UNA INTRODUCCIÓN A LOS MÉTODOS NUMÉRICOS PARA SISTEMAS LINEALES

F. VADILLO

Resumen. Los modelos matemáticos más sencillos en Matemática Aplicada son sistemas de ecuaciones lineales, son modelos que aparecen en muchas aplicaciones como se puede ver en el capítulo 2 de [16] y el capítulo 12 de [4]. En este documento se presentan los métodos clásicos: la eliminación gaussiana en sus distintas versiones, los métodos de eliminación compacta y finalmente se hace una breve presentación de los métodos iterativos.

Índice

Received by the editors 28 de septiembre de 2022.
www.ehu.es/~mepvaarf.
Departamento de Matemáticas de la UPV/EHU.

Eduardo Vadillo

1. Sistemas de ecuaciones lineales

Uno de los problemas más frecuentes en la computación científica es resolver sistemas de ecuaciones lineales simultáneas de m ecuaciones y n incógnitas de la forma

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1, \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2, \\ \vdots & & \vdots & & \dots & & \vdots & \vdots & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \dots & + & a_{mn}x_n & = & b_m, \end{array} \tag{1.1}$$

que abreviadamente se escribe en notación vectorial de la siguiente forma:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{1.2}$$

donde

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{n2} & \cdots & a_{mn} \end{pmatrix}, \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}, \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \tag{1.3}$$

con $\mathbf{A}$ y $\mathbf{b}$ datos conocidos y $\mathbf{x}$ las incógnitas que se tratan de calcular.

Los sistemas lineales se dividen en tres casos diferentes:

1. El caso de los sistemas cuadrados con el mismo número de ecuaciones que de incógnitas ($m = n$), sistema que se estudiaran en este capítulo.
2. El caso de sistemas sobre-determinados con más ecuaciones que incógnitas ($m > n$), sistemas que como se verá en el capítulo de ajustes de curvas calcula soluciones mínimo cuadráticas.
3. El caso $m < n$ con menos ecuaciones que incógnitas. En este caso evidentemente la solución no es única, suponiendo que $rang(A) = m$ la solución se puede descomponer en la forma
$$\mathbf{x} = \mathbf{x}^+ + \mathbf{x}^-,$$
donde $\mathbf{x}^+$ está en el rango de $\mathbf{A}^T$ por lo que existe $\mathbf{z}$ tal que $\mathbf{x}^+ = \mathbf{A}^T\mathbf{z}$ y $\mathbf{x}^-$ está en el núcleo de $\mathbf{A}$ por lo que $\mathbf{A}\mathbf{x}^- = \mathbf{0}$. Entonces
$$\mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{x}^+ + \mathbf{x}^-) = \mathbf{A}\mathbf{A}^T\mathbf{z} = \mathbf{b},$$
que un sistema cuadrado de $m \times m$ que se puede resolver para después calcular
$$\mathbf{x}^+ = \mathbf{A}^T[\mathbf{A}\mathbf{A}^T]^{-1}\mathbf{b}.$$

En este capítulo se resolverán los sistemas de ecuaciones lineales (1.1) o (1.2) con el mismo número de ecuaciones que de incógnitas ($m = n$).

Si la matriz $\mathbf{A}$ es invertible ($\det(\mathbf{A}) \neq 0$), la solución existe y es única: $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ y si no hiciera falta calcular dicha solución la discusión hubiera acabado. Pero en la mayoría de las aplicaciones es preciso resolver el sistema (1.1), de hecho resolver sistemas lineales es una de las tareas más frecuentes en la computación científica.

Desde el punto de vista teórico resolver el sistema no plantea ninguna dificultad, bastaría aplicar la regla de Cramer. Sin embargo, en la práctica existe una muy seria

Fernando Vadillo

dificultad que es su elevado número de operaciones. Para calcular un determinante de dimensión n se deben realizar

$$\begin{array}{cc} (n-1)\cdot n! & \text{multiplicaciones} \\ n!-1 & \text{adiciones.} \end{array}$$

Como para resolver el sistema por Cramer se deben calcular $n+1$ determinantes, en total el número de operaciones que resulta es

$$\begin{array}{cc} (n^2-1)\cdot n! & \text{multiplicaciones,} \\ (n+1)\cdot(n!-1) & \text{adiciones,} \\ n & \text{divisiones,} \end{array}$$

lo que supone que por ejemplo para resolver un sistema lineal de diez ecuaciones con diez incógnitas el número de operaciones es del orden de $4\cdot 10^8$ multiplicaciones, para $n=20$ resultan del orden de $9\cdot 10^{20}$; suponiendo que por ejemplo se realizan un millon de multiplicaciones por segundo, para resolver dicho sistema de 20×20 el tiempo necesario sería aproximadamente

$$9\cdot 10^{14}\text{segundos} \approx 10^{13}\text{minutos} \approx 2\cdot 10^{11}\text{horas} \approx 10^{10}\text{días} \approx 3\cdot 10^{7}\text{años}$$

cantidad astronómica para un sistema tan pequeño. Además, esta cantidad tan elevada de operaciones puede provocar soluciones incorrectas debido a los inevitables errores de redondeo. En consecuencia, es imprescindible conocer otros métodos para resolver sistemas lineales como (1.1), de hecho, el diseño de métodos numéricos que resuelvan eficazmente sistemas de ecuaciones lineales es todavía hoy un campo de investigación muy activo.

El método más importante para resolver sistema lineales es llamado de eliminación Gaussiana por haber sido Gauss(1777-1855) el primero en describirlo sistemáticamente aunque ya era conocía desde mucho antes.

