MÉTODOS NUMÉRICOS BÁSICOS PARA LA RESOLUCIÓN DE SISTEMAS DE ECUACIONES LINEALES

x

F. Javier Pérez Fernández

=

Servicio de Publicaciones
UNIVERSIDAD DE CÁDIZ

F. JAVIER PÉREZ FERNÁNDEZ

Métodos numéricos básicos para la resolución de sistemas de ecuaciones lineales

UNIVERSIDAD DE CÁDIZ
SERVICIO DE PUBLICACIONES
1998

Métodos numéricos básicos para la resolución de sistemas de ecuaciones lineales / F. Javier Pérez Fernández. -- Cádiz: Universidad, Servicio de Publicaciones, 1998. -- 148 p.

ISBN 84-7786-543-4

1. Ecuaciones diferenciales lineales. 2. Métodos iterativos (Matemáticas). I. Universidad de Cádiz. Servicio de Publicaciones, ed. II. Título

512.64

F. Javier Pérez Fernández

I.S.B.N. 84-7786-543-4
Diseño de Cubierta: Creasur
Fotomecánica: Fotocromía Lineal S.L.
Imprime: INGRASA Artes Gráficas
D.L.: CA-983/98

A mi familia

Prólogo

Las funciones lineales y sus sistemas constituyeron el punto inicial de estudio del Algebra Lineal. Hoy día el tratamiento de estas cuestiones involucra un gran número de conceptos y conocimientos de esta rama de las Matemáticas, que tuvieron su origen en otros tipos de problemas. De esta forma al analizar y resolver sistemas de ecuaciones lineales hay que poner en juego una amplia gama de conceptos y técnicas del Algebra Lineal.

Por otra parte, parece incuestionable que al aplicar las matemáticas a problemas de la técnica, dado el volumen de variables y de relaciones que entre ellas que suelen aparecer, se requiere un tratamiento discreto.

Muy particularmente, los sistemas de ecuaciones lineales que se presentan en la práctica suelen ser de grandes dimensiones, por ello es necesario conocer no sólo el tratamiento típicamente algebraico del problema, sino también los distintos métodos numéricos con los que podrá abordarse, así como discernir la idoneidad de cada uno de ellos, en función de las carcterísticas del problema planteado.

La cuestión de encontrar métodos sencillos y efectivos para resolver tales sistemas, cuando tienen un gran número de ecuaciones e incógnitas, sigue siendo objeto de interés, pues las soluciones numéricas de tales ecuaciones son de especial importancia en problemas técnicos cuya modelización es lineal.

Presentamos una panorámica general de los métodos básicos, tratando su fundamentación matemática y los problemas de índole práctico que pueden aparecer, así como criterios sobre los que efectuar una valoración de la oportunidad de hacer una elección determinada, dependiendo de la situación problemática planteada. Se incluye un apéndice con las "rutinas usuales" en código C.

Los conocimientos necesarios para abordar la lectura de este libro son los correspondientes a un curso de Álgebra Lineal. No obstante, deseando que el trabajo sea lo más autocontenido posible, se ha introducido un capítulo sobre el estudio general de los sistemas de ecuaciones lineales. Con el mismo criterio, se ha justificado con notas a pie de página aquellos resultados del Álgebra Lineal que pudieran estar más en el olvido, este es el caso de algunas propiedades relativas a los autovalores.

El estudio de errores y del condicionamiento de una matriz, así como el de la convergencia de los métodos iterativos requiere el conocimiento de normas matriciales y de sucesiones y series de matrices, por ello las cuatro primeras secciones del capítulo siete se han dedicado a abordar estos temas, con la finalidad, ya manifestada, de que el lector no tenga que detener la lectura, si no ha visto o no recuerda estas cuestiones.

Nuestro agradecimiento a los Profesores Benítez Trujillo, Díaz Moreno, Pérez Cuéllar, Romero Romero y Aizpuru Tomás, por sus acertadas sugerencias.

Índice

1. Introducción

1.1 Necesidad de los métodos numéricos del Algebra Lineal.

El objeto de los métodos numéricos del Algebra Lineal es la resolución de sistemas de ecuaciones lineales, la inversión de matrices y la búsqueda de los vectores y valores propios de las matrices. Estos tres problemas están intimamente relacionados y muy particularmente la determinación de la inversa de una matriz y la solución de sistemas de ecuaciones lineales.

Desde una perspectiva teórica y estrictamente algebraica, estos problemas se resuelven con facilidad. Así se resuelve un sistema desarrollando los determinantes por la fórmula de Cramer, y se hallan los valores propios de una matriz escribiendo la ecuación característica y calculando sus raices.

Pero en la práctica, esta forma de actuar se hace inviable incluso para matrices de orden pequeño. Consideremos, por ejemplo, la resolución de un sistema de n ecuaciones con n incógnitas por el método de Cramer, y veamos el número de operaciones necesarias para resolverlo.

- El determinante de la matriz del sistema es

$$|A| = \sum_{1}^{n!} (-1)^{\sigma} a_{1i_1} a_{2i_2} a_{3i_3} \dots a_{ni_n},$$

  con σ = número de inversiones de $\{i_1, i_2, i_3, \dots, i_n\}$. Como hay $n!$ sumandos, entonces habrá que realizar $n! - 1$ sumas, cada sumando con n factores y por tanto con $(n-1)$ productos, con lo que el número de multiplicaciones es $n! \cdot (n-1)$.

- Pero como hay que resolver un total de $n+1$ determinantes, tendremos:

  Sumas: $S = (n! - 1) \cdot (n+1)$

  Productos: $P = n! \cdot (n-1) \cdot (n+1)$ y Divisiones: $D = n$.

Consecuentemente el número total de operaciones será:

$$T = (n! - 1) \cdot (n+1) + n! \cdot (n-1) \cdot (n+1) + n =$$

$$n! \cdot (n+1) - (n+1) + n! \cdot (n-1) \cdot (n+1) + n = n! \left[(n+1) + n^2 - 1\right] - 1 =$$

$$n! \cdot \left(n^2 + n\right) - 1 = (n+1)! \cdot n - 1.$$

Si por ejemplo una operación se hiciera en $1.5 \cdot 10^{-6}$ segundos, el número de operaciones y el tiempo para un sistema con diez ecuaciones serían del orden de $4 \cdot 10^8$ y 10 minutos respectivamente. Si el sistema tuviese 20 ecuaciones, el número de operaciones y el tiempo se elevarían al orden de 10^{21} y más de $475 \cdot 10^6$ años.

Estos resultados ponen claramente de manifiesto la imposibilidad de resolver un sistema mediante el método de Cramer con n relativamente pequeño. Pero incluso cuando n es pequeño, la acumulación de los errores de redondeo por el ordenador conlleva la obtención de soluciones sensiblemente alejadas de las exactas.

Resulta pues imprescindible el estudio de métodos que nos permitan de forma efectiva la resolución de estos problemas.

La extraordinaria frecuencia con que se presentan en la ingeniería problemas cuya modelización matemática involucra modelos lineales, requiere no sólo el conocimiento de los métodos básicos, sino también la capacidad para discernir, en función del problema planteado, cuál de ellos resultará más óptimo.

En este trabajo nos centraremos en los métodos numéricos básicos del Algebra Lineal relacionados con la resolución de sistemas de ecuaciones lineales.

1.2 Un problema: "Distribución de temperaturas en equilibrio".

Presentamos un problema[1] de ingeniería, cuya modelización matemática es lineal y para cuya resolución necesitaremos hacer uso de métodos numéricos.

Algunos tipos de buques, particularmente los químicos y los frigoríficos, tienen dependencias que se encuentran sometidas a bajísimas temperaturas. Por ejemplo, un barco químico puede transportar gases licuados, para lo que necesitará mantener el recipiente que los almacena a temperaturas inferiores a los menos treinta grados.

Los aceros pueden presentar una gran fragilidad ante bajas temperaturas, de forma que ante una carga eventual, producida por el mismo oleaje, se pueden fracturar con gran facilidad. Esta circunstancia deriva de la estructura interna de la materia y de las leyes que rigen el movimiento molecular.

Resulta pues de gran importancia conocer cuál será la resistencia de un material concreto cuando se ve sometido a unas temperaturas determinadas.

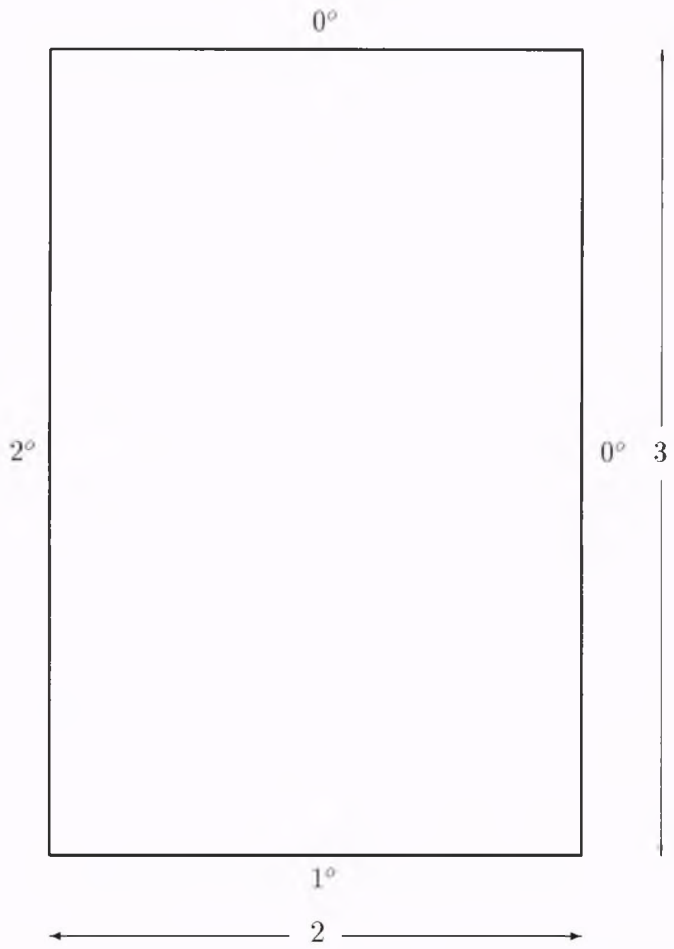
Supongamos por simplicidad una placa delgada cuyas caras están aisladas térmicamente, de manera que las únicas fuentes de calor se encuentran en su perímetro. Pasado un cierto momento inicial las temperaturas perimetrales se habrán distribuido por el interior de la placa, llegando a una situación de equilibrio. Nuestro interés radica en conocer la distribución de temperaturas en equilibrio en el interior de la placa; o más precisamente un esquema de su distribución, que lo proporciona las curvas que unen los puntos de igual temperatura (líneas isotérmicas).

La forma de la placa es indistinta para el problema, salvo la complejidad en un mayor número de incógnitas, como veremos pronto.

También son indistintas las temperaturas perimetrales, que supondremos siempre constantes. Por simplicidad consideraremos que la placa es rectangular, con lados de longitudes 2 y 3 unidades de medida respectivamente,

[1] Sobre una idea de RORRES & ANTON (1979), pgs. 95 – 109.

FIGURA 1.1:



y que las temperaturas perimetrales son de 2, 1 y 0 grados, tal y como se indica en la figura 1.1.

Para obtener un modelo matemático para el problema, aplicamos la siguiente propiedad de una distribución de temperaturas en equilibrio, consecuencia de algunas leyes de la termodinámica, para un problema como en el que estamos, en el que las fuentes de calor son externas y no hay ningún foco interior de calor.

