^{-1} \ \mathbf{b} \ ,$$

y si $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ converge a $\mathbf{x}$, entonces

$$\mathbf{x} = D^{-1} \ (L + U) \ \mathbf{x} + D^{-1} \ \mathbf{b} \ .$$

Esto implica que

$$D\ \mathbf{x} = (L+U)\ \mathbf{x} + \mathbf{b} \qquad \text{y} \qquad (D-L-U)\ \mathbf{x} = \mathbf{b}\ .$$

Ya que $D - L - U = A$, luego $\mathbf{x}$ satisface $A\ \mathbf{x} = \mathbf{b}$. De manera parecida se procede con el esquema de Gauss-Seidel dado por la ecuación $(XVII.7)$.

Podemos dar ahora condiciones de suficiencia fáciles de verificar para la convergencia de los métodos de Jacobi y de Gauss-Seidel.

Teorema XVII.4

Si A es una matriz estrictamente dominante diagonalmente, entonces, para cualquier elección de $\mathbf{x}^{(0)} \in \mathcal{R}^n$ ambos métodos, el de Jacobi o el de Gauss-Seidel, dan lugar a sucesiones $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ que convergen a la solución de $A\ \mathbf{x} = \mathbf{b}$.

La relación entre la rapidez de convergencia y el radio espectral de la matriz de iteración T se puede ver de la desigualdad $(XVII.8)$. Como $(XVII.8)$ se satisface para cualquier norma matricial natural se sigue, de la afirmación que siguió al Teorema XIII.14, que

$$||\mathbf{x}^{(k)} - \mathbf{x}|| \approx \rho(T)^k\ ||\mathbf{x}^{(0)} - \mathbf{x}||\ . \qquad (XVII.10)$$

Supongamos que $\rho(T) < 1$ y que se va a usar $\mathbf{x}^{(0)} = \mathbf{0}$ en una técnica iterativa para aproximar $\mathbf{x}$ con un error relativo máximo de 10^{-t}. Por la estimación $(XVII.10)$, el error relativo después de k iteraciones es aproximadamente $\rho(T)^k$, así que se espera una precisión de 10^{-t} si

$$\rho(T)^k \leq 10^{-t}\ ,$$

esto es, si

$$k \geq \frac{t}{-\log_{10}\rho(T)}\ .$$

Por lo tanto, es deseable escoger la técnica iterativa con el menor $\rho(T) < 1$ para el sistema particular $A\ \mathbf{x} = \mathbf{b}$.

En general no se conoce cuál de las dos técnicas, la de Jacobi o la de Gauss-Seidel, debe usarse. Sin embargo, en un caso especial, sí se conoce la respuesta.

Teorema XVII.5 (Stein-Rosenberg)

Si $a_{ij} \leq 0$ para cada $i \neq j$ y $a_{ii} > 0$ para cada $i = 1, 2, \ldots, n$, entonces se satisface una y solamente una de las siguientes afirmaciones:

a) $0 < \rho(T_{GS}) < \rho(T_J) < 1$;
b) $1 < \rho(T_J) < \rho(T_{GS})$;
c) $\rho(T_{GS}) = \rho(T_J) = 0$;
d) $\rho(T_J) = \rho(T_{GS}) = 1$;

Para el caso especial descrito en el Teorema XVII.5, vemos que cuando un método converge, entonces ambos convergen, siendo el método de Gauss-Seidel más rápido que el método de Jacobi.

4. LOS METODOS DE RELAJACION

Como la razón de convergencia de un procedimiento depende del radio espectral de la matriz asociada con el método, una manera de seleccionar un procedimiento que nos lleve a una convergencia acelerada consiste en escoger un método cuya matriz asociada tenga un radio espectral mínimo. Estos procedimientos nos llevan a los métodos de relajación. Pero antes de formular la teoría de los métodos de relajación, veamos las ideas fundamentales de la forma más simple. Supongamos que se dispone de un sistema de ecuaciones lineales

$$\begin{aligned} E_1: &\quad a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n = b_1\ , \\ E_2: &\quad a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n = b_2\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n = b_n\ . \end{aligned} \tag{$XVII$.11}$$

Transformaremos este sistema de la manera siguiente: pondremos los términos constantes a la izquierda y dividiremos la primera ecuación por $-a_{11}$, la segunda por $-a_{22}$, etc. Obtendremos entonces un sistema que está listo para la relajación:

$$\begin{aligned} E_1: &\quad -\ x_1 + b_{12}\ x_2 + \ldots + b_{1n}\ x_n + c_1 = 0\ , \\ E_2: &\quad b_{21}\ x_1 - \quad x_2 + \ldots + b_{2n}\ x_n + c_2 = 0\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ E_n: &\quad b_{n1}\ x_1 + b_{n2}\ x_2 + \ldots - \quad x_n + c_n = 0\ , \end{aligned} \tag{$XVII$.12}$$

donde

$$b_{ij} = -\frac{a_{ij}}{a_{ii}} \quad (i \neq j) \qquad \text{y} \qquad c_i = \frac{b_i}{a_{ii}}\ . \tag{$XVII$.13}$$