1.1. Un ejemplo de eliminación Gaussiana. Para ilustra el algoritmo de la eliminación Gaussiana comenzamos por un ejemplo sencillo que es sistema de dimensión tres:

$$\begin{array}{rcrcrcl} 10x_1 & - & 7x_2 & & & = & 7, \\ -3x_1 & + & 2x_2 & + & 6x_3 & = & 4, \\ 5x_1 & - & x_2 & + & 5x_3 & = & 6, \end{array} \tag{1.4}$$

que escrito en forma matricial es:

$$\begin{pmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 4 \\ 6 \end{pmatrix}. \tag{1.5}$$

El coeficiente 10 de x_1 en la primera ecuación es el pivote que usaremos en la primera etapa del algoritmo para eliminar la incógnita x_1 de la segunda y tercera ecuación, para ello a la segunda ecuación le restamos la primera multiplicada por $-3/10$ y a la tercera por $5/10$ con el resultado siguiente:

$$\begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 2.5 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 6.1 \\ 2.5 \end{pmatrix}. \tag{1.6}$$

Fernando Iadrino

En la segunda etapa podríamos usar el pivote -0.1 para eliminar la incógnita x_2 de la tercera ecuación pero por razones de estabilidad es conveniente hacer un pivotaje, intercambiando la segunda y tercera ecuaciones el nuevo pivote es 2.5 resultando el sistema:

$$\begin{pmatrix} 10 & -7 & 0 \\ 0 & 2.5 & 6 \\ 0 & -0.1 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 2.5 \\ 6.1 \end{pmatrix}. \tag{1.7}$$

Para eliminar ahora x_2 de la tercera ecuación la restamos la segunda multiplicada por $-0.1/2.5$ resultando el sistema triangular superior:

$$\begin{pmatrix} 10 & -7 & 0 \\ 0 & 2.5 & 5 \\ 0 & 0 & 6.2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 2.5 \\ 6.2 \end{pmatrix}, \tag{1.8}$$

cuya solución se obtiene por una sencillo proceso de marcha atrás. De la última ecuación $6.2x_3 = 6.2$ obtenemos que $x_3 = 1$, llevado este valor a la segunda ecuación resulta $2.5x_2 + 5 = 2.5$ de donde sale $x_2 = -1$ y finalmente de la primera ecuación $10x_1 - 7(-1) = 7$ obtenemos el valor de $x_1 = 0$.

Este algoritmo de forma comparta se puede escribir de otra manera, definiendo las matrices:

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ -0.3 & -0.04 & 1 \end{pmatrix}, \mathbf{U} = \begin{pmatrix} 10 & -7 & 0 \\ 0 & 2.5 & 5 \\ 0 & 0 & 6.2 \end{pmatrix}, \mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

donde $\mathbf{L}$ los multiplicados usados durante la eliminación, $\mathbf{U}$ es la matriz de los coeficientes final, y $\mathbf{P}$ es una matriz de permutación que describe el pivotaje. Con estas tres matrices tenemos:

$$\mathbf{L} \cdot \mathbf{U} = \mathbf{P} \cdot \mathbf{A}, \tag{1.9}$$

que indica que la matriz de los coeficientes $\mathbf{A}$ se puede expresar en términos de productos de matrices con estructura más sencilla.

2. Condicionamiento del problema

Antes de presentar los métodos numéricos para resolver sistemas de ecuaciones lineales es necesario estudiar el condicionamiento del propio sistema para evitar que leves modificaciones en los coeficientes provoquen grandes diferencias en su solución. El siguiente ejemplo ilustra la situación. Si se considera el sistema

$$\begin{aligned} 2x + 6y &= 8, \\ 2x + 6.00001y &= 8.00001, \end{aligned}$$

de solución exacta $x = y = 1$ y el sistema muy parecido

$$\begin{aligned} 2x + 6y &= 8, \\ 2x + 5.99999y &= 8.00002, \end{aligned}$$

ahora la nueva solución es $x = 10, y = -2$ por lo que se concluye que el sistema está mal condicionado y será difícil su tratamiento numérico.

Para cuantificar el condicionamiento del sistema (1.2) se define el número de condición de la matriz $\mathbf{A}$ de la siguiente forma

Fernando Vadillo

Definicion 2.1. El número de condición $\kappa(A)$ de una matriz $A \in \mathbb{C}^{n\times n}$ en la norma $||\cdot||$ es el número

$$\kappa(A) = \begin{cases} ||A||\cdot||A^{-1}||, & \text{si A es inversible;} \\ \infty, & \text{otros.} \end{cases}$$

Nota 2.2. Note que para cualquier norma inducidas por una norma vectorial,

$$1 = ||I|| = ||AA^{-1}|| \leq ||A||\cdot||A^{-1}|| = \kappa(A).$$

Ejemplo 2.3. Si la matriz A es real, simétrica y definida positiva con autovalores

$$\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n = \lambda_{min} > 0,$$

entonces $||A||_2 = \lambda_{max}$ y $||A^{-1}||_2 = 1/\lambda_{min}$ por lo que $\kappa(A) = \lambda_{max}/\lambda_{min}$.

Proposicion 2.4. *Sea $Ax = b$ y $A(x + \triangle x) = b + \triangle b$, suponiendo que $b \neq 0$*

$$\frac{||\triangle x||}{||x||} \leq \kappa(A)\frac{||\triangle b||}{||b||}.$$

Lema 2.5. *Suponiendo que la matriz $A \in \mathbb{C}^{n\times n}$ satisface $||A|| < 1$ para alguna norma matricial inducida, entonces la matriz $I + A$ es inversible y además*

$$||(I + A)^{-1}|| \leq (1 - ||A||)^{-1}.$$

Proposicion 2.6. *Sea $Ax = b$ y $(A + \triangle A)(x + \triangle x) = b$. Suponiendo que A es inversible tal que $\kappa(A)||\triangle A||/||A|| < 1$ en alguna norma matricial inducida, entonces*

$$\frac{||\triangle x||}{||x||} \leq \frac{\kappa(A)}{1 - \kappa(A)\frac{||\triangle A||}{||A||}} \cdot \frac{||\triangle A||}{||A||}.$$

Teorema 2.7. *Teorema del condicionamiento. Suponiendo que la matriz A es inversible con $\kappa(A)||\triangle A||/||A|| < 1$ en alguna norma matricial inducida, $Ax = b$ y $(A + \triangle A)(\hat{x}) = b + \triangle b$, entonces*

$$\frac{||\hat{x} - x||}{||x||} \leq \frac{\kappa(A)}{1 - \kappa(A)\frac{||\triangle A||}{||A||}} \cdot \left(\frac{||\triangle A||}{||A||} + \frac{||\triangle b||}{||b||}\right).$$

Para un estudio más detallado del condicionamiento de matrices se recomiendan las referencias [20], [6],[10], [3], [2], [7], [19], [9], [17] y [12].