La propiedad física que aplicaremos, que recibe el nombre de *Propiedad del Valor Medio,* podemos enunciarla de la forma siguiente[2]:

Sea una placa en equilibrio térmico y sea **P** *un punto interior de ella. Si* **C** *es un círculo cualquiera con centro en* **P**, *totalmente contenido en la placa, la temperatura en* **P** *es el valor medio de la temperatura del círculo.*

El problema de la determinación de las temperaturas interiores, con esta formulación continua se hace en la práctica inviable. Pero podemos discretizar el problema, calculando las temperaturas en "unos cuantos puntos interiores" y entonces el problema, en términos matemáticos, se reduce a un sistema de ecuaciones lineales. ¿Cuántos puntos interiores determinar?, dependerá de las dimensiones de la placa; pero en cualquier caso cuantos más se determinen más ajustada a la realidad serán las líneas isotérmicas de distribución de temperaturas.

Inicialmente discreticemos la propiedad física anterior y para ello pensemos previamente cómo determinar los puntos interiores, ¿cómo elegirlos?

Si sobre la placa superponemos una malla cuadriculada, tal y como indicamos en la figura 1.2, tendremos en el interior dos puntos sobre la malla.

Si la malla la subdividimos ahora, como aparece en la figura 1.3, tendremos 15. Estas subdivisiones podemos seguir haciéndolas, tantas como fina deseemos que sea nuestra aproximación a los datos reales.

Los puntos de intersección de la malla los llamaremos *puntos de encuentro.* Si están sobre el perímetro recibirán el nombre de puntos de encuentro perimetrales, y si se encuentran en el interior de la placa, puntos de encuentro interiores.

Nuestra preocupación consiste ahora en determinar las temperaturas en los puntos de encuentro interiores, de acuerdo con la *Propiedad del Valor Medio,* una vez discretizada y que es:

La temperatura en cada punto de encuentro interior es igual al promedio de las temperaturas de los cuatro puntos de encuentro circundantes más próximos.

Puesto que este principio sólo proporciona una aproximación al problema real, las temperaturas que mediante él se determinen sólo serán una aproximación a las reales. Cuánto más densa sea la malla, mejor será la apro-

[2]Para un análisis más detallado sobre su obtención, puede consultarse COSTA NOVELLA, E. (1986). *Ingeniería Química*, T. IV (Transmisión del Calor). Madrid, Alhambra Universidad; pág. 55 y siguientes.

FIGURA 1.2:

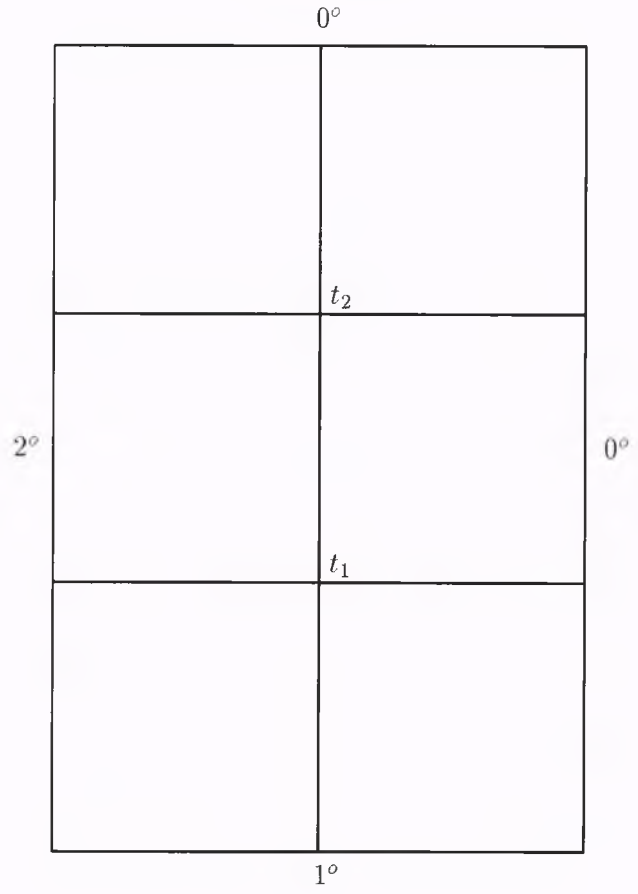
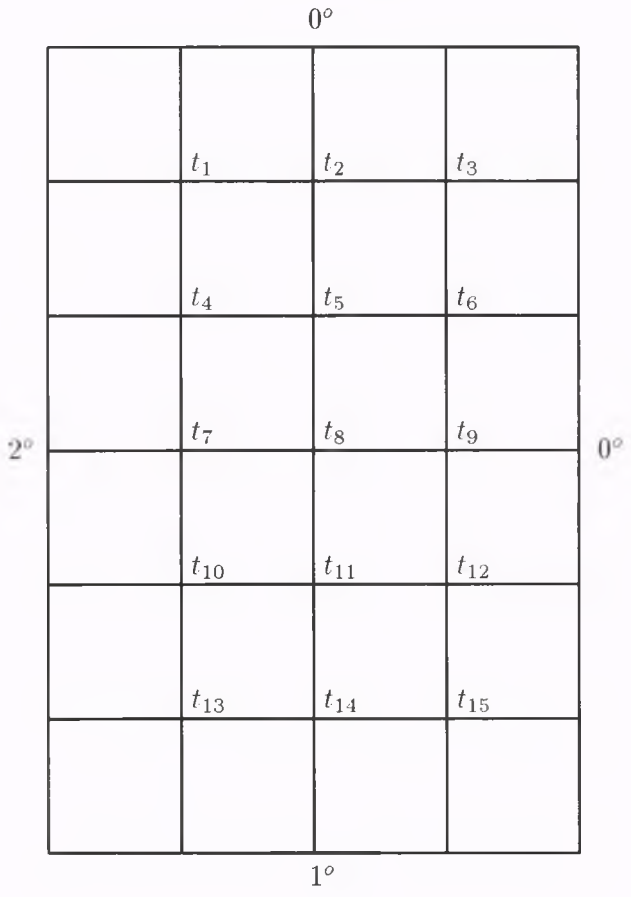


FIGURA 1.3:



ximación y también mayor el tamaño del sistema de ecuaciones lineales que habrá que resolver.

En el primer caso tenemos un sistema de dos ecuaciones con dos incógnitas:

$$\begin{aligned} t_1 &= 1/4\,(2+0+0+t_2) \\ t_2 &= 1/4\,(2+t_1+0+1) \end{aligned}$$

cuya matriz del sistema y cuyo vector de términos independientes son

$$A = \begin{bmatrix} 1 & -1/4 \\ 1/4 & 1 \end{bmatrix} \quad \text{y} \quad b = \begin{bmatrix} 1/2 \\ 3/4 \end{bmatrix}$$

y por tanto $t_1 = 0.6470$ y $t_2 = 0.5882$

En el segundo supuesto el sistema es de 15 ecuaciones con 15 incógnitas:

$$\begin{aligned} t_1 &= 1/4\,(2+0+t_2+t_4) & t_9 &= 1/4\,(t_8+t_6+0+t_{12}) \\ t_2 &= 1/4\,(t_1+0+t_3+t_5) & t_{10} &= 1/4\,(2+t_7+t_{11}+t_{13}) \\ t_3 &= 1/4\,(t_2+0+0+t_6) & t_{11} &= 1/4\,(t_{10}+t_8+t_{12}+t_{14}) \\ t_4 &= 1/4\,(2+t_1+t_5+t_7) & t_{12} &= 1/4\,(t_{11}+t_9+0+t_{15}) \\ t_5 &= 1/4\,(t_4+t_2+t_6+t_8) & t_{13} &= 1/4\,(2+t_{10}+t_{14}+1) \\ t_6 &= 1/4\,(t_5+t_3+0+t_9) & t_{14} &= 1/4\,(t_{13}+t_{11}+t_{15}+1) \\ t_7 &= 1/4\,(2+t_4+t_8+t_{10}) & t_{15} &= 1/4\,(t_{14}+t_{12}+0+1) \\ t_8 &= 1/4\,(t_7+t_5+t_9+t_{11}) \end{aligned}$$

cuya matriz es

$$\begin{bmatrix}
1 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{1}{4} & 1 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -\frac{1}{4} & 1 & 0 & 0 & -\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{1}{4} & 0 & 0 & 1 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 1 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 1 & 0 & 0 & -\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -\frac{1}{4} & 0 & 0 & 1 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 1 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 1 & 0 & 0 & -\frac{1}{4} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{4} & 0 & 0 & 1 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 1 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 1 & 0 & 0 & -\frac{1}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{4} & 0 & 0 & 1 & -\frac{1}{4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 1 & -\frac{1}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 1
\end{bmatrix}$$

y cuyo vector de términos independientes es

$$[1/2, 0, 0, 1/2, 0, 0, 1/2, 0, 0, 1/2, 0, 0, 3/4, 1/4, 1/4]^t$$

el proceso de eliminación gaussiana nos proporciona la solución:

$$\begin{aligned} &[t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}, t_{11}, t_{12}, t_{13}, t_{14}, t_{15}]^t = \\ &[0.9334, 0.4684, 0.2026, 1.265, 0.7377, 0.3422, 1.390, 0.8748, \\ &\qquad 0.4286, 1.420, 0.9226, 0.4975, 1.349, 0.9777, 0.6188]^t \end{aligned}$$

Si la malla no es muy densa en relación al tamaño de la placa, podemos seguir efectuando subdivisiones, tantas como queramos. La siguiente subdivisión proporciona un sistema de 77 ecuaciones con 77 incógnitas. Y para el siguiente caso, obtendremos un sistema cuya matriz es de orden 345, y hemos de pensar que aún en este caso la malla sea posiblemente muy grosera. En el siguiente paso tendremos ya 1457 ecuaciones e incógnitas.

Resolver un sistema de las dimensiones que estamos señalando requiere del uso de métodos más rápidos que el de Gauss. Piénsese que para una matriz de orden 345×345 la capacidad de un ordenador personal se vería seriamente comprometida. Para matrices tan especiales como la que nos ha aparecido[3] (con muchos ceros y los elementos no nulos simétricamente dispuestos en líneas paralelas a la diagonal de la matriz), ¿será posible valerse de su estructura "semi vacía" para gastar menos memoria del computador y poder así manejar matrices más grandes? En el supuesto de que ello fuera posible, ¿habrá otros métodos también, para otras matrices de características especiales?, ¿para cuáles?

Pero incluso para casos tan singulares como éste, cuando la matriz sea de orden 1457×1457 o incluso más, las posibilidades del método de Gauss o de otros similares, basados en la misma idea, ¿serán viables?, ¿podremos pensar en otra forma de abordar la cuestión?

Pensemos en el caso del sistema de quince ecuaciones. La matriz A del sistema podemos escribirla también de esta otra forma

$$A = I - B$$

donde I es la matriz identidad y B la siguiente matriz:

$$\begin{pmatrix}
0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0
\end{pmatrix}$$

El sistema tiene la forma:

$$At = b$$

o equivalentemente: $(I - B)t = b$, y por lo tanto podemos escribirlo como: $t = Bt + b$.

El vector t incógnita, aparece ahora en los dos miembros de la ecuación matricial. Pues bien, tomaremos un valor arbitrario inicial $t^{(0)}$, se sustituye en el segundo miembro de la igualdad y al resultado lo denominamos $t^{(1)}$.

[3] Es este un caso de matriz "banda" que posteriormente se estudiará.