Supongamos que $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)})$ es la aproximación inicial a la solución del sistema dado. Sustituyendo estos valores en el sistema tendremos los **restos**

$$\begin{aligned} R_1^{(0)} &= c_1 - x_1^{(0)} + \sum_{j=2}^{n} b_{1j}\ x_j^{(0)} = x_1^{(1)} - x_1^{(0)}\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ R_k^{(0)} &= c_k - x_k^{(0)} + \sum_{\substack{j=1 \\ j\neq k}}^{n} b_{kj}\ x_j^{(0)} = x_k^{(1)} - x_k^{(0)}\ , \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ R_n^{(0)} &= c_n - x_n^{(0)} + \sum_{j=1}^{n-1} b_{nj}\ x_j^{(0)} = x_n^{(1)} - x_n^{(0)}\ . \end{aligned} \tag{$XVII$.14}$$

Si damos un incremento $\delta x_s^{(0)}$ a una de las incógnitas $x_s^{(0)}$, el resto correspondiente $R_s^{(0)}$ quederá disminuido en $\delta x_s^{(0)}$ y todos los otros restos $R_i^{(0)}$ $(i \neq s)$ quedarán aumentados en

$b_{is}\ \delta x_s^{(0)}$. De este modo, para hacer que desaparezca el resto siguiente $R_i^{(1)}$ es suficiente dar a $x_s^{(1)}$ un incremento $\delta x_s^{(1)} = R_s^{(0)}$ y tendremos

$$R_s^{(1)} = 0 \qquad \text{y} \qquad R_i^{(1)} = R_i^{(0)} + b_{is}\ \delta x_s^{(0)} \quad \text{para} \quad i \neq s\ . \tag{$XVII$.15}$$

Así el **método de relajación**, en su forma más simple, consiste en reducir el resto numéricamente más elevado a cero, en cada etapa, cambiando el valor del componente apropiado de la aproximación. El proceso acaba cuando todos los restos del último sistema transformado son iguales a cero con la exactitud requerida.

Vamos ahora a describir los métodos de relajación. Antes de describir un procedimiento para seleccionar tales métodos, necesitamos introducir una manera nueva de medir la cantidad por la cual una aproximación a la solución de un sistema lineal difiere de la solución real del sistema. El método hace uso del denominado **vector residual**.

Definición. Si $\tilde{\mathbf{x}} \in \mathcal{R}^n$ es una aproximación a la solución del sistema lineal definido por $A\ \mathbf{x} = \mathbf{b}$, el **vector residual** de $\tilde{\mathbf{x}}$ con respecto a este sistema se define como $\mathbf{r} = \mathbf{b} - A\ \tilde{\mathbf{x}}$.

En procedimientos como los métodos de Jacobi o de Gauss-Seidel se asocia un vector residual con cada cálculo de una componente aproximada del vector solución. El objetivo del método consiste en generar una sucesión de aproximaciones que hagan que los vectores residuales asociados converjan a cero. Supongamos que tomamos

$$\mathbf{r}_i^{(k)} = (r_{1i}^{(k)}, r_{2i}^{(k)}, \ldots, r_{ni}^{(k)})^t$$

para denotar al vector residual para el método de Gauss-Seidel correspondiente al vector solución aproximado

$$(x_1^{(k)}, x_2^{(k)}, \ldots, x_{i-1}^{(k)}, x_i^{(k-1)}, \ldots, x_n^{(k-1)})^t\ .$$

La $m-$ésima componente de $\mathbf{r}_i^{(k)}$ es

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj}\ x_j^{(k)} - \sum_{j=i}^{n} a_{mj}\ x_j^{(k-1)} \tag{$XVII$.16}$$

ó

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj}\ x_j^{(k)} - \sum_{j=i+1}^{n} a_{mj}\ x_j^{(k-1)} - a_{mi}\ x_i^{(k-1)}$$

para cada $m = 1, 2, \ldots, n$. En particular, la $i-$ésima componente de $\mathbf{r}_i^{(k)}$ es

$$r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij}\ x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\ x_j^{(k-1)} - a_{ii}\ x_i^{(k-1)}\ ;$$

así que

$$a_{ii}\ x_i^{(k-1)} + r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij}\ x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\ x_j^{(k-1)}\ . \tag{$XVII$.17}$$

Recuérdese, sin embargo, que en el método de Gauss-Seidel $x_i^{(k)}$ se escoge como

$$x_i^{(k)} = \frac{-\sum_{j=1}^{i-1}(a_{ij}\ x_j^{(k)}) - \sum_{j=i+1}^{n}(a_{ij}\ x_j^{(k-1)}) + b_i}{a_{ii}}\ , \qquad (XVII.6)$$

así que la ecuación $(XVII.17)$ puede escribirse como $a_{ii}\ x_i^{(k-1)} + r_{ii}^{(k)} = a_{ii}\ x_i^{(k)}$ ó

$$x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}}\ . \qquad (XVII.18)$$

Podemos derivar otra conexión entre los vectores residuales y la técnica de Gauss-Seidel. De $(XVII.16)$, la $i-$ésima componente de $\mathbf{r}_{i+1}^{(k)}$ es

$$\begin{aligned} r_{i,i+1}^{(k)} &= b_i - \sum_{j=1}^{i} a_{ij}\ x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\ x_j^{(k-1)} \\ &= b_i - \sum_{j=1}^{i-1} a_{ij}\ x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\ x_j^{(k-1)} - a_{ii}\ x_i^{(k)}\ . \end{aligned} \qquad (XVII.19)$$

La ecuación $(XVII.6)$ implica que $r_{i,i+1}^{(k)} = 0$. Entonces, en cierto sentido, la técnica de Gauss-Seidel está ideada para requerir que la $i-$ésima componente de $\mathbf{r}_{i+1}^{(k)}$ sea cero.