3. La eliminación Gaussiana sin pivotaje

3.1. El algoritmo general. La eliminación gaussiana en sus diferentes versiones es el método directo más utilizado para resolver sistemas de ecuaciones lineales. Como ya se comentó en el ejemplo, básicamente consiste en ir anulando por columnas los términos que se encuentran por debajo de la diagonal principal, sumando para ello a la fila correspondiente la que contiene el elemento diagonal (denominado pivote) multiplicado por el factor adecuado.

Sean $A = A^{(1)} = (a_{ij}^{(1)})$ la matriz de los coeficientes del sistema y sea $\mathbf{b}^{(1)} = (b_1^{(1)}, \cdots, b_n^{(1)})^T$ los segundos miembros. Suponiendo que $a_{11}^{(1)} \neq 0$, calculamos los multiplicadores:

$$m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \qquad i = 2, 3, \cdots, n \tag{3.1}$$

Fernando Vadillo

para obtener un sistema equivalente $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$ con:

$$a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i1}a_{11}^{(1)}, \quad i,j = 2,\cdots,n \tag{3.2}$$

$$b_{i}^{(2)} = b_{i}^{(1)} - m_{i1}b_{1}^{(1)}, \quad i = 2,\cdots,n \tag{3.3}$$

con la matriz:

$$A^{(2)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}. \tag{3.4}$$

Siguiendo con este procedimiento para las columnas segunda, tercera... en el paso k-ésimo el sistema que tenemos es:

$$A^{(k)}\mathbf{x} = \mathbf{b}^{(k)}, \qquad k = 1,\cdots,n \tag{3.5}$$

tales que:

$$A^{(k)} = \begin{pmatrix} a_{11}^{(1)} & \cdots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & \cdots & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}. \tag{3.6}$$

Suponiendo ahora que el pivote $a_{kk}^{(k)} \neq 0$ se calculan los multiplicadores:

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \quad i = k+1,\cdots,n, \tag{3.7}$$

para obtener la nueva matriz $A^{(k+1)}$ con los nuevos valores:

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)}, \tag{3.8}$$

$$b_{i}^{(k+1)} = b_{i}^{(k)} - m_{ik}b_{k}^{(k)}, \tag{3.9}$$

para $i,j = k+1,\cdots,n$.

Después de de $n-1$ etapas llegamos al sistema triangular superior:

$$\begin{pmatrix} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{nn}^{(n)} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1^{(1)} \\ \vdots \\ b_n^{(n)} \end{pmatrix}. \tag{3.10}$$

Llamando $U = A^{(n)}$ y $\mathbf{g} = \mathbf{b}^{(n)}$, el sistema $U\mathbf{x} = \mathbf{g}$ es triangular superior que resolvemos por un proceso de marcha atrás con:

$$x_n = \frac{g_n}{u_{nn}}, \tag{3.11}$$

$$x_k = \frac{1}{u_{kk}}\Big(g_k - \sum_{j=k+1}^{n} u_{kj}x_j\Big), \quad k = n-1, n-2,\cdots,1, \tag{3.12}$$

que finaliza el algoritmo de la eliminación Gaussiana.

Fernando Vadillo

En la práctica es frecuente que se tenga que resolver un sistema $\mathbf{Ax} = \mathbf{b}$ con la misma matriz $\mathbf{A}$ y distintos vectores $\mathbf{b}$. Por ejemplo, la columna j de la matriz inversa A^{-1} es la solución del sistema $A\mathbf{x} = \mathbf{e}_j$ y para calcular la inversa de una matriz dada tenemos que resolver los sistema para $j = 1, ..., n$. En estos caso es conveniente conservar los multiplicadores m_{ij}, introduciendo la matriz triangular inferior:

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & \cdots & 1 \end{pmatrix}, \tag{3.13}$$

con el siguiente resultado importante

Teorema 3.1. *Suponiendo que se puede lleva a cabo la eliminación Gaussiana y $U\mathbf{x} = \mathbf{c}$ es el sistema triangular superior resultante de la eliminación gaussiana,entonces $A = LU$, denomina factorización LU de la matriz A, y $\mathbf{b} = L\mathbf{c}$.*

Demostración. Definiendo la matriz

$$M^{(k)} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & -m_{k+1,k} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & -m_{k+2,k} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & -m_{n,k} & \cdots & 1 \end{pmatrix}, \tag{3.14}$$

resulta que

$$A^{(k+1)} = M^{(k)}A^{(k)}, \qquad \mathbf{b}^{(k+1)} = M^{(k+1)}\mathbf{b}^{(k)}$$

y por tanto

$$U = M^{(n-1)}M^{(n-2)} \cdots M^{(1)}A,$$

y

$$\mathbf{c} = M^{(n-1)}M^{(n-2)} \cdots M^{(1)}\mathbf{b}.$$

De esta manera

$$L = \left[M^{(n-1)}M^{(n-2)} \cdots M^{(1)}\right]^{-1} = \left[M^{(1)}\right]^{-1} \cdots \left[M^{(n-1)}\right]^{-1},$$

con

$$\left[M^{(k)}\right]^{-1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & m_{k+1,k} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & m_{k+2,k} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & m_{n,k} & \cdots & 1 \end{pmatrix}.$$

Fernando Vadillo

□

Después de calcular la factorización LU de la matriz de los coeficientes, el sistema $A\mathbf{x} = \mathbf{b}$ es equivalente a los dos sistemas triangulares siguientes:

$$L\mathbf{c} = \mathbf{b} \qquad \text{y} \qquad U\mathbf{x} = \mathbf{c}, \tag{3.15}$$

el primero triangular inferior y el segundo triangular superior. Además,

$$\det(A) = \det(L) \cdot \det(U) = u_{11} \cdots u_{nn}. \tag{3.16}$$

La tabla siguiente recoge el numero de sumas y productos necesarios:

	Factorización	segundo miembros	sistema triangular
Productos	$\frac{n(n^2-1)}{3}$	$\frac{n(n-1)}{2}$	$\frac{n(n+1)}{2}$
Sumas	$\frac{n(n-1)(2n-1)}{6}$	$\frac{n(n-1)}{2}$	$\frac{n(n-1)}{2}$

El coste de una algoritmo se evalúa en flops. Un flop (floating-point operations) consiste en una multiplicación y una adición en coma o punto flotante.