Ahora volvemos a repetir el proceso sucesivamente, obteniendose la sucesión:

$$\begin{aligned} t^{(1)} &= Bt^{(0)} + b, \\ t^{(2)} &= Bt^{(1)} + b, \\ t^{(3)} &= Bt^{(2)} + b, \\ &\cdots \\ t^{(n)} &= Bt^{(n-1)} + b, \\ &\cdots \end{aligned}$$

Si en cada paso el valor $t^{(i)}$ se aproxima más a la solución exacta que su inmediato anterior, de manera que

$$\lim_{n\to\infty} t^{(n)} = t,$$

habremos encontrado un método que resulta muy efectivo, pero ¿será posible siempre partir de cualquier valor inicial y tener la seguridad de que las iteraciones convergerán a la solución exacta? El Teorema del punto fijo para aplicaciones contractivas en espacios métricos completos nos proporciona la solución, ya que estamos ante una ecuación del tipo $x = f(x)$; no obstante, aquí nos aproximaremos a esta cuestión sin apoyarnos directamente en este resultado. Y en el caso de que tal convergencia esté garantizada, ¿cuándo deberemos detenernos? ¿Cómo podremos conocer el grado de aproximación a la solución exacta en el momento que deseemos parar? ¿Habrá otras posibilidades de efectuar las iteraciones?

A todas estas interrogantes y a otras muchas que aprecerán, intentaremos dar respuesta en las páginas que siguen.

2. Estudio general de los sistemas de ecuaciones lineales

2.1 Generalidades.

Una ecuación lineal o de primer grado es una expresión de la forma

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b$$

tal que $a_i \in \mathbb{K}$ donde $i = 1, \ldots, n$ y $b \in \mathbb{K}$, siendo $\mathbb{K}$ un cuerpo, usualmente el de los reales o el de los complejos. Los a_i reciben el nombre de coeficientes y b el de término independiente. Y los x_i son variables que se suelen llamar incógnitas. Cuando el término independiente es 0, la ecuación se dice homogénea.

Un conjunto $\{r_1, r_2, \ldots, r_n\}$ de elementos del cuerpo $\mathbb{K}$, tales que

$$a_1r_1 + a_2r_2 + \cdots + a_nr_n = b,$$

se dirá que es una solución de la ecuación[1].

El estudio de estas ecuaciones requiere su transformación en otras más sencillas que tengan las mismas soluciones. Dos de tales ecuaciones se dirán equivalentes.

Un *sistema de ecuaciones lineales* es un conjunto de m ecuaciones lineales con n incógnitas:

$$\left.\begin{array}{lcl} a_{11}x_1 + a_{12}x_2 + & \cdots & + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + & \cdots & + a_{2n}x_n = b_2 \\ \vdots & \ddots & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + & \cdots & + a_{mn}x_n = b_m \end{array}\right\} \qquad (2.1)$$

donde todos los coeficientes y términos independientes pertenecen al mismo cuerpo $\mathbb{K}$: $a_{ij}, b_i \in \mathbb{K}$, con $1 \leq i \leq m$ y $1 \leq j \leq n$.

Un sistema se dice homogéneo cuando todas sus ecuaciones son homogéneas.

Una solución del sistema lo constituirá un conjunto de n elementos del cuerpo $\mathbb{K}$: $r_1, r_2, \ldots, r_n \in \mathbb{K}$ que satisfacen o son solución de todas y cada una de las ecuaciones del sistema.

Dos sistemas que tengan las mismas soluciones diremos que son equivalentes. Y en ello radica el proceso de resolución, en buscar sistemas equivalentes al inicial cuyas soluciones sean fáciles de determinar.

Aparte de la expresión analítica (2.1) del sistema, éste puede adoptar también estas otras formas:

[1] Un caso especialmente interesante de estas ecuaciones lo constituye las conocidas ecuaciones en congruencia módulo p, con p primo, que de forma equivalente podría escribirse como $ax + py = b$, que es la ecuación diofántica lineal, en la que se buscan soluciones enteras.

Matricial:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{2.2}$$

o simplificadamente $A \cdot x = b$, donde A es la llamada matriz del sistema, formada por los coeficientes del mismo, x es la matriz columna de las incógnitas y b la matriz columna de los términos independientes.

Vectorial:

$$x_1 A_1 + x_2 A_2 + \cdots + x_n A_n = b \tag{2.3}$$

donde $A_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{mi} \end{bmatrix}$, para $i = 1, \ldots, n$, y $b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$ son vectores del espacio vectorial $\mathbb{K}^m$.

Cuando un sistema tiene al menos una solución se dice compatible. En el caso contrario se llama incompatible. Si el sistema compatible tiene una solución única se llama determinado, e indeterminado si tiene más de una.

En lo que sigue consideraremos únicamente sistemas cuya matriz sea real, ya que si fuera complejo se podrá reducir a uno real, de la forma siguiente:

Sea $A \cdot x = b$, tal que tanto la matriz A como los vectores columnas están sobre el cuerpo de los complejos, por tanto: $A = M + iN$, $x = x_1 + ix_2$, $b = b_1 + ib_2$ con M, N, x_1, x_2, b_1, b_2 reales.

Tendremos $(M + iN)(x_1 + ix_2) = (b_1 + ib_2)$ o equivalentemente

$$Mx_1 - Nx_2 = b_1 \qquad \text{y} \qquad Nx_1 + Mx_2 = b_2,$$

es decir

$$\begin{bmatrix} M & -N \\ N & M \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

Si el sistema original era de orden n, es decir tenía n ecuaciones y n incógnitas, el nuevo será de orden $2n$. No obstante, hay que indicar que algunos lenguajes como FORTRAN permiten trabajar con aritmética compleja, por lo que es posible el tratamiento del sistema sin necesidad de reducirlo a uno real, aunque no suele hacerse, debido a que el tiempo necesario de esta forma es mayor que el empleado mediante la referida reducción.

2.2 Estructura de las soluciones.

Como $\mathbb{R}^n$ y $\mathbb{R}^m$ son espacios vectoriales reales y como, fijadas unas bases, toda matriz real A, de dimensión $m \times n$, puede identificarse con una aplicación lineal: $f: \mathbb{R}^n \to \mathbb{R}^m$ tal que $f(x) = A \cdot x$ podemos, entonces, estudiar los sistemas de ecuaciones lineales sobre la base de esta estrecha relación con las aplicaciones lineales. Así el sistema (2.1), cuya expresión matricial es $A \cdot x = b$, puede interpretarse como $f(x) = b$.

Desde esta perspectiva, el conjunto solución del sistema de ecuaciones lineales es la preimagen por f de b y, por tanto, $r = (r_1, \ldots, r_n)$ es una solución del sistema si y sólo si $r \in f^{-1}(b)$. Naturalmente, si la preimagen de b es el conjunto vacío entonces el sistema $A \cdot x = b$ será incompatible; es más, el sistema es incompatible si y sólo si $b \notin \operatorname{Im} f$, donde con $\operatorname{Im} f$ denotamos a la imagen de la aplicación lineal f. Es también evidente que $A \cdot x = b$ es compatible si y sólo si $b \in \operatorname{Im} f$.

Como $f(0) = 0$, un sistema lineal homogéneo $A \cdot x = 0$ será siempre compatible y el vector nulo $0 \in \mathbb{R}^n$ se llama solución trivial del mismo.

Si denotamos por $\mathbb{S}$ el conjunto solución de $A \cdot x = 0$, es claro que se trata del núcleo de la aplicación f:

$$\mathbb{S} = N(f) = f^{-1}(0),$$

por tanto, $\mathbb{S}$ es un subespacio vectorial de $\mathbb{R}^n$, llamado espacio solución del sistema lineal homogéneo $A \cdot x = 0$.

Para sistemas homogéneos, la existencia de soluciones distintas de la trivial dependerá de que la aplicación f no sea inyectiva.

Sea $r = r(A)$ el rango de la matriz del sistema; entonces, de la relación entre las dimensiones del Núcleo y la Imagen de una aplicación lineal, se tiene que la dimensión de $\mathbb{S}$ es:

$$\dim \mathbb{S} = n - r.$$

Por consiguiente, si el rango de la matriz del sistema es r, entonces el número de soluciones linealmente independientes es $n - r$, y las restantes soluciones son combinaciones lineales de aquellas[2].

Se denomina sistema homogéneo asociado al sistema no homogéneo (2.1), cuya expresión matricial es $A \cdot x = b$, al sistema de ecuaciones lineales homogéneo $A \cdot x = 0$, cuya matriz del sistema es la misma que la del sistema (2.1).

Consideremos un sistema no homogéneo (2.1), en la forma $A \cdot x = b$, compatible, y sea x_0 una solución del mismo. Si x es una solución cualquiera de (2.1), entonces

$$A \cdot (x - x_0) = 0.$$

[2]Es más, si el sistema homogéneo es cuadrado $m = n$:

$$\left.\begin{array}{llll} a_{11}x_1 + a_{12}x_2 + & \cdots & +a_{1n}x_n = 0 \\ a_{21}x_1 + a_{22}x_2 + & \cdots & +a_{2n}x_n = 0 \\ \vdots & \ddots & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + & \cdots & +a_{nn}x_n = 0 \end{array}\right\}$$

puede interpretarse en la forma:

$$[a_{i1}, \ldots, a_{in}] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = 0 \quad \text{con} \quad 1 \leq i \leq n$$

lo que en términos de un espacio euclídeo se interpreta como que toda solución x es ortogonal a $\{F_1, \ldots, F_n\}$, donde $F_i = (a_{i1}, \ldots, a_{in})$, con lo que se tiene que $\mathbb{S}$ es ortogonal a $L = L(F_1, \ldots, F_n)$. Y como $\dim(\mathbb{S})$ es $n - r$ y $\dim(L) = r$, entonces $\mathbb{S}$ es complemento ortogonal de L.

luego $x - x_0$ es solución de $A \cdot x = 0$. Y como $x = x_0 + (x - x_0)$, entonces toda solución de $A \cdot x = b$ se puede expresar como suma de x_0 y una solución del sistema homogéneo asociado, de donde resulta que el conjunto de soluciones de $A \cdot x = b$ está contenido en $x_0 + \mathbb{S}$.

Recíprocamente, si $y \in \mathbb{S}$ entonces $A \cdot (x_0 + y) = b$, por lo que $x_0 + \mathbb{S}$ está contenido en el conjunto de soluciones de (2.1).

■

Por lo que hemos obtenido el siguiente resultado.

TEOREMA 1 *Sea $A \cdot x = b$ un sistema no homogéneo compatible, y sea x_0 una solución del mismo. Entonces el conjunto de las soluciones de $A \cdot x = b$ es precisamente $x_0 + \mathbb{S}$, donde $\mathbb{S}$ es el espacio solución del sistema homogéneo asociado.*

Es evidente que un sistema compatible no homogéneo o tiene una única solución (S.C.D.) o si tiene más de una, tiene infinitas (S.C.I.)[3].

2.3 Análisis de los sistemas.

Dado el sistema $A \cdot x = b$, si $b \in \mathrm{Im} f$, el sistema es compatible, y como $\mathrm{Im} f$ es el subespacio generado por las columnas de A, $\mathrm{Im} f = L(A_1, \ldots, A_n)$, dependerá de que los vectores $A_i \in \mathbb{R}^m$, columnas de la matriz A, constituyan una familia libre o ligada, el que el sistema sea determinado o indeterminado.

Desde luego, si $\{A_i\}_{1 \leq i \leq n}$ es una familia libre, entonces la expresión de b será única y, por tanto, el sistema compatible determinado (S.C.D.).

Si $\{A_i\}_{1 \leq i \leq n}$ es una familia ligada, evidentemente la expresión de b respecto de ella no será única, por lo que estaremos ante un sistema compatible indeterminado (S.C.I.).

Es ya, por tanto, evidente, que para que el sistema tenga solución es condición necesaria y suficiente que $b \in L(A_1, \ldots, A_n)$ y por tanto, si y sólo si los sistemas de vectores $\{A_1, A_2, \ldots, A_n\}$ y $\{A_1, A_2, \ldots, A_n, b\}$ tienen el mismo rango.