Reducir una coordenada del vector residual a cero, sin embargo, no es necesariamente la manera más eficiente de reducir la norma del vector $\mathbf{r}_{i+1}^{(k)}$. En realidad, modificando el procedimiento de Gauss-Seidel en la forma de la ecuación $(XVII.18)$ a:

$$x_i^{(k)} = x_i^{(k-1)} + \omega\ \frac{r_{ii}^{(k)}}{a_{ii}} \qquad (XVII.20)$$

para ciertas elecciones de ω positivo nos llevará a una convergencia significativamente más rápida.

Los métodos que emplean la ecuación $(XVII.20)$ se conocen como **métodos de relajación**. Para $0 < \omega < 1$, los procedimientos se llaman **métodos de sub-relajación** y se pueden emplear para obtener la convergencia de algunos sistemas que no son convergentes por el método de Gauss-Seidel. Para $\omega > 1$, los procedimientos se llaman **métodos de sobre-relajación** y se pueden usar para acelerar la convergencia de sistemas que son convergentes por el método de Gauss-Seidel. Estos métodos se abrevian frecuentemente como **SOR** (de **Successive Over-Relaxation**) y son particularmente útiles para resolver los sistemas lineales que aparecen en la solución numérica de ciertas ecuaciones diferenciales parciales.

Antes de ilustrar las ventajas del método SOR notamos que usando la ecuación $(XVII.17)$, la ecuación $(XVII.20)$ se puede reformular para propósitos de cómputo como

$$x_i^{(k)} = (1-\omega)\ x_i^{(k-1)} + \frac{\omega}{a_{ii}}\Big[b_i - \sum_{j=1}^{i-1} a_{ij}\ x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\ x_j^{(k-1)}\Big]\ . \qquad (XVII.21)$$

Para determinar la forma matricial del método SOR reescribimos $(XVII.21)$ como

$$a_{ii}\ x_i^{(k)} + \omega \sum_{j=1}^{i-1} a_{ij}\ x_j^{(k)} = (1-\omega)\ a_{ii}\ x_i^{(k-1)} - \omega \sum_{j=i+1}^{n} a_{ij}\ x_j^{(k-1)} + \omega\ b_i$$

así que

$$(D - \omega\ L)\ \mathbf{x}^{(k)} = [(1-\omega)\ D + \omega\ U]\ \mathbf{x}^{(k-1)} + \omega\ \mathbf{b}$$

ó

$$\mathbf{x}^{(k)} = (D - \omega\ L)^{-1}\ [(1-\omega)\ D + \omega\ U]\ \mathbf{x}^{(k-1)} + \omega\ (D - \omega\ L)^{-1}\ \mathbf{b}\ .$$

Algoritmo iterativo Successive Over-Relaxation (SOR).

==========

Para resolver el sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dados el parámetro ω y una aproximación inicial $\mathbf{x}^{(0)}$.

Entrada: número de incógnitas y de ecuaciones n; las componentes de la matriz $A = (a_{ij})$ donde $1 \le i, j \le n$; las componentes b_i, con $1 \le i \le n$, del término no homogéneo $\mathbf{b}$; las componentes XO_i, con $1 \le i \le n$, de la aproximación inicial $\mathbf{XO} = \mathbf{x}^{(0)}$; el parámetro ω; la tolerancia TOL; el número máximo de iteraciones N_0.

Salida: solución aproximada $x_1, x_2, \ldots, x_n$ ó mensaje de que el número de iteraciones fue excedido.

Paso 1: Tomar $k = 1$.

Paso 2: Mientras que $k \le N_0$ seguir los pasos 3–6.

Paso 3: Para $i = 1, 2, \ldots, n$ tomar

$$x_i = (1-\omega)\ XO_i + \frac{\omega}{a_{ii}}\big[-\sum_{j=1}^{i-1}(a_{ij}\ x_j) - \sum_{j=i+1}^{n}(a_{ij}\ XO_j) + b_i\big]\ .$$

Paso 4: Si $||\mathbf{x} - \mathbf{XO}|| < TOL$ entonces SALIDA $(x_1, x_2, \ldots, x_n)$; *(procedimiento completado satisfactoriamente)* PARAR.

Paso 5: Tomar $k = k + 1$.

Paso 6: Para $i = 1, 2, \ldots, n$ tomar $XO_i = x_i$.

Paso 7: SALIDA *(número máximo de iteraciones excedido)*; *(procedimiento completado sin éxito)* PARAR.