Para resolver un sistema de n ecuaciones lineales $A\mathbf{x} = \mathbf{b}$ el numero más importante de operaciones se utilizan en construir la factorización LU que precisa del orden de $\frac{n^3}{3}$ flops, porque para resolver un sistema triangular se realizan del orden de $\frac{n^2}{2}$ flops.

La matriz inversa de una matriz no singular ($\det(A) \neq 0$) se puede calcular resolviendo los n sistemas lineales

$$A\mathbf{x} = \mathbf{e}_i \qquad i = 1, \cdots, n, \tag{3.17}$$

donde los $\mathbf{e}_i$ son los vectores de la base canónica de $\mathbb{R}^n$. El numero de operaciones que necesitan es del orden de $\frac{n^3}{3}$ flops para factorizar la matriz A y $n \cdot n^2$ flops para resolver los sistemas triangulares, lo que hace un total de $\frac{4}{3}n^3$. En conclusión, calcular la inversa de una matriz cuesta 4 veces lo que supone resolver un sistema lineal y no n veces como podríamos pensar en un principio.

3.2. Factorización LU de una matriz.

Definicion 3.2. La j-ésima submatriz principal de una matriz $A \in \mathbb{C}^{n\times n}$ es la matriz $A_j \in \mathbb{C}^{j\times j}$ con $(A_j)_{kl} = a_{kl}$ para $1 \leq k, l \leq j$.

Teorema 3.3. *Teorema de existencia del factorización LU. Una matriz $A \in \mathbb{C}^{n\times n}$ tiene factorización LU si y sólo si todas sus submatrices principales A_j tienen matriz inversa. Además, dicha factorización es única y coincide con la obtenida por la eliminación Gaussiana.*

Demostración. Página 153 de [15]. □

Ejemplo 3.4. La siguiente matriz no es singular y tampoco tiene una factorización LU.

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

Teorema 3.5. *Factorización LU de matrices diagonal dominantes. Si una matriz $A \in \mathbb{C}^{n\times n}$ es regular y diagonal dominante, entonces siempre tiene una factorización LU.*

Fernando Vadillo

Definicion 3.6. Una matriz $\mathbf{P} \in \mathbb{R}^{n\times n}$ es una matriz de permutación si cada fila y cada columna contiene $n-1$ ceros y un uno.

Teorema 3.7. *Factorización LU con permutaciones. Si la matriz $A \in \mathbb{C}^{n\times n}$ es inversible, existe una matriz de permutación P tal que PA tiene factorización LU.*

Demostración. Por inducción sobre n. □

4. Métodos de eliminación compacta

Cuando la matriz A tiene todos sus menores principales distintos de cero, se puede plantear la posibilidad de hallar su factorización LU directamente sin pasar por todas la fases intermedias de la eliminación Gaussiana para lo que bastaría escribir la ecuación matricial

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ \cdots & \cdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{22} & \cdots & u_{2n} \\ & & \ddots & \cdots \\ & & & u_{nn} \end{pmatrix},$$

donde los segundo miembros son las incógnitas.

Multiplicando la primera fila de $\mathbf{L}$ de por las columnas de $\mathbf{U}$ se tiene la primera fila de $\mathbf{U}$ porque

$$u_{11} = a_{11}, \cdots\cdots, u_{1n} = a_{1n}.$$

En la siguiente etapa multiplicamos las filas de $\mathbf{L}$ por la primer columna de $\mathbf{U}$ se tiene que

$$l_{21}u_{11} = a_{21}, \cdots\cdots, l_{n1}u_{11} = a_{n1},$$

y como $u_{11} \neq 0$ se puede obtener la primera columna de $\mathbf{L}$.

A continuación multiplicamos la segunda fila de $\mathbf{L}$ por las columnas de $\mathbf{U}$ para obtener la segunda fila de $\mathbf{U}$.

En general, conocidas las $k-1$ filas de $\mathbf{U}$ y las $k-1$ columnas de $\mathbf{L}$ el algoritmo de Crout calcula la k-ésima fila de $\mathbf{U}$ de la forma

$$u_{kj} = a_{kj} - \sum_{p=1}^{k-1} l_{kp}u_{pj}, \quad j \geq k,$$

y la k-ésima columna de $\mathbf{L}$

$$l_{ik} = \frac{1}{u_{kk}}\left[a_{ik} - \sum_{p=1}^{k-1} l_{ip}u_{pk}\right], \quad i \geq k.$$

El método de Crout siguiendo el esquema adjunto, primero calcula la fila 1, después lo que queda de la columna 2, más tarde el tres de la fila 3, columna 4 ...

			1
			3
			5
			7
2	4	6	

Fernando Vaquilo

Otro método de eliminación comparta utilizado es el método de Doolittle (ver por ejemplo pg 167 de [3])

Es importante destacar que estos métodos de eliminación compacta tienen exactamente las mismas propiedades que la eliminación Gaussiana, sólo difieren en el orden de las operaciones, pero con dos ventajas importantes, la primera no necesitan resultados intermedios y después los productos escalares se pueden realizar directamente sin acumular errores.