Dado el sistema $A \cdot x = b$, denominamos matriz ampliada del sistema, y la denotaremos por $\widehat{A}$, a la matriz del sistema ampliada con la columna de los términos independientes:

$$\widehat{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{bmatrix}$$

En términos de la matriz A del sistema y de su matriz ampliada $\widehat{A}$, tenemos el siguiente resultado, conocido como *Teorema de Rouché-Fröbenius.*

[3] Si no estuviésemos trabajando sobre $\mathbb{R}$ y el cuerpo $\mathbb{K}$ soporte de nuestro sistema es un cuerpo de Galois, esta afirmación no resulta adecuada. Por ejemplo si $\mathbb{K} = \mathbb{Z}_2$ entonces los elementos de $\mathbb{Z}_2^n$ son n-uplas formadas por "0" y "1", luego habrá un máximo de 2^n vectores, por lo que $\mathbb{S} \subseteq \mathbb{Z}_2^n$ tendrá un número finito de elementos. Por tanto, si el sistema es compatible indeterminado tendrá más de una solución, pero no en número infinito.

TEOREMA 2 *El sistema es compatible si y sólo si* $r(A) = r\left(\widehat{A}\right)$. *En el caso de que existan soluciones:*

1. *Si* $r(A) = r\left(\widehat{A}\right) = n$, *entonces* $\{A_1, A_2, \ldots, A_n\}$ *es linealmente independiente y consecuentemente el sistema es compatible determinado.*
2. *Si* $r(A) = r\left(\widehat{A}\right) < n$, *entonces,* $\{A_1, A_2, \ldots, A_n\}$ *es un sistema ligado y por tanto el sistema será compatible indeterminado.*

2.4 Sistemas equivalentes.

Obvio es decir que la resolución de un sistema que tuviese forma diagonal:

$$\left.\begin{array}{rrrr} a_{11}x_1 & & & = b_1 \\ & a_{22}x_2 & & = b_2 \\ & & \ddots & \\ & & a_{nn}x_n & = b_n \end{array}\right\}$$

es inmediata.

Algo menos inmediata, pero también muy cómoda resultaría la resolución de un sistema triangular:

$$\left.\begin{array}{rrrrr} a_{11}x_1 & +a_{12}x_2 & +\cdots & +a_{1n}x_n & = b_1 \\ & +a_{22}x_2 & +\cdots & +a_{2n}x_n & = b_2 \\ & & \ddots & \vdots & \vdots \\ & & & +a_{nn}x_n & = b_n \end{array}\right\}$$

pues bastaría ir de abajo hacia arriba calculando $x_n, x_{n-1}, \ldots, x_1$.

En la resolución de un sistema convendrá con frecuencia, transformarlo en otro equivalente diagonal o triangular, cuya resolución resulte inmediata.

Para transformar un sistema en otro equivalente utilizaremos transformaciones elementales sobre la matriz ampliada del sistema. A continuación señalamos algunos resultados conocidos sobre transformaciones elementales, que nos serán de utilidad en el desarrollo del tema.

Se dice que sobre una matriz cualquiera A se ha realizado una transformación elemental de línea (fila o columna), si se ha efectuado alguna de las siguientes operaciones entre las líneas de la misma:

1. Permutación de líneas. Operación que simbolizaremos por F_{ij}, para indicar que se ha intercambiado la i-ésima fila por la j-ésima fila; o bien por C_{ij}, cuando se trate de columnas.
2. Adición a una línea de otra multiplicada por un escalar. Operación que simbolizaremos por $F_{ij}(\lambda)$, para indicar que la i-ésima fila se reemplaza por si misma más λ veces la j-ésima fila. Análogamente $C_{ij}(\lambda)$ para el caso en que las líneas sean columnas.
3. Producto de una línea por un escalar no nulo λ. Operación que simbolizaremos por $F_i(\lambda)$ para indicar que la i-ésima fila se ha multiplicado por λ. Análogamente $C_i(\lambda)$ para cuando se trate de columnas.

La matriz resultante de realizar una transformación elemental sobre la matriz identidad I, recibe el nombre de matriz elemental. Y el efecto de realizar una transformación elemental sobre una matriz A de orden mxn, es el mismo que el de multiplicar A por la matriz elemental que resultaría de efectuar esa transformación elemental sobre la matriz identidad I (a la derecha de A si se trata de transformación elemental de columna y a la izquierda de A si se trata de transformación elemental de fila). Desde luego toda matriz elemental es invertible y la inversa es también una matriz elemental.

Con estos conocimientos, a nivel de prerrequisitos, estamos en condiciones de obtener algunos resultados necesarios para la transformación se un sistema en otro equivalente.

TEOREMA 3 *Si un sistema lo multiplicamos, por la izquierda, por una matriz regular, el nuevo sistema que se obtiene es equivalente con el primero.*

Demostración. En efecto. Sea M una matriz regular de orden $m \times m$. Si x es solución de $A \cdot x = b$ entonces es obvio que también lo es de $M \cdot A \cdot x = M \cdot b$. Y recíprocamente, si x es solución de $M \cdot A \cdot x = M \cdot b$, entonces se tendrá que:

$$M^{-1} \cdot M \cdot A \cdot x = M^{-1} \cdot M \cdot b$$

y consecuentemente: $A \cdot x = b$. ■

TEOREMA 4 *Si se realizan k transformaciones elementales de fila sobre un sistema de ecuaciones lineales $Ax = b$, el sistema resultante es equivalente al inicial.*

Demostración. En efecto. Ya que ello es equivalente a multiplicar por la izquierda de A por k matrices elementales, y por tanto regulares, con lo que el nuevo sistema será equivalente al inicial, como consecuencia del Teorema 3. ■

TEOREMA 5 *Si en un sistema una ecuación es combinación lineal de otras, entonces el sistema que se obtiene al suprimir dicha ecuación, es equivalente al primero.*

Demostración. En efecto. Supongamos que es la k-ésima ecuación la que es combinación de las restantes. Designando la i-ésima ecuación por E_i, podemos simbolizarlo de la siguiente manera:

$$E_k = \sum_{\substack{i=1 \\ i \neq k}}^{m} \lambda_i E_i.$$

Realizando sobre E_k las $m-1$ transformaciones elementales de fila siguientes:

$$F_{ki}(-\lambda_i) \quad \text{con } 1 \leq i \leq m \text{ e } i \neq k,$$

obtendremos en el lugar de la k-ésima ecuación la identidad: $0 = 0$.

Y desde luego, el sistema resultante es equivalente al inicial, por haberse obtenido de él mediante transformaciones elementales. Obviamente, si suprimimos en el sistema resultante la citada identidad, el nuevo sistema de $m - 1$ ecuaciones sigue siendo equivalente al inicial. ■

- *Este resultado podemos ampliarlo a dos sistemas cualesquiera, con número distinto de ecuaciones, en la forma siguiente:*

Sean

$$\left.\begin{array}{lcl} a_{11}x_1 + a_{12}x_2 + & \cdots & + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + & \cdots & + a_{2n}x_n = b_2 \\ \vdots & \ddots & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + & \cdots & + a_{mn}x_n = b_m \end{array}\right\} \tag{2.4}$$

y

$$\left.\begin{array}{lcl} c_{11}x_1 + c_{12}x_2 + & \cdots & + c_{1n}x_n = d_1 \\ c_{21}x_1 + c_{22}x_2 + & \cdots & + c_{2n}x_n = d_2 \\ \vdots & \ddots & \vdots \\ c_{p1}x_1 + c_{p2}x_2 + & \cdots & + c_{pn}x_n = d_p \end{array}\right\} \tag{2.5}$$

dos sistemas con el mismo número de incógnitas, pero con distinto número de ecuaciones.

Consideremos los siguientes vectores de $\mathbb{R}^{n+1}$ asociados respectivamente a ambos sistemas:

$$\left\{\begin{array}{l} u_1 = (a_{11}, \ldots, a_{1n}, b_1) \\ \vdots \\ u_m = (a_{m1}, \ldots, a_{mn}, b_m) \end{array}\right. \quad \text{y} \quad \left\{\begin{array}{l} v_1 = (c_{11}, \ldots, c_{1n}, d_1) \\ \vdots \\ v_p = (c_{p1}, \ldots, c_{pn}, d_p) \end{array}\right.$$

y las variedades lineales engendradas por ambos grupos:

$$L(u_1, u_2, \ldots, u_m) \quad \text{y} \quad L(v_1, v_2, \ldots, v_p).$$

Pues bien:

TEOREMA 6 *Si $L(u_1, u_2, \ldots, u_m) = L(v_1, v_2, \ldots, v_p)$, entonces los sistemas (2.4) y (2.5) son equivalentes.*

Es decir, si las combinaciones lineales de las ecuaciones del sistema (2.4) coinciden con las del sistema (2.5) entonces los sistemas son equivalentes, con lo que un sistema se podrá sustituir por otro obtenido del primero mediante combinaciones lineales de sus ecuaciones, sin alterar las soluciones. Las combinaciones lineales buscadas serán aquellas que simplifiquen el sistema de ecuaciones original.

Demostración. Al ser $L(u_1, u_2, \ldots, u_m) = L(v_1, v_2, \ldots, v_p)$, entonces el sistema formado uniendo las ecuaciones (2.4) y (2.5) es equivalente a cualquiera de los dos iniciales (2.4) o (2.5), por el Teorema 5. Consecuentemente serán equivalentes entre si. ■

2.5 Consideraciones generales sobre la resolución de Sistemas de Ecuaciones Lineales.

Teniendo en cuenta los resultados obtenidos sobre sistemas equivalentes, para resolver un sistema compatible es conveniente, previamente, eliminar las ecuaciones dependientes puesto que son redundantes. Este proceso de "limpieza" puede de hecho simultanearse con el de resolución.

Si después de este proceso de "limpieza" quedan menos ecuaciones independientes que incógnitas, supuesta la compatibilidad, estaremos ante un caso de sistema compatible indeterminado. Parametrizando tantas incógnitas como la diferencia entre el número total de incógnitas y el número de ecuaciones obtendremos un sistema cuadrado compatible determinado.

A las incógnitas parametrizadas las denominaremos incógnitas libres y a las no parametrizadas, cuyos valores vendrán dados en función de las anteriores, las llamaremos incógnitas básicas.

En lo sucesivo sólo consideraremos sistemas cuadrados, con el mismo número de ecuaciones que de incógnitas, una vez "limpiado" y parametrizado, y tal que la matriz del sistema resultante sea regular.

Nos interesará, pues, estudiar métodos para la resolución de este tipo de sistemas.

3. Métodos directos de Gauss y Gauss-Jordan

3.1 Introducción.

Como ya indicamos anteriormente el uso de métodos numéricos para la resolución de sistemas de ecuaciones lineales es una necesidad incluso para sistemas con un número muy pequeño de ecuaciones.

Abordaremos en este epígrafe y en los tres siguientes el estudio de algunos métodos directos.

Los métodos directos pretenden, en general, obtener una solución exacta del sistema con un número finito de operaciones elementales. Aunque desde una perspectiva teórica esto siempre es posible, en la práctica, desgraciadamente no siempre es así como podremos ver más adelante.

En los métodos iterativos nos aproximamos a la solución partiendo de un valor arbitrario, mediante una sucesión de vectores cuyo límite es la solución del sistema original.

Por consiguiente hay una sustancial diferencia de índole cualitativa y metodológica, que se pondrá totalmente en evidencia tras el estudio de diversos métodos de ambas naturalezas.

Tras él podrá comprenderse cómo, en el caso de computadores pequeños, los métodos directos son válidos para sistemas de orden no superior a 250. Y aunque el avance en la estructura de la memoria de los ordenadores, el aumento de su rapidez de funcionamiento y el desarrollo de los procesos de cálculo hacen válidos aquellos métodos para órdenes mayores, del orden de 10^3 e incluso más; no obstante, las exigencias de memoria y la cantidad de operaciones hacen que los métodos directos resulten demasiado lentos, costosos e incluso inviables.