==========

Ejemplo. El sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dado por

$$\begin{array}{lrcrcrcr} E_1: & 4\ x_1 & + & 3\ x_2 & & & = & 24\ , \\ E_2: & 3\ x_1 & + & 4\ x_2 & - & x_3 & = & 30\ , \\ E_3: & & - & x_2 & + & 4\ x_3 & = & -24\ , \end{array}$$

tiene por solución $\mathbf{x} = (3, 4, -5)^t$. Se usarán los métodos de Gauss-Seidel y el SOR con $\omega = 1.25$ para resolver este sistema usando $\mathbf{x}^{(0)} = (1, 1, 1)^t$ para ambos métodos. Las ecuaciones para el método de Gauss-Seidel son

$$\begin{aligned} x_1^{(k)} &= -0.75\ x_2^{(k-1)} + 6\ , \\ x_2^{(k)} &= -0.75\ x_1^{(k)} + 0.25\ x_3^{(k-1)} + 7.5\ , \\ x_3^{(k)} &= 0.25\ x_2^{(k)} - 6\ , \end{aligned}$$

para cada $k = 1, 2, \ldots$, y las ecuaciones para el método SOR con $\omega = 1.25$ son

$$\begin{aligned} x_1^{(k)} &= -0.25\, x_1^{(k-1)} - 0.9375\, x_2^{(k-1)} + 7.5\ , \\ x_2^{(k)} &= -0.9375\, x_1^{(k)} - 0.25\, x_2^{(k-1)} + 0.3125\, x_3^{(k-1)} + 9.375\ , \\ x_3^{(k)} &= 0.3125\, x_2^{(k)} - 0.25\, x_3^{(k-1)} - 7.5\ . \end{aligned}$$

Las primeras siete iteraciones de cada método se muestran en las tablas 3 y 4.

Para obtener una precisión de siete lugares decimales el método de Gauss-Seidel requiere de 34 iteraciones en contra de las 14 que se necesitan en el método de sobrerelajación con $\omega = 1.25$.

Tabla 3

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	1.000000	1.000000	1.000000
1	5.250000	3.812500	−5.046875
2	3.1406250	3.8828125	−5.0292969
3	3.0878906	3.9267578	−5.0183105
4	3.0549317	3.9542236	−5.0114441
5	3.0343323	3.9713898	−5.0071526
6	3.0214577	3.9821186	−5.0044703
7	3.0134111	3.9888241	−5.0027940

Tabla 4

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	1.000000	1.000000	1.000000
1	6.312500	3.5195313	−6.6501465
2	2.6223144	3.9585266	−4.6004238
3	3.1333027	4.0102646	−5.0966864
4	2.9570513	4.0074838	−4.9734897
5	3.0037211	4.0029250	−5.0057135
6	2.9963275	4.0009263	−4.9982822
7	3.0000498	4.0002586	−5.0003486

Un problema que se presenta al usar el método SOR, es cómo escoger el valor apropiado de ω. Aún cuando no se conoce una respuesta completa a esta pregunta para un sistema lineal general $n \times n$, los siguientes resultados pueden usarse en ciertas situaciones.

Teorema XVII.6 (Kahan)

Si $a_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$, entonces $\rho(T_\omega) \geq |\omega - 1|$. Esto implica que $\rho(T_\omega) < 1$ sólo si $0 < \omega < 2$, donde $T_\omega = (D - \omega\, L)^{-1}\, [(1 - \omega)\, D + \omega\, U]$ es la matriz de iteración del método SOR.

Teorema XVII.7 (Ostrowski-Reich)

Si A es una matriz positiva definida y $0 < \omega < 2$, entonces el método SOR converge para cualquier elección de la aproximación inicial $\mathbf{x}^{(0)}$ del vector solución.

Teorema XVII.8

Si A es una matriz positiva definida y tridiagonal, entonces $\rho(T_{GS}) = [\rho(T_J)]^2 < 1$, la elección óptima de ω para el método SOR es

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(T_J)]^2}} \ , \qquad (XVII.22)$$

y con este valor de ω, $\rho(T_\omega) = \omega - 1$.

5. ELECCION DEL METODO PARA RESOLVER SISTEMAS LINEALES

Cuando el sistema lineal es lo suficientemente pequeño para que sea fácilmente acomodado en la memoria principal de un ordenador, es en general más eficaz usar una técnica directa que minimice el efecto del error de redondeo. Específicamente, es adecuado el algoritmo de eliminación Gaussiana con pivoteo escalado de columna.

Los sistemas lineales grandes cuyos coeficientes son entradas básicamente de ceros y que aparecen en patrones regulares se pueden resolver generalmente de una manera eficiente usando un procedimiento iterativo como el discutido en este capítulo. Los sistemas de este tipo aparecen naturalmente, por ejemplo, cuando se usan técnicas de diferencias finitas para resolver problemas de valor en la frontera, una aplicación común en la solución numérica de ecuaciones diferenciales parciales.