5. La eliminación Gaussiana con pivotáje

Los elementos de la diagonal de la matriz U son los pivotes de la eliminación Gaussiana y si an algún momento el pivote es cero, evidentemente el algoritmo no se puede concluir. Incluso si en alguna etapa del algoritmo el pivote es pequeño comparado al resto de los elementos de la matriz, aunque el algoritmo se puede continuas el resultado final puede ser erróneo, veamos un ejemplo.

Se Considera el sistema (1.5) ligeramente modificado, el elemento (2,2) pasa de 2 a 2.099 y el segundo elemento de b_2 de 4 a 3.901, el nueva sistema es:

$$\begin{pmatrix} 10 & -7 & 0 \\ -3 & 2.099 & 6 \\ 5 & -1 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 3.901 \\ 6 \end{pmatrix}, \tag{5.1}$$

cuya solución exacta es la misma que antes $x_1 = 0, x_2 = -1, x_3 = 1$.

En la primera etapa con el pivote 10 el resultado es

$$\begin{pmatrix} 10 & -7 & 0 \\ -0 & -0.001 & 6 \\ 0 & 2.5 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 6.001 \\ 2.5 \end{pmatrix}, \tag{5.2}$$

donde el pivote es muy pequeño comparado a los otros elementos de la matriz. Si no se cambiara de pivote en el siguiente paso sumaríamos $2.5 \cdot 10^3$ veces la segunda fila a la tercera quedando

$$(5 + (2.5 \cdot 10^3)(6))x_3 = 2.5 + (2.5 \cdot 10^3)(6.001). \tag{5.3}$$

Si se supone que se opera con cinco dígitos, el resultado es $x_3 = 0.99993$ lo que no parece ser un error muy serio teniendo en cuenta que antes $x_3 = 1$. Sin embargo, cuando se calcula la segunda variable se tiene

$$-0.001x_2 + (6)(0.99993) = 6.001, \tag{5.4}$$

que da $x_2 = -1.5$ y finalmente $x_1 = -0.35$. Ahora los errores son enormes. El error ha surgido porque se ha tomado un pivote pequeño, si hubiéramos cambiado el pivote nada de esto hubiera ocurrido.

En la práctica conviene pivotar siempre que $a_{kk}^{(k)}$ sea pequeño, como se ha visto un pivote pequeño conduce a un multiplicador grande que amplifica los errores de redondeo.

Habitualmente se utilizan estrategias de pivotaje parcial en las que se lleva al pivote el elemento máximo en la columna de aquellos que toca aniquilar, es decir

$$pivote = \max_{k \le i \le n} |a_{ik}^{(k)}|. \tag{5.5}$$

La interpretación matricial del algoritmo de Gauss con pivotaje parcial intercambia filas que equivale a pre-multiplicar la matriz A por una matriz de permutación

Fernando Vadillo

P, matrices que tiene exactamente un 1 en cada fila y cada columna y son ortogonales por lo que $P^{-1} = P^T$. En consecuencia la eliminación gaussiana conpivotaje parcial obtiene una factorización de la forma

$$L \cdot U = P \cdot A. \tag{5.6}$$

6. La factorización de Cholesky

Además de la factorización LU, las matrices definidas positivas admiten otra otro tipo de factorización más barata que se estudia a continuación.

Teorema 6.1. *Teorema de la factorización de Cholesky . Si $A \in \mathbb{C}^{n\times n}$ es una matriz definida positiva, existe una matriz triangular superior $R \in \mathbb{C}^{n\times n}$ con elementos diagonales positivos tal que $A = R^*R$.*

Demostración. Por inducción sobre n. □

Teorema 6.2. *Matrices reales Si $A \in \mathbb{R}^{n\times n}$ no singular, entonces existe una matriz G triangular inferior con los $g_{ii} > 0, i = 1, \cdots, n$ tal que $A = GG^T$ si y sólo si A es definida positiva.*

Demostración. Notas de clase. □

Después de probar la existencia de la factorización de Cholesky ahora se explica como se calcula. Escribiendo la igualdad

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

multiplicando las filas de G por las columnas de G^T se tiene

$$\begin{aligned} g_{11} &= \sqrt{a_{11}}, \\ g_{i1} &= \frac{a_{i1}}{g_{11}}, \qquad i = 2, \cdots n, \end{aligned}$$

y después para $k = 2, \cdots, n$

$$\begin{aligned} g_{kk} &= \sqrt{a_{kk} - \sum_{p=1}^{k-1} g_{kp}^2}, \\ g_{ik} &= \frac{a_{ik} - \sum_{p=1}^{k-1} g_{ip} g_{kp}}{g_{kk}}, \qquad i = k+1, \cdots n. \end{aligned}$$

En cuanto al número de operaciones se necesitan:

- n raíces cuadradas, una por cada término diagonal.
- $\frac{n(n-1)}{2}$ divisiones, una por cada término no diagonal.

Fernando Vadillo

- Multiplicaciones

$$\begin{aligned}\sum_{k=2}^{n}(k-1)(n-k+1) &= n\sum_{k=2}^{n}(k-1)-\sum_{k=2}^{n}(k-1)^2,\\ &= n\frac{(n-1)n}{2}-\frac{(n-1)n(2n-1)}{6}\approx\frac{n^3}{6}.\end{aligned}$$

que es la mitad de las $\frac{n^3}{3}$ necesarias en la eliminación Gaussiana.

7. Métodos iterativos

Un método iterativo para resolver un sistema lineal $\mathbf{Ax}=\mathbf{b}$ se puede describir como aquel en el que partiendo de un valor aproximado $\mathbf{x}^{(0)}$, se genera una sucesión $\mathbf{x}^{(1)},\mathbf{x}^{(2)},...\mathbf{x}^{(\mathbf{k})}...$ que converge a la solución exacta del problema. Dicha convergencia se demostrará en ausencia de errores de redondeo, porque en la práctica bastará con elegir k suficientemente grande para que los errores de aproximación sean del mismo orden que los errores de redondeo.