En lo que sigue consideraremos sistemas cuadrados de matriz A regular:

$$A \cdot x = b.$$

3.2 Sistemas Triangulares.

Los métodos directos suelen transformar la resolución del sistema en la de un sistema triangular del tipo $U \cdot x = b$,

$$\left.\begin{array}{lllll} u_{11}x_1 & +u_{12}x_2 & +\cdots & +u_{1n}x_n & = b_1 \\ & +u_{22}x_2 & +\cdots & +u_{2n}x_n & = b_2 \\ & & \ddots & \vdots & \vdots \\ & & & +u_{nn}x_n & = b_n \end{array}\right\},$$

donde U es una matriz triangular superior $U = (u_{ij})$, con $u_{ij} = 0$ si $i > j$, que se resuelve fácilmente por recurrencia, por sustitución desde detrás hacia adelante:

$$x_n = \frac{b_n}{u_{nn}}, \quad x_i = \frac{b_i - \sum_{j=i+1}^{n} u_{ij}x_j}{u_{ii}}, \quad \text{con } i = n-1, \ldots, 1.$$

3.2.1 Coste en número de operaciones.

Teniendo en cuenta que para calcular x_i se necesitan:

1. $n - i$ productos, al sustituir los valores $x_n, \ldots, x_{i+1}$ ya calculados,
2. $n - i$ sumas/restas, resultado de "despejar" $u_{ii}x_i$,
3. y una división

en total serán precisas $\sum_{i=1}^{n} n-i$ sumas $+ \sum_{i=1}^{n} n-i$ productos $+n$ divisiones, lo que hace un total de n^2 operaciones.

Obviamente este número es sensiblemente inferior al que se obtuvo para un sistema ordinario mediante Cramer.

Los métodos directos se basan en transformar el sistema inicial en otro triangular y equivalente a aquél. O si se prefiere en hallar una matriz M invertible, que resulte fácil de calcular y tal que $M \cdot A$ sea triangular, de forma que el sistema equivalente $M \cdot Ax = Mb$ se puede resolver por sustitución hacia atrás.

3.3 Método de Gauss.

3.3.1 Proceso de eliminacion de Gauss.

El método de Gauss es uno de estos métodos directos que procede en dos pasos, uno en el que se triangulariza el sistema (produciendo ceros debajo de la diagonal de A) y otro en el que se resuelve el sistema triangular resultante. Al primer paso mediante el que se reduce la matriz inicial a una triangular superior lo denominaremos "eliminación gaussiana" , y al segundo, como ya hemos hecho antes, "sustitución hacia atrás".[1]

El procedimiento no es más que una generalización del conocido método escolar de reducción.

Sea $A \cdot x = b$ reescrito de la forma: $A^{(1)}x = b^{(1)}$ y sea la matriz ampliada del sistema $\widehat{A}$, escrita ahora como $\widehat{A}^{(1)}$. Realizamos los siguientes pasos:

1. Hacemos las transformaciones elementales:
   1.1 $F_{i1}\left(-a_{i1}^{(1)}/a_{11}^{(1)}\right)$ para $i = 2, \ldots, n$ en la matriz $\widehat{A}^{(1)}$, suponiendo que $a_{11}^{(1)} \neq 0$, con lo que habremos conseguido pasar a un sistema

[1] Hay autores que con "eliminación de Gauss" denominan conjuntamente a ambos pasos.

equivalente, de matriz ampliada:

$$\widehat{A}^{(2)} = \left[\begin{array}{cccc|c} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{array}\right]$$

siendo $a_{ij}^{(2)} = a_{ij}^{(1)} + m_{i1}a_{1j}^{(1)}$, con $2 \leq i, j \leq n$ y $m_{i1} = -a_{i1}^{(1)}/a_{11}^{(1)}$ y $b_i^{(2)} = b_i^{(1)} + m_{i1}b_1^{(1)}$, $2 \leq i \leq n$. El elemento $a_{11}^{(1)}$ recibe el nombre de primer pivote de la eliminación.

Matricialmente podemos escribirlo en la forma: $A^{(2)}x = b^{(2)}$, donde, como es sabido, la matriz $A^{(2)}$ obtenida de $A^{(1)}$ tras $n-1$ transformaciones elementales, puede escribirse en la forma:

$$\widehat{A}^{(2)} = M^{(1)} \cdot A^{(1)}$$

con

$$M^{(1)} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ m_{21} & 1 & 0 & \cdots & 0 \\ m_{31} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{n1} & 0 & 0 & \cdots & 1 \end{bmatrix}$$

y $b^{(2)} = M^{(1)}b^{(1)}$.

Y desde luego ambos sistemas, el inicial y el obtenido, son equivalentes, sin más que tener en cuenta el Teorema 4.

1.2 Si $a_{11}^{(1)} = 0$, bastará efectuar una permutación, F_{1i}, de la fila primera por cualquier otra fila i, en la que el elemento $a_{i1}^{(1)}$ sea no nulo, y aplicar luego las transformaciones elementales indicadas en 1.1.

Dos cuestiones tendremos que comprobar: una, que una tal permutación es siempre posible, es decir, que siempre existe un elemento de la matriz $a_{i1}^{(1)} \neq 0$ con $1 < i \leq n$; y dos, cómo afecta el intercambio de filas al proceso de resolución.

La primera cuestión es inmediata, pues de no existir en la primera columna ningún elemento no nulo, resultaría obviamente que la matriz $A^{(1)} = A$ sería singular, en contra de nuestra hipótesis de partida de que A es una matriz regular.

En cuanto a la segunda cuestión, es también evidente que el nuevo sistema obtenido tras la permutación de filas es equivalente al inicial, pues es el resultado de una transformación elemental de fila sobre el inicial (Teorema 4).

2. Realizamos ahora las transformaciones elementales:

2.1 $F_{i2}\left(-a_{i2}^{(2)}/a_{22}^{(2)}\right)$ para $i = 3, \ldots, n$ en la matriz $\widehat{A}^{(2)}$, suponiendo que $a_{22}^{(2)} \neq 0$, con lo que habremos conseguido pasar a un sistema equivalente, de matriz ampliada:

$$\widehat{A}^{(3)} = \left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{2n}^{(3)} & b_3^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \cdots & a_{nn}^{(3)} & b_n^{(3)} \end{array}\right]$$

siendo $a_{ij}^{(3)} = a_{ij}^{(2)} + m_{i2}a_{2j}^{(2)}$, con $3 \leq i, j \leq n$ y $m_{i2} = -a_{i2}^{(2)}/a_{22}^{(2)}$ y $b_i^{(3)} = b_i^{(2)} + m_{i2}b_2^{(2)}$, $3 \leq i \leq n$. El elemento $a_{22}^{(2)}$ recibe el nombre de segundo pivote de la eliminación.

Matricialmente podemos escribirlo en la forma: $A^{(3)}x = b^{(3)}$, donde como es sabido la matriz $A^{(3)}$obtenida de $A^{(2)}$ tras $n-2$ transformaciones elementales, puede escribirse como el siguiente producto $A^{(3)} = M^{(2)} \cdot A^{(2)}$, con

$$M^{(2)} = \left[\begin{array}{ccccc} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & m_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & m_{n2} & 0 & \cdots & 1 \end{array}\right]$$

y $b^{(3)} = M^{(2)}b^{(2)}$.

Naturalmente, como antes, el sistema $A^{(3)}x = b^{(3)}$ será equivalente al $A^{(2)}x = b^{(2)}$ y, por tanto, equivalente al $A^{(1)}x = b^{(1)}$.

2.2 Si $a_{22}^{(2)} = 0$, bastará efectuar una permutación, F_{2i}, de la fila segunda por cualquier otra fila i tal que $2 < i \leq n$, en la que el elemento $a_{i2}^{(2)}$ sea no nulo, y aplicar luego las transformaciones elementales indicadas en 2.1.

Desde luego tal permutación es siempre posible, pues caso contrario sería porque en la matriz:

$$A^{(2)} = \left[\begin{array}{cccc} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{array}\right]$$

$a_{i2}^{(2)} = 0$, para $i = 2, \ldots, n$. Pero en este caso se tendría que:

$$\left|A^{*(2)}\right| = \left|\begin{array}{ccc} a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \ddots & \vdots \\ a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{array}\right| = 0,$$

al ser la primera columna nula, de donde

$$\left|A^{(2)}\right| = a_{11}^{(1)} \cdot \left|A^{*(2)}\right| = 0.$$

Y como $A^{(2)}$ se ha obtenido de A mediante transformaciones elementales, resultaría que necesariamente también A habría de ser singular, en contra de nuestra hipótesis inicial.

Además el nuevo sistema sigue siendo equivalente al inicial, pues es el resultado de una transformación elemental de fila.

3. Continuamos este proceso sucesivamente:

3.1 Generando un total de $n-1$ ecuaciones de la forma: $A^{(r)}x = b^{(r)}$, con $r = 2, \ldots, n$ y donde $\widehat{A}^{(r)} = \left(A^{(r)}|b^{(r)}\right)$ es

$$\widehat{A}^{(r)} = \left[\begin{array}{ccccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & \cdots & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & \cdots & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{rr}^{(r)} & \cdots & a_{rn}^{(r)} & b_r^{(r)} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nr}^{(r)} & \cdots & a_{nn}^{(r)} & b_n^{(r)} \end{array}\right]$$

obteniéndose $A^{(r+1)}x = b^{(r+1)}$ a partir de $A^{(r)}x = b^{(r)}$ mediante las transformaciones elementales de fila $F_{ir}\left(-a_{ir}^{(r)}/a_{rr}^{(r)}\right)$, siendo $r < i \leq n$, suponiendo que $a_{rr}^{(r)} \neq 0$.

De esta forma obtenemos $\widehat{A}^{(r+1)}$, donde $a_{ij}^{(r+1)} = a_{ij}^{(r)} + m_{ir}a_{rj}^{(r)}$, con $r < i, j \leq n$ y $m_{ir} = -a_{ir}^{(r)}/a_{rr}^{(r)}$ con $r < i \leq n$, y siendo $b_i^{(r+1)} = b_i^{(r)} + m_{ir}b_r^{(r)}$, $r < i \leq n$.

De esta forma, $A^{(r+1)} = M^{(r)} \cdot A^{(r)}$ y $b^{(r+1)} = M^{(r)}b^{(r)}$, con

$$M^{(r)} = \left[\begin{array}{ccccc} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & m_{(r+1)r} & \\ & & & \vdots & \ddots \\ & & & m_{nr} & 1 \end{array}\right].$$

Tras los $n-1$ pasos se llega a $A^{(n)} \cdot x = b^{(n)}$ que es un sistema triangular superior, que se resuelve de forma elemental mediante un proceso de sustitución hacia atrás.

3.2 Si en algún paso $a_{rr}^{(r)} = 0$, bastará efectuar una permutación, F_{ri}, de la fila r-ésima por cualquier otra fila i, con $r < i \leq n$, en la que el elemento $a_{ir}^{(r)}$ sea no nulo, y aplicar luego las transformaciones elementales indicadas en 3.1.

Como en los casos anteriores, esta transformación es siempre posible, pues de lo contrario se deduciría que

$$\left|A^{*(r)}\right| = \left|\begin{array}{ccc} a_{rr}^{(r)} & \cdots & a_{rn}^{(r)} \\ \vdots & \ddots & \vdots \\ a_{nr}^{(r)} & \cdots & a_{nn}^{(r)} \end{array}\right| = 0$$

y consecuentemente que

$$|A^{(r)}| = a_{11}^{(1)} \cdot a_{22}^{(2)} \cdot \ldots \cdot a_{r-1r-1}^{(r-1)} \cdot |A^{*(r)}| = 0, \qquad (3.1)$$

por lo que A también sería singular, pues $A^{(r)}$ se ha obtenido a partir de A mediante transformaciones elementales, lo que obviamente es imposible.