CAPITULO XVIII. ESTIMACIONES DE ERROR Y REFINAMIENTO ITERATIVO

1. ESTIMACIONES DE ERROR

Parece razonable intuitivamente que si $\tilde{\mathbf{x}}$ es una aproximación a la solución $\mathbf{x}$ de $A\ \mathbf{x} = \mathbf{b}$ y el vector residual $\mathbf{r} = \mathbf{b} - A\ \tilde{\mathbf{x}}$ tiene la propiedad de que $||\mathbf{r}||$ es pequeño, entonces $||\mathbf{x} - \tilde{\mathbf{x}}||$ será también peque˜no. Aún cuando éste es frecuentemente el caso, ciertos sistemas especiales, que aparecen bastante en la práctica, no tienen esta propiedad.

Ejemplo. El sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dado por

$$\begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 3.0001 \end{pmatrix} ,$$

tiene la solución única $\mathbf{x} = (1, 1)^t$. La aproximación a esta solución $\tilde{\mathbf{x}} = (3, 0)^t$ tiene vector residual

$$\mathbf{r} = \mathbf{b} - A\ \tilde{\mathbf{x}} = \begin{pmatrix} 3 \\ 3.0001 \end{pmatrix} - \begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -0.0002 \end{pmatrix} ,$$

así que $||\mathbf{r}||_\infty = 0.0002$.

Aunque la norma del vector residual es pequeña, la aproximación $\tilde{\mathbf{x}} = (3, 0)^t$ es obviamente bastante pobre; en realidad, $||\mathbf{x} - \tilde{\mathbf{x}}||_\infty = 2$.

Esta dificultad se puede explicar muy simplemente si se observa que la solución del sistema representa la intersección de las rectas

$$l_1 : \quad x_1 + 2\ x_2 = 3 \quad \text{y} \quad l_2 : \quad 1.0001\ x_1 + 2\ x_2 = 3.0001\ .$$

El punto $(3, 0)$ se encuentra en l_1 y las rectas son casi paralelas. Esto implica que $(3, 0)$ se encuentra también cerca de l_2, aún cuando difiere significativamente del punto de intersección $(1, 1)$. Si las rectas no hubieran sido casi paralelas, se esperería que un vector residual peque˜no implicara una aproximación precisa.

En general, no podemos depender de la geometría del sistema para obtener una indicación de cúando pueden presentarse problemas. Sin embargo, podemos extraer esta información considerando las normas de la matriz A y de su inversa.

Definición. El **número de condición** $K(A)$ de la matriz no singular A relativo a la norma $||\ \cdot\ ||$ se define como

$$K(A) = ||A||\ ||A^{-1}|| .$$

Teorema XVIII.1

Si $\tilde{\mathbf{x}}$ es una aproximación a la solución de $A\ \mathbf{x} = \mathbf{b}$ y A es una matriz no singular, entonces para cualquier norma natural,

$$||\mathbf{x} - \tilde{\mathbf{x}}|| \leq ||\mathbf{r}||\ ||A^{-1}|| = K(A)\frac{||\mathbf{r}||}{||A||} \qquad (XVIII.1)$$

y

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||}{||\mathbf{x}||} \leq ||A^{-1}||\ ||A||\ \frac{||\mathbf{r}||}{||\mathbf{b}||} = K(A)\ \frac{||\mathbf{r}||}{||\mathbf{b}||}\ , \qquad (XVIII.2)$$

siempre que $\mathbf{x} \neq 0$ y $\mathbf{b} \neq 0$, donde $\mathbf{r}$ es el vector residual de $\tilde{\mathbf{x}}$ con respecto al sistema $A\ \mathbf{x} = \mathbf{b}$.

Demostración: como $\mathbf{r} = \mathbf{b} - A\ \tilde{\mathbf{x}} = A\ \mathbf{x} - A\ \tilde{\mathbf{x}}$ y A no es singular:

$$||\mathbf{x}-\tilde{\mathbf{x}}|| = ||A^{-1}\ \mathbf{r}|| \leq ||A^{-1}||\ ||\mathbf{r}||\ .$$

Además, como $\mathbf{b} = A\ \mathbf{x}$, $||\mathbf{b}|| \leq ||A||\ ||\mathbf{x}||$; así que

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||}{||\mathbf{x}||} \leq ||A^{-1}||\ ||A||\ \frac{||\mathbf{r}||}{||\mathbf{b}||}\ .$$

c.q.d.

Las desigualdades $(XVIII.1)$ y $(XVIII.2)$ implican que las cantidades $||A^{-1}||$ y $K(A) = ||A||\ ||A^{-1}||$ pueden ser usadas para dar una indicación de la conexión entre el vector residual y la precisión de la aproximación. En general, el error relativo $||\mathbf{x} - \tilde{\mathbf{x}}||/||\mathbf{x}||$ es de mayor interés y por la desigualdad $(XVIII.2)$ este error está acotado por el producto del número de condición $K(A) = ||A||\ ||A^{-1}||$ con el residual relativo para esta aproximación $||\mathbf{r}||/||\mathbf{b}||$. Para esta aproximación puede usarse cualquier norma que sea conveniente, el único requisito es que se use consistentemente desde el principio hasta el final.