El principal motivo para usar buscar nuevos métodos para resolver sistemas lineales, es que para dimensión muy grande $n\gg 1$ los métodos directos , que necesitan del orden de n^3 operaciones son muy caros. Además, cuando la matriz $\mathbf{A}$ tiene alguna estructura especial, por ejemplo matrices tridiagonales o parecidas, los métodos iterativos pueden tener una importante ventaja respecto de los métodos directos conservando muchos de sus ceros.

En esta sección se estudian métodos iterativos lineales porque se calcula la iteración $\mathbf{x}^{(k)}$ con operaciones lineales de la aproximación previa $\mathbf{x}^{(k-1)}$, en caso contrario el método iterativo no lineal, es decir, métodos de optimación, Krylov y multigrid. Para los alumnos interesados se recomienda el capítulo 7 de [1].

7.1. Matrices convergentes.

Para estudiar las propiedades de convergencia de los métodos iterativos es necesario conocer las matrices convergentes.

Definicion 7.1. Matriz convergente Una matriz $\mathbf{A}$ se dice que es convergente si $\lim_{k\to\infty}\mathbf{A}^k=\mathbf{0}$, donde $\mathbf{0}$ indica la matriz cero.

Teorema 7.2. *Una matriz $\mathbf{A}$ es convergente si y sólo si $\rho(\mathbf{A})<1$.*

Corolario 7.3. *Si para alguna norma matricial $||\mathbf{A}||<1$, entonces la matriz $\mathbf{A}$ es convergente.*

Teorema 7.4. *Teorema de Gerschgorin Dada una matriz $A\in\mathbb{C}^{n\times n}$, se define para cada $r_i=\sum_{j=1,j\neq i}^{n}|a_{ij}|$ y las bolas cerradas $\mathcal{B}_i(a_{ii},r_i)$ para $i=1,...,n$. Entonces*

(a) *Todos los autovalores de la matriz A se encuentra en la unión de la bolas $\mathcal{R}=\bigcup_{i\leq i\leq n}\mathcal{B}_i$.*
(b) *Cada componente conexa de $\mathcal{R}$ contiene tanto autovalores como bolas, contando cada autovalor tantas veces como su multiplicidad indica.*

Este teorema permite localizar los valores propios de una matriz.

Fernando Vadillo

Ejemplo 7.5. Considere la matriz simetrica

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

los discos de Gerschgorin son $\mathcal{B}(3,1/2), \mathcal{B}(1,1), \mathcal{B}(1,1/2)$ por lo se puede afirmar que hay dos autovalores en $[0,2]$ y el otra está en $[2.5,3.5]$.

Conviene notar que como los autovalores de la matriz A y $D^{-1}AD$ son los mismos, este teorema se puede aprovechar para localizar mejor alguno de sus autovalores. En el ejemplo anterior si se toma la matriz diagonal $D = diag(\alpha,1,1)$ con $\alpha > 0$

$$D^{-1}AD = \begin{pmatrix} 3 & -\frac{1}{2\alpha} & 0 \\ -\frac{\alpha}{2} & 1 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & 1 \end{pmatrix},$$

ahora las bolas son $\mathcal{B}(3,1/(2\alpha)), \mathcal{B}(1,(\alpha+1)/2)$ y entonces se puede deducir que $\rho(A) < 3.191$.

7.2. Descripción general de los métodos iterativos lineales. La idea básica de los métodos iterativos lineales escribe la matriz $\mathbf{A = M + N}$ tal que la matriz M es fácil de invertir. Entonces se define la iteración

$$\mathbf{M x^{(k)} = b - N x^{(k-1)}}, \tag{7.1}$$

y si después se comprueba que existe $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}$, evidentemente $\mathbf{x}$ es la solución.

Además de elegir las matrices $\mathbf{M}$ y $\mathbf{N}$ para que el método sea convergente, se debe determinar algún criterio de parada de la iteración para lo que se define el error: $\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$. Lo razonable ahora sería pedir que $||\mathbf{e}^{(k)}|| \leq \varepsilon$, para alguna tolerancia ε pequeña. Sin embargo, en la práctica como no se conoce el valor exacta $\mathbf{x}$, es imposible evaluar dicho error y en su lugar se define el residuo $\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A x}^{(k)} = \mathbf{A e}^{(k)}$ y se utiliza un criterio de parada del tipo $||\mathbf{r}^{(k)}|| \leq \varepsilon_r$ porque se acota el error de la forma

$$||\mathbf{e}^{(k)}|| = ||\mathbf{A}^{-1}\mathbf{r}^{(k)}|| \leq ||\mathbf{A}^{-1}|| \cdot ||\mathbf{r}^{(k)}|| \leq ||\mathbf{A}^{-1}|| \cdot \varepsilon_r.$$

El residuo también se utiliza para expresar el método iterativo (7.1) de forma más cómoda porque

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \mathbf{M}^{-1}\mathbf{r}^{(k-1)}. \tag{7.2}$$

Para analizar el error del método (7.1), como $M\mathbf{x} = \mathbf{b} - N\mathbf{x}$, es evidente que

$$\mathbf{e}^{(k)} = -\mathbf{M}^{-1}\mathbf{N} \cdot \mathbf{e}^{(k-1)} = \mathbf{R} \cdot \mathbf{e}^{(k-1)}, \tag{7.3}$$

donde $\mathbf{R = I - M^{-1}A}$ se denomina matriz de convergencia del método iterativo se demuestra el siguiente teorema de convergencia

Teorema 7.6. *Teorema de convergencia El método iterativo* (7.1) *es convergente si y sólo su matriz de convergencia es una matriz convergente.*