Y, desde luego, el sistema que se obtiene en cada caso tras una permutación de filas sigue siendo equivalente al anterior.

Por consiguiente, finalmente, tanto si se han tenido que realizar permutaciones de filas como si no, se llega a un sistema triangular, $A^{(n)} \cdot x = b^{(n)}$, equivalente al inicial, tal que se puede poner

$$A^{(n)} = M \cdot A \qquad \text{y} \qquad b^{(n)} = M \cdot b, \qquad (3.2)$$

donde M es la matriz que se obtiene de multiplicar todas las matrices elementales, correspondientes a las transformaciones elementales realizadas sobre la matriz ampliada del sistema.

Una vez se ha obtenido el sistema triangular equivalente al sistema inicial, se resuelve mediante "sustitución hacia atrás", en la forma indicada en el epígrafe **3.2.**

Cuando no es preciso realizar permutaciones de filas, debido a que los sucesivos pivotes que van apareciendo son todos no nulos, el proceso recibe el nombre de eliminación "simple" de Gauss. En caso contrario se le suele denominar proceso de eliminación de Gauss "con intercambios".

Por tanto, en ocasiones es necesario efectuar permutación de filas para evitar los ceros en lugares pivotales. Más adelante veremos que el intercambio de filas de la matriz suele ser necesario en el cálculo práctico, aun cuando no aparezcan pivotes ceros, para obtener buenas soluciones[2].

3.3.2 Caracterización del proceso de eliminacion gaussiana.

Acabamos de comprobar que si la matriz del sistema $A \cdot x = b$ es regular entonces es siempre posible transformar este sistema en otro equivalente triangular superior.

Recíprocamente, si un sistema se puede triangularizar mediante el procedimiento de Gauss, es porque siempre es posible encontrar pivotes $a_{rr}^{(r)}$ no nulos.

Mas como $|A^{(n)}| = a_{11}^{(1)} \cdot a_{22}^{(2)} \cdot \ldots\ldots \cdot a_{nn}^{(n)}$, tal como (3.1) pone en evidencia; es claro que $A^{(n)}$ es regular.Por otra parte, por (3.2), $A^{(n)} = M \cdot A$, donde M es una matriz regular producto de matrices elementales; luego entonces, A también es regular.

■

Consecuentemente hemos obtenido el siguiente resultado.

[2]Lo más aproximadas posibles a las exactas teóricas, ya que la aritmética computacional no es exacta.

TEOREMA 7 *Un sistema es triangularizable por eliminación de Gauss si y sólo si la matriz A del mismo es regular.*

Mas ¿es posible, a priori, conocer si será o no necesario efectuar permutaciones de filas?

Relación entre los pivotes y los menores principales de A.

La pregunta anterior también se puede formular en estos otros términos: ¿cuándo los pivotes que van apareciendo en la eliminación de Gauss son no nulos?

Si A es regular y se puede triangularizar, por Gauss, sin necesidad de efectuar permutaciones de filas entonces se tendrá que:

$$M \cdot A = M^{(n-1)} \cdot M^{(n-2)} \cdot \ldots \cdot M^{(2)} \cdot M^{(1)} \cdot A = A^{(n)}.$$

Por tanto $A = M^{-1} \cdot A^{(n)}$, donde

$$M^{-1} = \left(M^{(1)}\right)^{-1} \cdot \ldots \cdot \left(M^{(n-1)}\right)^{-1}.$$

Ahora bien

$$M^{(r)} = \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & m_{(r+1)r} & & \\ & & & \vdots & \ddots & \\ & & & m_{nr} & & 1 \end{bmatrix},$$

por tanto

$$\left(M^{(r)}\right)^{-1} = \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & -m_{(r+1)r} & & \\ & & & \vdots & \ddots & \\ & & & -m_{nr} & & 1 \end{bmatrix}.$$

Luego

$$M^{-1} = \left(M^{(1)}\right)^{-1} \cdot \ldots \cdot \left(M^{(n-1)}\right)^{-1}$$

$$= \begin{bmatrix} 1 & & & & & & \\ -m_{21} & 1 & & & & & \\ -m_{31} & -m_{32} & \ddots & & & & \\ \vdots & & & 1 & & & \\ & & & -m_{(r+1)r} & & & \\ \vdots & & & \vdots & \ddots & & \\ -m_{n1} & -m_{n2} & \cdots & -m_{nr} & \cdots & 1 \end{bmatrix},$$

que es una matriz triangular inferior con "unos" en la diagonal, por lo que $|M^{-1}| = 1$.

Podemos considerar A por bloques en la forma

$$A = \left[\begin{array}{c|c} A_{r\times r} & \\ \hline & \end{array}\right] = \left[\begin{array}{c|c} M^{-1}_{r\times r} & \theta \\ \hline & \end{array}\right] \cdot \left[\begin{array}{c|c} A^{(n)}_{r\times r} & \\ \hline \theta & \end{array}\right].$$

Al efectuar el producto de estas matrices particionadas tendremos que

$$A_{r\times r} = M^{-1}_{r\times r} \cdot A^{(n)}_{r\times r} \quad \text{de donde} \quad |A_{r\times r}| = |M^{-1}_{r\times r}| \cdot |A^{(n)}_{r\times r}|.$$

Como los elementos de la la diagonal de $M^{-1}_{r\times r}$ son iguales a 1, $|M^{-1}_{r\times r}| = 1$ y por tanto[3] $|A_{r\times r}| = |A^{(n)}_{r\times r}|$.

Por otra parte, como $A^{(n)}_{r\times r}$ es triangular superior obtenida mediante eliminación de Gauss, su determinante será el producto de su elementos diagonales, es decir

$$|A_{r\times r}| = |A^{(n)}_{r\times r}| = a^{(1)}_{11} \cdot a^{(2)}_{22} \cdot \ldots \cdot a^{(r)}_{rr}. \tag{3.3}$$

Además tenemos que $|A| \neq 0$, siendo

$$|A| = |M^{-1}| \cdot |A^{(n)}| = |A^{(n)}| = a^{(1)}_{11} \cdot a^{(2)}_{22} \cdot \ldots \cdot a^{(r)}_{rr} \cdot \ldots \cdot a^{(n)}_{nn}.$$

Entonces, para $1 \leq i \leq n$, se tiene que

$$a^{(i)}_{ii} \neq 0. \tag{3.4}$$

De (3.3) y (3.4) concluimos que $|A_{r\times r}| \neq 0$, para $1 \leq r \leq n$. Es decir, cada uno de estos determinantes, formado por las primeras r filas y las primeras r columnas de la matriz A, llamado menor principal de orden r de la matriz A, es no nulo.

Recíprocamente, si todos los menores principales de la matriz A, del sistema, son no nulos, entonces el método de Gauss se puede aplicar sin intercambio de filas; o lo que es lo mismo, todos los pivotes que aparecen en la triangularización son no nulos.

En efecto:

- $|A_{1\times 1}| \neq 0$ luego $a_{11} \neq 0$, con lo que ya tenemos que el primer pivote es no nulo.
- Supongamos que $a^{(1)}_{11}, a^{(2)}_{22}, a^{(3)}_{33}, \ldots, a^{(r-1)}_{r-1r-1} \neq 0$. Veamos que entonces $a^{(r)}_{rr} \neq 0$.

Al ser todos estos pivotes no nulos, significa que no fue preciso realizar intercambios de filas para obtener $A^{(r)}$, por lo que podemos poner

$$A^{(r)} = M^{(r-1)} \cdot M^{(n-2)} \cdot \ldots \cdot M^{(2)} \cdot M^{(1)} \cdot A.$$

Llamando $M_{r-1} = M^{(r-1)} \cdot M^{(r-2)} \cdot \ldots \cdot M^{(2)} \cdot M^{(1)}$, tendremos que

[3] Este resultado ya nos asegura que las transformaciones elementales no alteran la nulidad de los menores principales. Es más para las del tipo $F_{ij}(\lambda)$, son invariantes.

$$A = M_{r-1}^{-1} \cdot A^{(r)}.$$

Escribiendo A por bloques $A = \left[\begin{array}{c|c} A_{r\times r} & \\ \hline & \end{array}\right] =$

$$= M_{r-1}^{-1} \cdot A^{(r)} = \left[\begin{array}{c|c} \left(M_{r-1}^{-1}\right)_{r\times r} & \theta \\ \hline & \end{array}\right] \cdot \left[\begin{array}{c|c} A_{r\times r}^{(r)} & \\ \hline \theta & \end{array}\right],$$

por lo que $A_{r\times r} = \left(M_{r-1}^{-1}\right)_{r\times r} \cdot A_{r\times r}^{(r)}$, de donde

$$|A_{r\times r}| = |\left(M_{r-1}^{-1}\right)_{r\times r}| \cdot |A_{r\times r}^{(r)}| = |A_{r\times r}^{(r)}|.$$

Pero

$$|A_{r\times r}^{(r)}| = a_{11}^{(1)} \cdot a_{22}^{(2)} \cdot \cdots \cdot a_{rr}^{(r)} \neq 0$$

y $a_{11}^{(1)}, a_{22}^{(2)}, a_{33}^{(3)}, \ldots, a_{r-1r-1}^{(r-1)}$ son todos no nulos. Luego $a_{rr}^{(r)} \neq 0$.

■

Consecuentemente hemos obtenido el siguiente resultado.

TEOREMA 8 *Un sistema se puede triangularizar por Gauss,* sin *intercambios de filas si y sólo si todos los menores principales de la matriz A del sistema son no nulos.*

Este resultado responde por completo a nuestra pregunta inicial.

Los resultados obtenidos tienen interés desde el punto de vista teórico, pero no son de inmediata utilidad en un caso práctico, toda vez que examinar tal cantidad de determinantes supondría tener que efectuar un elevado número de operaciones y emplear, consecuentemente, mucho tiempo. Más tarde se estudiará como podemos proceder en la práctica ante una matriz de la que desconocemos si es regular o no y por supuesto el carácter de la nulidad de sus distintos menores principales.

3.3.3 Coste en número de operaciones del proceso de eliminación de Gauss.

Supongamos que estamos ante el proceso de eliminación simple de Gauss.

1. Transformación de $\widehat{A}^{(r)}$ en $\widehat{A}^{(r+1)}$:

Si estamos en la fila r y en la columna r en el proceso de eliminación simple de Gauss, tenemos que:

$$\widehat{A} \longrightarrow \widehat{A}^{(r)} = \left[\begin{array}{ccccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & \cdots & \cdots & a_{1n}^{(1)} & b_{(1)}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & \cdots & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{rr}^{(r)} & \cdots & a_{rn}^{(r)} & b_r^{(r)} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nr}^{(r)} & \cdots & a_{nn}^{(r)} & b_n^{(r)} \end{array}\right].$$

Entonces, para pasar al siguiente paso con $\widehat{A}^{(r+1)}$, nos fijamos en las siguientes transformaciones.

1.1.- Transformación de $A^{(r)}$ en $A^{(r+1)}$:

1. Por cada fila subpivotal $f_{r+1}, \ldots, f_n$ hay que realizar una división, para obtener m_{ir}, con $r < i \leq n$, con lo que habrá que realizar $n - r$ divisiones.

2. A cada fila subpivotal se le suma un múltiplo de la fila r para obtener ceros en la columna r debajo de $a_{rr}^{(r)}$. Estos ceros sabemos que van a obtenerse, con lo que se pueden colocar directamente, no siendo necesaria ni la suma ni el producto para los elementos que ocupan las posiciones $(r+1, r), \ldots, (n, r)$, donde el primer número del par indica la fila y el segundo la columna. Por consiguiente, para cada fila subpivotal hay que realizar $n - r$ productos y $n - r$ sumas y como son $n - r$ las filas subpivotales, entonces serán necesarias, en este paso, $(n - r)^2$ sumas y otros tantos productos.