Ya que para cualquier matriz no singular A

$$1 = ||I|| = ||A \cdot A^{-1}|| \leq ||A||\ ||A^{-1}|| = K(A)\ ,$$

se espera que la matriz A tenga un buen comportamiento (llamada formalmente una **matriz bien condicionada**) si $K(A)$ está cerca de uno y un comportamiento defectuoso (llamada una **matriz mal condicionada**) cuando $K(A)$ sea significativamente mayor que uno. El comportamiento en esta situación se refiere a la relativa seguridad de que un vector residual peque˜no implique correspondientemente una solución aproximada precisa.

Ejemplo. La matriz del sistema considerado en el ejemplo anterior es

$$A = \begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix}\ ,$$

que tiene $||A||_\infty = 3.0001$. Esta norma no se considera grande, sin embargo

$$A^{-1} = \begin{pmatrix} -10000 & 10000 \\ 5000.5 & -5000 \end{pmatrix}\ ,$$

y $||A^{-1}||_\infty = 20000$ y para la norma infinita $K(A) = 20000 \times 3.0001 = 60002$. El tama˜no del número de condición para este ejemplo seguramente nos detendría al tomar decisiones apresuradas acerca de la precisión, basadas en el residual de la aproximación.

Mientras que, en teoría, el número de condición de una matriz depende totalmente de las normas de la matriz y de su inversa, en la práctica, el cálculo de la inversa está sujeto a errores de redondeo y es dependiente de la exactitud con la que se estén haciendo los cálculos. Si hacemos la suposición de que la solución aproximada al sistema lineal $A\ \mathbf{x} = \mathbf{b}$ se determina usando aritmética de t dígitos y eliminación Gaussiana, se puede demostrar que el vector residual $\mathbf{r}$ para la aproximación $\tilde{\mathbf{x}}$ tiene la propiedad

$$||\mathbf{r}|| \approx 10^{-t}\ ||A||\ ||\tilde{\mathbf{x}}|| \ . \qquad (XVIII.3)$$

De esta ecuación aproximada, se puede obtener una estimación del número de condición efectivo para la aritmética de t dígitos, sin la necesidad de invertir la matriz A. [La aproximación en la ecuación $(XVIII.3)$ supone que todas las operaciones aritméticas en la técnica de eliminación Gaussiana se efectúan usando aritmética de t dígitos, pero que las operaciones que se necesitan para determinar el residual se hacen en doble precisión, es decir, $2t$ dígitos, para eliminar la pérdida de precisión involucrada en la sustracción de números casi iguales que ocurre en los cálculos del residual].

La aproximación del número de condición $K(A)$ a t dígitos viene de considerar el sistema lineal $A\ \mathbf{y} = \mathbf{r}$. La solución de este sistema puede aproximarse fácilmente ya que los multiplicadores para el método de eliminación Gaussiana han sido ya calculados y supuestamente retenidos. De hecho $\tilde{\mathbf{y}}$, la solución aproximada de $A\ \mathbf{y} = \mathbf{r}$, satisface que

$$\tilde{\mathbf{y}} \approx A^{-1}\ \mathbf{r} = A^{-1}\ (\mathbf{b} - A\ \tilde{\mathbf{x}}) = A^{-1}\ \mathbf{b} - A^{-1}\ A\ \tilde{\mathbf{x}} = \mathbf{x} - \tilde{\mathbf{x}}\ ; \qquad (XVIII.4)$$

así que $\tilde{\mathbf{y}}$ es una estimación del error cometido al aproximar la solución del sistema original. Consecuentemente la ecuación $(XVIII.3)$ puede usarse para deducir que

$$\begin{aligned} ||\tilde{\mathbf{y}}|| &\approx ||\mathbf{x} - \tilde{\mathbf{x}}|| = ||A^{-1}\ \mathbf{r}|| \leq \\ &\leq ||A^{-1}||\ ||\mathbf{r}|| \approx ||A^{-1}||\ (10^{-t}\ ||A||\ ||\tilde{\mathbf{x}}||) = 10^{-t}\ ||\tilde{\mathbf{x}}||\ K(A)\ . \end{aligned}$$

Esto proporciona una aproximación para el número de condición involucrado en la solución del sistema $A\ \mathbf{x} = \mathbf{b}$ usando eliminación Gaussiana y el tipo de aritmética de t dígitos descrito anteriormente:

$$K(A) \approx 10^{t}\ \frac{||\tilde{\mathbf{y}}||}{||\tilde{\mathbf{x}}||}\ . \qquad (XVIII.5)$$

Ejemplo. El sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dado por

$$\begin{pmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.71 & 9.612 \\ 1.5611 & 5.1791 & 1.6852 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 15913 \\ 28.544 \\ 8.4254 \end{pmatrix}\ ,$$

tiene la solución exacta $\mathbf{x} = (1, 1, 1)^t$.