El radio espectral de la matriz $\mathbf{R}$ determina la velocidad de convergencia del método iterativo. Es evidente de (7.3) que

$$\mathbf{e}^{(k)} = \mathbf{R}^k \cdot \mathbf{e}^{(0)}.$$

Fernando Vadillo

Si se toma $\mathbf{e}^{(0)}$ un autovector asociado al autovalor dominante de la matriz $\mathbf{R}$, es decir

$$\mathbf{R}\mathbf{e}^{(0)} = \lambda\mathbf{e}^{(0)}, \quad \text{con} \quad |\lambda| = \rho(\mathbf{R}),$$

entonces $\mathbf{R}^k\mathbf{e}^{(0)} = \lambda^k\mathbf{e}^{(0)}$ y

$$||\mathbf{e}^{(k)}|| = \rho(\mathbf{R})^k \cdot ||\mathbf{e}^{(0)}||.$$

Esto significa que si se quisiera reducir la norma del error en 10^{-m}, bastaría con que $k\log_{10}\rho(R) \leq -m$, es decir, si el método es convergente $\log_{10}\rho(R) < 0$ y debemos realizar un número de iteraciones $k \geq m/(-\log_{10}\rho(R))$, por tanto, cuando $\rho(R) \ll 1$, $-\log_{10}\rho(R) \gg 1$, entonces k será pequeño y se precisarán pocas iteraciones y al reves cuando $\rho(R) \approx 1$. En conclusión, cuanta menor sea $\rho(R)$ más rápida será la convergencia.

Para construir los métodos que estudiaremos a continuación, dada la $\mathbf{A} = (a_{ij})$, se definen la matriz diagonal

$$\mathbf{D} = \begin{pmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{pmatrix},$$

y las dos matrices estrictamente triangulares

$$L = \begin{pmatrix} & & & \\ a_{21} & & & \\ a_{31} & a_{32} & & \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} \end{pmatrix}, U = \begin{pmatrix} a_{12} & a_{13} & \cdots & a_{1n} \\ & a_{23} & \cdots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{n-1n} \\ & & & \end{pmatrix},$$

tales que $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$.

7.3. El método de Jacobi. En el método de Jacobi: $\mathbf{M} = \mathbf{D}, \mathbf{N} = \mathbf{L} + \mathbf{U}$ por lo que el método iterativo es

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \mathbf{D}^{-1}\mathbf{r}^{(k-1)}. \tag{7.4}$$

y como $\mathbf{D}$ es diagonal su inversa es inmediata para lo que es necesario que los términos diagonales $a_{ii} \neq 0$. Por componentes se expresa de la forma

$$x_i^{(k)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k-1)}\right), \qquad i = 1, ..., n.$$

Para estudiar la convergencia de este método se aplica el teorema 7.6 por lo que dependerá del radio de convergencia de la matriz $\mathbf{J} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$.

Teorema 7.7. (a) *El método de Jacobi es convergente para toda matriz* $\mathbf{A}$ *estrictamente diagonal dominante por filas, es decir, tal que*

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}|, \qquad i = 1, \cdots n.$$

(b) *El método de Jacobi es convergente para toda matriz* $\mathbf{A}$ *estrictamente diagonal dominante por columnas, es decir, tal que*

$$|a_{jj}| > \sum_{i\neq j} |a_{ji}|, \qquad j = 1, \cdots n.$$

Fernando Vadillo

Definicion 7.8. Matrices irreducibles Una matriz $\mathbf{A} \in \mathbb{C}^{n\times n}$ se dice irreducible, si no existe una matriz de permutación $\mathbf{P}$ tal que

$$\mathbf{P}^T\mathbf{A}\mathbf{P} = \begin{pmatrix} \overline{A}_{11} & \overline{A}_{12} \\ 0 & \overline{A}_{22} \end{pmatrix},$$

donde $\overline{A}_{11} \in \mathbb{C}^{p\times p}$ y $\overline{A}_{22} \in \mathbb{C}^{q\times q}$ tal que $p+q=n$ con $p,q>0$, $\overline{A_{12}} \in \mathbb{C}^{p\times q}$ y $0 \in \mathbb{C}^{q\times p}$ es la matrices de ceros.

Existe una definición alternativa de matriz irreducible que es más fácil de comprobar. Dada la matriz $\mathbf{A}$, se le asocia una grafo $G(A)$ con los vertices $1,...n$ y los caminos $i \to j$ si $a_{ij} \neq 0$. Entonces se puede demostrar que una matriz es irreducible si y sólo si el grafo $G(A)$ está conectado, es decir, siempre se puede encontrar una camino entre dos vértices cualesquiera.

Teorema 7.9. (a) *El método de Jacobi es convergente para toda matriz* $\mathbf{A}$ *irreducible y diagonal dominante por filas, es decir, tal que*

$$|a_{ii}| \geq \sum_{j\neq i} |a_{ij}|, \qquad i = 1, \cdots, n$$

con una desigualdad estricta para al menos un índice.

(b) *El método de Jacobi es convergente para toda matriz* $\mathbf{A}$ *irreducible y diagonal dominante por columnas, es decir, tal que*

$$|a_{jj}| \geq \sum_{i\neq j} |a_{ji}|, \qquad j = 1, \cdots, n$$

con una desigualdad estricta para al menos un índice.

7.4. El método de Gauss-Seidel. En el método de Gauss-Seidel se toma la matriz triangular inferior $\mathbf{M} = \mathbf{E} = \mathbf{L} + \mathbf{D}$ por lo que resulta la iteración

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \mathbf{E}^{-1}\mathbf{r}^{(k-1)}, \tag{7.5}$$

que expresada por componentes queda

$$x_i^{(k)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}x_j^{(k)} - \sum_{j>i} a_{ij}x_j^{(k-1)}\right), \qquad i = 1, ..., n.$$

Y la convergencia ahora depende de la matriz $\mathbf{G} = -\mathbf{E}^{-1}\mathbf{U}$.

Teorema 7.10. (a) *El método de Gauss-Seidel es convergente para toda matriz estrictamente diagonal dominante por filas o columnas.*

(b) *El método de Gauss-Seidel es convergente para toda matriz irreducible y diagonal dominante por filas o columnas.*

Fernando Vadillo

7.5. Métodos de relajación. Para cada uno de los dos métodos anteriores se puede un sencilla modificación para obtener un ω-Jacobi y un ω-Gauss-Seidel simplemente haciendo que al final de cada iteración

$$\mathbf{x}^{(k)} = \omega \mathbf{x}^{(k)} + (1-\omega)\mathbf{x}^{(k-1)}, \tag{7.6}$$

con el parámetro $\omega > 0$ y con la idea de acelerar la convergencia. Cuando $0 < \omega < 1$ se tiene una sub-relajación y en caso contrario una sobre-relajación.