1.2.- Transformación de $b^{(r)}$ en $b^{(r+1)}$:

1. Por cada fila subpivotal hay que realizar una suma y un producto, como hay $n - r$ de estas filas, entonces habrá que realizar $n - r$ productos y otras tantas sumas.

2. Transformación de $\widehat{A}$ en $\widehat{A}^{(n)}$:

2.1.- Transformación de A en $A^{(n)}$:

Teniendo en cuenta el análisis efectuado en 1.1., tendremos que:

- El número de divisiones será $\sum_{r=1}^{n-1} (n - r)$, suma de las de cada fila.
- El número tanto de productos como de sumas será $\sum_{r=1}^{n-1} (n - r)^2$, suma de los que hay que realizar en cada paso.

Por tanto, teniendo en cuenta que la suma[4] de los n primeros cuadrados

$$1^2 + 2^2 + \ldots + n^2 = [n\,(n+1)\,(2n+1)]\,/6,$$

se tendrá:

- Sumas: $\sum_{r=1}^{n-1} (n - r)^2 = \frac{(n-1)\cdot n\cdot(2n-1)}{6} = \frac{2n^3-3n^2+n}{6}$.
- Productos y divisiones:

  $\sum_{r=1}^{n-1} \left[(n - r)^2 + (n - r)\right] = \frac{(n-1)\cdot n\cdot(2n-1)}{6} + \frac{(n-1)n}{2} = \frac{n^3-n}{3}$.

En total el número de operaciones es

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

[4]Toda vez que $1, 4, \ldots, n^2$ se trata de una progresión aritmética de orden tres, la suma de los n primeros términos de la misma se puede obtener como consecuencia de aplicar la fórmula de Newton para diferencias finitas. No obstante, la cuestión es abordable sin necesidad de esta "artilleria", desde una perspectiva heurística; puede verse una bonita demostración de ello, así como la generalización a la suma de las primeras n potencias naturales, de cualquier exponente k, en la pág. 73 de POLYA, G.: *La Decouverte des Mathematiques: les modèles, une méthode générale*. 1967. Paris, Dunod.

2.2.- Transformación de b en $b^{(n)}$:

Teniendo en cuenta el análisis realizado en 1.2., tendremos que en total serán necesarias $2\sum_{r=1}^{n-1}(n-r) = n^2 - n$ operaciones para transformar b en $b^{(n)}$.

De 2.1. y 2.2. se tiene que

El número total de operaciones en la eliminación de Gauss en $\widehat{A}$ *es* $\frac{4n^3+3n^2-7n}{6}$.

Considerando un sistema de orden 10, el número de operaciones realizadas en la eliminación será 705 y 100 las necesarias en el proceso de sustitución hacia atrás; es decir, se necesitan 805 operaciones para resolver el sistema por el método de Gauss, cantidad considerablemente inferior a las $4 \cdot 10^8$ necesarias mediante el método de Cramer.

Si se hubiese tenido que realizar alguna permutación de filas en el proceso de eliminación, como las permutaciones de filas no implican ninguna operación aritmética, es claro que el coste en número de operaciones de la eliminación gaussiana con intercambio de filas es el mismo que el de la eliminación simple.

3.3.4 Almacenamiento en la computadora.

El método de Gauss es fácilmente programable. Como hemos visto se procede en dos pasos: eliminación y sustitución hacia atrás. Sin embargo, desde un punto de vista estrictamente computacional, conviene ocupar la mínima memoria posible.

Con frecuencia se presenta la necesidad de resolver multisistemas; es decir, sistemas con distintos términos independientes, pero con la misma matriz del sistema: $A \cdot x = b_i$.

Sería poco rentable trabajar con la matriz ampliada del sistema en cada caso, puesto que el proceso de triangularización es el mismo en todos ellos, variando $b_i^{(n)}$, pero resultando siempre de las mismas transformaciones elementales que se efectuaron sobre A.

Por ello se procede en dos pasos:

1. triangularización de la matriz A,
2. resolución del sistema.

Pero esto plantea un problema de construcción suplementario: ¿cómo y dónde se almacenan las transformaciones elementales que hay que aplicar a cada vector de términos independientes b_i?

La solución óptima estriba en utilizar la misma matriz A para almacenar estas transformaciones, puesto que en el proceso de triangulación media matriz queda con ceros (vacía).

Ejemplo 1:

Sea el sistema

$$\begin{cases} x + y + z = 6 \\ 2x - y + 3z = 9 \\ x + 2y - z = 2 \end{cases}$$

cuya matriz es $\begin{bmatrix} 1 & 1 & 1 \\ 2 & -1 & 3 \\ 1 & 2 & -1 \end{bmatrix}$.

La primera transformación que es necesario realizar es $F_{21}\,(-a_{21}/a_{11})$, es decir $F_{21}\,(-2)$, quedando

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

Como el elemento $a_{21}^{(2)}$ es cero, con lo que a los efectos prácticos queda inutilizado, almacenamos en su lugar el multiplicador de la transformación elemental, cambiado de signo, para utilizarlo en el futuro con el vector b, resultando:

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

El hecho de que se cambie de signo el multiplicador proviene de que computacionalmente las transformaciones se entienden restando; es decir $F_{21}(2) =$ fila 2– fila 1 por 2.

Ahora se realizaría la transformación elemental de fila $F_{31}\,(-a_{31}/a_{11})$; es decir $F_{31}\,(-1)$, resultando, tras almacenar en el lugar del cero que ocupa $a_{31}^{(2)}$ el multiplicador (-1), cambiado de signo, que lo produce:

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

con estas transformaciones se ha conseguido hacer ceros en la primera columna debajo del primer pivote. Realizando ahora $F_{32}\left(-a_{32}^{(2)}/a_{22}^{(2)}\right)$, es decir $F_{32}\,(1/3)$, y almacenando en el lugar del cero que ocupará el lugar (3,2) el multiplicador cambiado de signo que lo produce, tendremos

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & -3 & 1 \\ 1 & -1/3 & -5/3 \end{bmatrix},$$

que es la matriz que almacena la computadora en el proceso de triangulación.

Para resolver el sistema para la columna término independiente

$$b = \begin{bmatrix} 6 \\ 9 \\ 2 \end{bmatrix}$$

habrá que realizar las mismas transformaciones sobre b, y por tanto:

$$\begin{aligned} b_1^{(3)} &= b_1^{(2)} = b_1^{(1)} = 6, \\ b_2^{(3)} &= b_2^{(2)} = b_2^{(1)} - b_1^{(1)} \cdot 2 = -3, \\ b_3^{(2)} &= b_3^{(1)} - b_1^{(1)} \cdot 1 = -4, \\ b_3^{(3)} &= b_3^{(2)} - b_2^{(2)} \cdot (-1/3) = -5, \end{aligned}$$

que se corresponden con

$$b^{(1)} = \begin{bmatrix} 6 \\ 9 \\ 2 \end{bmatrix}, \qquad b^{(2)} = \begin{bmatrix} 6 \\ -3 \\ -4 \end{bmatrix} \qquad \text{y} \qquad b^{(3)} = \begin{bmatrix} 6 \\ -3 \\ -5 \end{bmatrix}.$$

Luego el sistema queda:

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 0 & -5/3 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 6 \\ -3 \\ -5 \end{bmatrix},$$

cuya solución es ya inmediata:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

Es de notar que en la matriz que almacena la computadora

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & -3 & 1 \\ 1 & -1/3 & -5/3 \end{bmatrix}$$

la parte triangular superior, incluida la diagonal, es la matriz del sistema triangularizado y que la parte triangular inferior

$$\left[\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 1 & -1/3 & 0 & 0 \end{array}\right]$$

coincide[5] con la matriz M^{-1} sin los "1" de la diagonal, del apartado **3.3.2**. □

Insistimos una vez más en la ventaja de proceder de esta forma, pues si necesitáramos resolver un sistema con la misma matriz y distinto vector de términos independientes b', ya la tendríamos triangularizada y por lo tanto el coste en tiempo de resolución se reduce considerablemente.

3.3.5 Inconvenientes del metodo de Gauss

Cero en lugar pivotal.

Ya hemos indicado, al analizar la eliminación de Gauss con intercambio de filas, que un inmediato inconveniente es la aparición de pivotes nulos. Este problema lo hemos solventado efectuando una permutación de filas.

Errores de redondeo.

Puesto que la aritmética computacional no es exacta, el intercambio de ecuaciones en el cálculo práctico suele ser necesario, aún cuando no aparezcan

[5] Esta situación tiene relación con la descomposición LU de la que se hablará más tarde.

pivotes nulos, para obtener buenas soluciones, en el sentido de que sean lo más aproximadas posible a las exactas teóricas.

Ejemplo 2:

Sea el sistema cuya matriz ampliada es

$$\left[\begin{array}{cc|c} 0.01 & 1 & 1 \\ 1 & -1 & 0 \end{array}\right].$$

Efectuamos las operaciones en aritmética computacional en punto flotante de dos dígitos.

Haciendo la transformación $F_{21}\,(-1/0.01)$ tendremos

$$\left[\begin{array}{cc|c} 0.01 & 1 & 1 \\ 0 & -100 & -100 \end{array}\right],$$

ya que al realizar las operaciones en aritmética en punto flotante, si con el símbolo $\prec\ \cdot\ \succ$ expresamos la versión redondeada que la computadora toma del número al que abarca, se verificará que[6]:

$$\begin{array}{l} a_{22}^{(2)} = -1 + 1 \cdot (-1/0.01) = -1 - \prec 100 \succ = \prec -101 \succ = \\ = \prec -0.101 \cdot 10^3 \succ = \prec -0.10 \cdot 10^3 \succ = -100, \end{array}$$

$$\begin{array}{l} b_2^{(2)} = 0 + 1 \cdot (-1/0.01) = - \prec 100 \succ = - \prec 0.100 \cdot 10^3 \succ = \\ = - \prec 0.10 \cdot 10^3 \succ = -100. \end{array}$$

Resolviendo ahora este sistema triangular, el vector solución es:

$$\left[\begin{array}{c} 0 \\ 1 \end{array}\right].$$

Pero la solución exacta es:

$$x_1 = x_2 = 1/1.01 = 0.990099\ldots \simeq 0.99,$$

con lo que la solución más exacta en esta aritmética computacional sería

$$x_1 = x_2 = 0.99,$$

obteniéndose un error muy importante.

Sin embargo, efectuando una permutación de las filas; es decir, la transformación F_{21}, tendremos

$$\left[\begin{array}{cc|c} 1 & -1 & 0 \\ 0.01 & 1 & 1 \end{array}\right].$$

Realizando la transformación $F_{21}\,(-0.01)$, queda

$$\left[\begin{array}{cc|c} 1 & -1 & 0 \\ 0 & 1.0 & 1 \end{array}\right],$$

[6] Hay que observar que al estar trabajando en aritmética en punto flotante de dos dígitos $\prec -0.101 \cdot 10^3 \succ$ es interpretado con sólo dos decimales $\prec 0.10 \cdot 10^3 \succ$.

ya que, en este caso

$$a_{22}^{(2)} = 1 + (-1) \cdot (-0.01) = \prec 1.01 \succ = \prec 0.101 \cdot 10^1 \succ = \\ = \prec 0.10 \cdot 10^1 \succ = 1.0,$$

$b_2^{(2)} = 1$, al no alterarse puesto que hemos multiplicado por 0.