Usando eliminación Gaussiana y aritmética de redondeo de 5 dígitos llegamos a la matriz ampliada

$$\left(\begin{array}{ccc|c} 3.3330 & 15920 & -10.333 & 15913 \\ 0 & -10596 & 16.501 & -10580 \\ 0 & 0 & -5.079 & -4.7 \end{array}\right)\ .$$

La solución aproximada a este sistema es

$$\tilde{\mathbf{x}} = (1.2001, 0.99991, 0.92538)^t \ .$$

El vector residual correspondiente a $\tilde{\mathbf{x}}$ calculado con doble precisión (y luego redondeado a cinco dígitos) es

$$\begin{aligned}
\mathbf{r} &= \mathbf{b} - A\,\tilde{\mathbf{x}} = \\
&= \begin{pmatrix} 15913 \\ 28.544 \\ 8.4254 \end{pmatrix} - \begin{pmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.71 & 9.612 \\ 1.5611 & 5.1791 & 1.6852 \end{pmatrix} \begin{pmatrix} 1.2001 \\ 0.99991 \\ 0.92538 \end{pmatrix} = \\
&= \begin{pmatrix} -0.0051818 \\ 0.27413 \\ -0.18616 \end{pmatrix} \ ;
\end{aligned}$$

así que

$$||\mathbf{r}||_\infty = 0.27413 \ .$$

La estimación del número de condición dada en la discusión anterior se obtiene resolviendo primero el sistema $A\ \mathbf{y} = \mathbf{r}$:

$$\begin{pmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.71 & 9.612 \\ 1.5611 & 5.1791 & 1.6852 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -0.0051818 \\ 0.27413 \\ -0.18616 \end{pmatrix} \ ,$$

lo cual implica que $\tilde{\mathbf{y}} = (-0.20008, 8.9989 \times 10^{-5}, 0.074607)^t$. Usando la estimación dada por la ecuación $(XVIII.5)$:

$$K(A) \approx 10^5\ \frac{||\tilde{\mathbf{y}}||_\infty}{||\tilde{\mathbf{x}}||_\infty} = \frac{10^5\ (0.20008)}{1.2001} = 16672 \ .$$

Las cotas de error dadas en el Teorema XVIII.1 para estos valores son

$$||\mathbf{x} - \tilde{\mathbf{x}}||_\infty \leq K(A)\frac{||\mathbf{r}||_\infty}{||A||_\infty} = \frac{(16672)(0.27413)}{15934} = 0.28683$$

y

$$\frac{||\mathbf{x} - \tilde{\mathbf{x}}||_\infty}{||\mathbf{x}||_\infty} \leq K(A)\ \frac{||\mathbf{r}||_\infty}{||\mathbf{b}||_\infty} = \frac{(16672)(0.27413)}{15913} = 0.28721 \ .$$

Para determinar el número de condición exacto de A, necesitamos construir primero A^{-1}. Usando aritmética de redondeo de 5 dígitos para los cálculos se obtiene la aproximación:

$$A^{-1} = \begin{pmatrix} -1.1701 \times 10^{-4} & -1.4983 \times 10^{-1} & 8.5416 \times 10^{-1} \\ 6.2782 \times 10^{-5} & 1.2124 \times 10^{-4} & -3.0662 \times 10^{-4} \\ -8.6631 \times 10^{-5} & 1.3846 \times 10^{-1} & -1.9689 \times 10^{-1} \end{pmatrix} \ .$$

El Teorema XIII.13 puede usarse para demostrar que $||A^{-1}||_\infty = 1.0041$ y $||A||_\infty = 15934$. Como consecuencia la matriz A mal condicionada tiene

$$K(A) = (1.0041)\ (15934) = 15999\ .$$

La aproximación que habíamos obtenido antes está bastante cerca de este $K(A)$ y ha requerido un esfuerzo computacional considerablemente menor.

Como la solución real $\mathbf{x} = (1, 1, 1)^t$ de este sistema es conocida, podemos calcular ambos

$$||\mathbf{x} - \tilde{\mathbf{x}}||_\infty = 0.2001$$

y

$$\frac{||\mathbf{x} - \tilde{\mathbf{x}}||_\infty}{||\mathbf{x}||_\infty} = 0.2001\ .$$

Las cotas de error dadas en el Teorema XVIII.1 para estos valores son

$$||\mathbf{x} - \tilde{\mathbf{x}}||_\infty \leq K(A)\frac{||\mathbf{r}||_\infty}{||A||_\infty} = \frac{(15999)(0.27413)}{15934} = 0.27525$$

y

$$\frac{||\mathbf{x} - \tilde{\mathbf{x}}||_\infty}{||\mathbf{x}||_\infty} \leq K(A)\ \frac{||\mathbf{r}||_\infty}{||\mathbf{b}||_\infty} = \frac{(15999)(0.27413)}{15913} = 0.27561\ .$$

2. REFINAMIENTO ITERATIVO

En la ecuación $(XVIII.4)$ usamos la estimación $\tilde{\mathbf{y}} \approx \mathbf{x} - \tilde{\mathbf{x}}$, en la que $\tilde{\mathbf{y}}$ es la solución aproximada al sistema $A\ \mathbf{y} = \mathbf{r}$. Sería razonable sospechar, a partir de este resultado, que $\tilde{\mathbf{x}} + \tilde{\mathbf{y}}$ fuese una mejor aproximación a la solución del sistema lineal $A\ \mathbf{x} = \mathbf{b}$ que la aproximación inicial $\tilde{\mathbf{x}}$.