En el caso de Gauss-Seidel si se elige el parámetro $1 < \omega < 2$ se tiene lo conocidos como métodos de sobre relajación SOR (successive over-relaxation) para los cuales la matriz $\mathbf{M} = \frac{1-\omega}{\omega}\mathbf{D} + \mathbf{E}$ de expresión matricial

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \omega \left((1-\omega)\mathbf{D} + \omega \mathbf{E}\right)^{-1} \mathbf{r}^{(k-1)}, \tag{7.7}$$

que por componentes se expresa de la forma

$$x_i^{(k)} = (1-\omega)x_i^{(k-1)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}x_j^{(k)} - \sum_{j>i} a_{ij}x_j^{(k-1)}\right), \qquad i = 1, ..., n.$$

La cuestión que ahora se plantea es que valor del parámetro ω se debe tomar y como elegirlo. En algunos ejemplos en los que se conocen explícitamente de los autovalores de las matrices y se puede optimizar ω (ejemplo 7.1 de [1] o [3, sección 8.3]). En general y en la práctica, se suelen dar unos pocos pasos con varios valores de ω y después se continua con el valor más prometedor.

En general, el método de Gauss-Seidel es más rápido pero también más caro que el método de Jacobi [3, capítulo 8], aunque sin duda la referencia más completa para los métodos iterativos es el clásico [18] y más reciente [13].

8. Algunos comentarios finales

Son muchos las cuestiones importantes que no se han tratado por obvias razones de espacio que merecen algunos comentarios.

8.1. Análisis del error de los métodos directos. Los métodos directos obtendría la solución exacta si no hubiera errores de redondeo, y por tanto es muy importante conocer y estudiar el comportamiento de dichos errores. Estos análisis son regresivos en el sentido que Wilkinson inicio en su famoso libro [20] y resultan bastante laboriosos. Para los alumnos interesados se recomienda la introducción de [19] y después quizá profundizar en las referencias [7], [6] y [9].

8.2. Refinamiento iterativo. Suponiendo que se ha obtenido una solución aproximada $\widehat{\mathbf{x}}$ del sistema $\mathbf{Ax} = \mathbf{b}$ utilizando una factorización LU, el refinamiento iterativo calcula una nueva aproximación $\mathbf{x}$ en los tres pasos siguientes

$$\begin{aligned} \text{Calcula el residuo} &: \mathbf{r} = \mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}, \\ \text{Resuelve el sistema} &: \mathbf{Ad} = \mathbf{r}, \\ \text{Suma} &: \mathbf{x} = \widehat{\mathbf{x}} + \mathbf{d}, \end{aligned}$$

Este proceso se puede iterar hasta que el residuo sea suficientemente pequeño, aunque también necesitaría un estudio de la convergencia (ver por ejemplo el capítulo 10 de [9], ó [6]).

Fernando Vadillo

8.3. Matrices huecas. Una matriz hueca (también conocidas como sparse) es una matriz con un gran porcentaje de elementos cero, un ejemplo típico son las matrices tridiagonales que tienen la forma

$$\begin{pmatrix} b_1 & c_1 & & & \\ a_1 & b_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & a_{n-1} & b_n \end{pmatrix}.$$

También son habituales matrices muy grandes y con bandas más anchas en los métodos numéricos de aproximación de ecuaciones diferenciales (ver por ejemplo pg. 72 de [12]). En MATLAB tiene un tipo de dato llamado `sparse` que identifica y manipula las matrices huecas (ver capítulo 15 de [8] o [5] o [11]).

Existen métodos tanto directos como iterativos para estos tipos de matrices que aprovechan la estructura de la matriz, para el lector interesado se recomienda la introducción de [14], aunque la referencia más completa sea [13].

Referencias

1. U.M. Ascher and C. Greif, *A Firts Course in Numerical Methods*, SIAM, 2011.
2. K.E. Atkinson, *An Introduction to Numerical Analysis*, Wiley, 1978.
3. R. Bulirsch and J. Stoer, *Introduction to Numerical Analysis*, Springer, 1980.
4. S.C. Chapra and R.P. Canale, *Métodos numéricos para ingenieros*, McGraw-Hill Interamericana, 2002.
5. J.W. Demmel, *Applied Numerical Linear Algebra*, SIAM, 1997.
6. G. Forsythe and C.B. Moler, *Computer solution of linear algebraic systems*, Prentice-Hall, 1967.
7. G.H. Golub and C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, 1989.
8. D.J. Higham and N.J. Higham, MATLAB *guide*.
9. N.J. Higham, *Accuracy and Stability of Numerical Algoritms*, SIAM, 1996.
10. E. Isaacson and H. Keller, *Analysis of Numerical Methods*, Wiley, 1966.
11. C.D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, 2000.
12. C.B. Moler, *Numerical Computing with* MATLAB, SIAM, 2004.
13. Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, 1997.
14. R. Sacco, F. Saleri, and A. Quarteroni, *Numerical Mathematics*, Springer, 2000.
15. J.M. Sanz-Serna, *Diez lecciones de Cálculo Numérico*, Universidad de Valladolid, 2010.
16. G. Strang, *Introduction to Applied Matematic*, Wellesley-Cambridge Press, 1986.
17. L.N. Trefethen and D. Bau, *Numerical Linear Algebra*, SIAM, 1997.
18. R.S. Vargas, *Matrix Iterative Analysis*, Prentice Hall, 1962.
19. D.S. Walkins, *Fundaments of Matrix Computions. Second Edition*, John Wiley and Sons, 2002.
20. J.H. Wilkinson, *Rounding error in algebraic proceses*, Prehtice-Hall, 1963.

Leonardo Vadillo