El método que usa esta suposición se llama **refinamiento iterativo**, o mejora iterativa y consiste en llevar a cabo iteraciones sobre el sistema cuyo lado derecho es el vector residual para las aproximaciones sucesivas, hasta que se obtiene una precisión satisfactoria. El procedimiento se usa generalmente sólo en los sistemas en que se sospecha que la matriz involucrada es mal condicionada, debido a que esta técnica no mejora mucho la aproximación para un sistema bien condicionado.

Algoritmo de refinamiento iterativo.

==========

Para aproximar la solución al sistema lineal $A\ \mathbf{x} = \mathbf{b}$ cuando se sospecha que A sea mal condicionada.

Entrada: número de incógnitas y de ecuaciones n; las componentes de la matriz $A = (a_{ij})$ donde $1 \leq i, j \leq n$; las componentes b_i, con $1 \leq i \leq n$, del término no homogéneo $\mathbf{b}$; la tolerancia TOL; el número máximo de iteraciones N_0.

Salida: solución aproximada $\mathbf{xx} = (xx_1, xx_2, \ldots, xx_n)$ ó mensaje de que el número de iteraciones fue excedido.

Paso 0: Resolver el sistema $A\ \mathbf{x} = \mathbf{b}$ para $x_1, x_2, \ldots, x_n$ por eliminación Gaussiana guardando los multiplicadores m_{ji}, $j = i+1, i+2, \ldots, n$, $i = 1, 2, \ldots, n-1$ y haciendo notar los intercambios de filas.

Paso 1: Tomar $k = 1$.

Paso 2: Mientras que $k \leq N_0$ seguir los pasos 3–8.

Paso 3: Para $i = 1, 2, \ldots, n$ (*calcular* $\mathbf{r}$*, realizando los cálculos con doble precisión aritmética*), tomar

$$r_i = b_i - \sum_{j=1}^{n} (a_{ij}\ x_j).$$

Paso 4: Resolver el sistema lineal $A\ \mathbf{y} = \mathbf{r}$ usando eliminación Gaussiana en el mismo orden que en el paso 0.

Paso 5: Para $i = 1, 2, \ldots, n$ tomar

$$xx_i = x_i + y_i.$$

Paso 6: Si $||\mathbf{x} - \mathbf{xx}|| < TOL$ entonces SALIDA $(xx_1, xx_2, \ldots, xx_n)$; *(procedimiento completado satisfactoriamente)* PARAR.

Paso 7: Tomar $k = k + 1$.

Paso 8: Para $i = 1, 2, \ldots, n$ tomar $x_i = xx_i$.

Paso 9: SALIDA *(número máximo de iteraciones excedido)*; *(procedimiento completado sin éxito)* PARAR.

==========

Si se está usando aritmética de t dígitos, un procedimiento recomendable para parar en el paso 6 consiste en iterar hasta que $|y_i^{(k)}| \leq 10^{-t}$ para cada $i = 1, 2, \ldots, n$.

Debe enfatizarse que la técnica de refinamiento iterativo no da resultados satisfactorios para todos los sistemas que contienen matrices mal condicionadas. En particular, si $K(A) \geq 10^t$, es probable que el procedimiento falle y que la única alternativa sea el uso de mayor precisión en los cálculos.

Ejemplo. En el ejemplo anterior encontramos que la aproximación al problema que habíamos estado considerando, usando aritmética de cinco dígitos y la eliminación Gaussiana, era $\tilde{\mathbf{x}}^{(1)} = (1.2001, 0.99991, 0.92538)^t$ y que la solución a $A\ \mathbf{y}^{(1)} = \mathbf{r}^{(1)}$ era $\tilde{\mathbf{y}}^{(1)} = (-0.20008, 8.9989 \times 10^{-5}, 0.074607)^t$. Usando el paso 5 del algoritmo, tenemos que

$$\tilde{\mathbf{x}}^{(2)} = \tilde{\mathbf{x}}^{(1)} + \tilde{\mathbf{y}}^{(1)} = (1.0000, 1.0000, 0.99999)^t$$

y el error real en esta aproximación es

$$||\mathbf{x} - \tilde{\mathbf{x}}^{(2)}||_\infty = 1.0 \times 10^{-5}\ .$$

Usando la técnica de paro sugerida para el algoritmo, calculamos $\mathbf{r}^{(2)} = \mathbf{b} - A\ \tilde{\mathbf{x}}^{(2)}$, y resolvemos el sistema $A\ \mathbf{y}^{(2)} = \mathbf{r}^{(2)}$, obteniéndose

$$\tilde{\mathbf{y}}^{(2)} = (-2.7003 \times 10^{-8}, 1.2973 \times 10^{-8}, 9.9817 \times 10^{-6})^t\ .$$

Puesto que $||\tilde{\mathbf{y}}^{(2)}||_\infty \leq 10^{-5}$, concluímos que

$$\tilde{\mathbf{x}}^{(3)} = \tilde{\mathbf{x}}^{(2)} + \tilde{\mathbf{y}}^{(2)} = (1.0000, 1.0000, 1.0000)^t$$

es suficientemente preciso. De hecho es claramente correcto.