Resolviendo ahora el sistema triangular tenemos obviamente como solución $x_1 = x_2 = 1$, bastante más aproximada y correcta que la obtenida con anterioridad. □

La causa del fracaso del método de Gauss en el primer caso radica en el error de redondeo cometido al transformar 101 en 100, debido a la utilización de un pivote muy pequeño.

Comprobamos cómo pivotes distintos pueden producir respuestas drásticamente diferentes, pudiéndose obtener resultados absolutamente inaceptables.

Aunque el uso de dos cifras significativas en el ejemplo, lo hace absolutamente irreal, el problema de los errores de redondeo sí es muy importante al resolver sistemas grandes, ya que cada resultado depende de todos los anteriores y el error se puede propagar fuera de todo control[7].

Una primera medida para subsanar este problema será realizar intercambio de filas, o más bien esta cuestión pone de relieve la necesidad de realizar dichas permutaciones. El problema ahora es ¿cuáles son las que habrá que realizar?

Precisión sobre el tamaño de los pivotes.

Ejemplo 3:

Consideremos el sistema[8] de matriz ampliada

$$\left[\begin{array}{cc|c} 1.00 \cdot 10^{-4} & 1.00 & 1.00 \\ 1.00 & 1.00 & 2.00 \end{array}\right],$$

cuya solución redondeada a cinco decimales es

$$x_1 = 1.00010 \qquad \text{y} \qquad x_2 = 0.99990.$$

Si se resuelve por el método de Gauss en aritmética computacional de 3 dígitos se tiene

$$a_{21}^{(1)} / a_{11}^{(1)} = 10^4,$$

con lo que

$$\begin{aligned} a_{22}^{(2)} &= \prec 1.00 \succ - \prec 1.00 \cdot 10^4 \succ \\ &= \prec 0.0001 \cdot 10^5 \succ - \prec 0.100 \cdot 10^5 \succ \end{aligned}$$

[7] Es muy difícil predecir en qué medida puede influir el tamaño de un sistema en la relevancia de los errores de redondeo, entre otras razones porque dependerá del tipo de computadora. No obstante en sistemas de 50 o más ecuaciones los errores de redondeo pueden ser suficientemente significativos.

[8] Ejemplo propuesto por Forsyte en 1967, citado por GASCA, 1987, pg. 31.

$$
\begin{aligned}
&= \prec 0.00 \cdot 10^5 \succ - \prec 0.10 \cdot 10^5 \succ \\
&= -1.00 \cdot 10^4, \\
b_2^{(2)} &= \prec 2.00 \succ - \prec 1.00 \cdot 10^4 \succ = -1.00 \cdot 10^4,
\end{aligned}
$$

resultando el sistema triangular

$$
\left[\begin{array}{rr|r} 1.00 \cdot 10^{-4} & 1.00 & 1.00 \\ 0.00 & -1.00 \cdot 10^4 & -1.00 \cdot 10^4 \end{array}\right],
$$

cuya solución obviamente es $x_2 = 1$, $x_1 = 0$. □

Podría pensarse que el grave error cometido se debe a los errores de redondeo producidos al utilizar como pivote un número extremadamente pequeño: 10^{-4}, pero ello no es del todo cierto, como seguidamente veremos.

Ejemplo 4:

Si multiplicamos la primera ecuación del ejemplo anterior por 10^4, nos queda el sistema equivalente

$$
\left[\begin{array}{rr|r} 1.00 & 1.00 \cdot 10^4 & 1.00 \cdot 10^4 \\ 1.00 & 1.00 & 2.00 \end{array}\right];
$$

ahora el pivote es 1 y $a_{21}^{(1)}/a_{11}^{(1)} = 1.00$, resultando

$$
a_{22}^{(2)} = \prec 1.00 - 1.00 \cdot 10^4 \succ = -1.00 \cdot 10^4,
$$

$$
b_2^{(2)} = \prec 2.00 - 1.00 \cdot 10^4 \succ = -1.00 \cdot 10^4,
$$

con lo que nuevamente obtenemos $x_2 = 1$, $x_1 = 0$. □

Como podemos comprobar, no siempre evitando pivotes pequeños se obtiene un resultado adecuado para la solución. Y es que realmente el origen del fracaso no está en el tamaño *per se* del pivote, sino en el desequilibrio del tamaño de los términos de la matriz.

Sistemas mal condicionados.

Ejemplo 5:

Consideremos ahora el siguiente sistema:

$$
\left[\begin{array}{rr|r} 2.000 & 0.6667 & 2.000 \\ 1.000 & 0.3333 & 1.000 \end{array}\right].
$$

Al hacer $F_{21}\,(-0.500)$ obtenemos:

$$
\left[\begin{array}{rr|r} 2.000 & 0.6667 & 2.000 \\ 0.000 & -0.00005 & 0.000 \end{array}\right],
$$

con lo que la solución es $x_1 = 1.000$ y $x_2 = 0.000$ y ello independientemente del número de dígitos utilizados.

Consideremos ahora este otro sistema ligeramente modificado

$$
\left[\begin{array}{rr|r} 2.000 & 0.6666 & 2.000 \\ 1.000 & 0.3333 & 1.000 \end{array}\right].
$$

Las dos ecuaciones son claramente proporcionales, con lo que ahora tiene infinitas soluciones. Si $x_2 = t$ entonces $x_1 = 1.000 - 0.3333 \cdot t$.

La alteración de una diezmilésima en un dato, hace que de un sistema compatible determinado pasemos a otro incompatible. Se trata de lo que se denomina un problema mal condicionado y que más adelante estudiaremos. □

Ejemplo 6:

Consideremos ahora este otro sistema

$$\left[\begin{array}{cc|c} 1 & 2 & 10 \\ 1.1 & 2 & 10.4 \end{array}\right].$$

La solución exacta de este sistema es $x_1 = 4$ y $x_2 = 3$.

Si efectuamos ahora una pequeña perturbación en el elemento (2,1), modificándolo levemente a 1.05, la solución exacta es ahora $x_1 = 8$ y $x_2 = 1$. □

Por tanto, una pequeña variación en un elemento de la matriz produce una enorme variación en el resultado. Podríamos pensar que sustituyendo los valores obtenidos en las ecuaciones originales detectaríamos el problema. Desafortunadamente no es así:

$$\begin{cases} 8 + 2 \cdot 1 = 10 \\ 1.1 \cdot 8 + 2 \cdot 1 = 10.8. \end{cases}$$

Resulta, pues, que siendo $x_1 = 8$ y $x_2 = 1$ valores muy alejados de la solución real, sin embargo no se detecta por la prueba del error, apareciendo como bastante aceptables ($10.08 \simeq 10.4$).

Estamos, en ambos casos ante sistemas mal condicionados, pequeñas alteraciones en los datos producen grandes alteraciones en los resultados.

Los ejemplos propuestos permiten una imagen geométrica del mal condicionamiento, representando las ecuaciones. Debido a que, en ambos casos, las pendientes de las rectas son casi iguales, es difícil detectar el punto de intersección, lo que cuantitativamente se traduce en una amplia región de incertidumbre. Las figuras 3.1 y 3.2 representan geométricamente dos sistemas, uno bien y otro mal condicionado, respectivamente.

Si escribimos las ecuaciones de forma genérica:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2, \end{cases}$$

tendremos que

$$\begin{cases} x_2 = -\frac{a_{11}}{a_{12}}x_1 + \frac{b_1}{a_{12}} \\ x_2 = -\frac{a_{21}}{a_{22}}x_1 + \frac{b_2}{a_{22}}, \end{cases}$$

por lo tanto, si las pendientes son casi idénticas, entonces

$$-\frac{a_{11}}{a_{12}} \simeq -\frac{a_{21}}{a_{22}}$$

FIGURA 3.1: Sistema bien condicionado

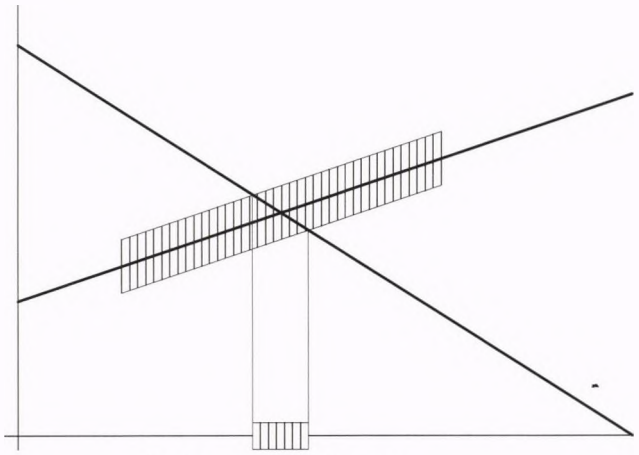
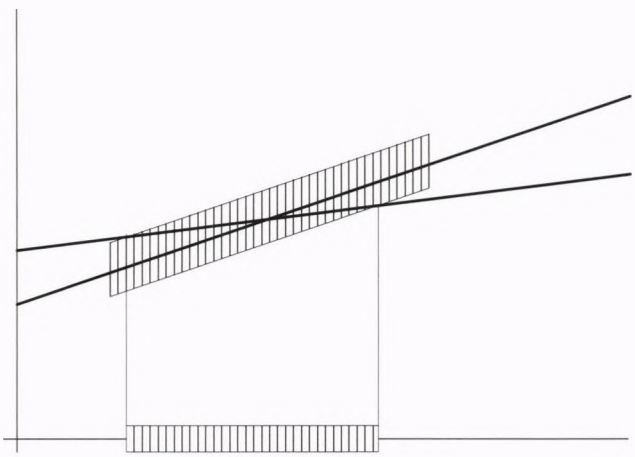


FIGURA 3.2: Sistema mal condicionado



y de ese modo $a_{11}a_{22} - a_{21}a_{12} \simeq 0$, o lo que es lo mismo, su determinante es casi nulo.

Así que podemos establecer una primera idea intuitiva de sistemas mal condicionados como aquellos cuyo determinante es casi nulo.

El determinante de la matriz del sistema no es cero, pero es muy pequeño (matriz casi-singular). Desde una perspectiva teórica hay una gran diferencia entre un sistema con determinante nulo y otro no nulo. Desde el punto de vista computacional, una matriz casi-singular puede arrastrarnos a cálculos desastrosos.

¿Cuál ha de ser la proximidad a cero para considerar que estamos ante un problema mal condicionado? La respuesta no es sencilla, e incluso la pregunta carece de sentido, desde la óptica intuitiva de que este carácter es medible por el determinante de la matriz del sistema, si se tiene en cuenta que si en un sistema multiplicamos una o más de una ecuación por un escalar el resultado es otro equivalente, pero sin embargo el valor del determinante queda alterado por completo.

Ejemplo 7:

Si la segunda ecuación del sistema perturbado del ejemplo anterior la multiplicamos por 10:

$$\left[\begin{array}{cc|c} 1 & 2 & 10 \\ 10.5 & 20 & 10.4 \end{array}\right],$$

el determinante de la matriz del sistema es ahora menos uno y sin embargo la solución exacta sigue siendo $x_1 = 8$ y $x_2 = 1$.

Análogamente si se hubiese multiplicado por 100, el determinante de la matriz del sistema sería ahora -10, y sin embargo la solución exacta seguiría siendo: $x_1 = 8$ y $x_2 = 1$.

Así, pues, la pequeña perturbación inicial del sistema en el elemento (2,1), modificándolo en cinco centésimas, produce una enorme variación en el resultado, y este hecho no es imputable al valor del determinante de la matriz del sistema, como acabamos de comprobar. □

Dado el carácter relativo del determinante es claro que deberemos desviar nuestro punto de mira para buscar una caracterización de las matrices mal condicionadas.

No obstante, tenemos una primera aproximación "bastante gráfica" al problema.

Este concepto se precisará más adelante, en el epígrafe "Error y condicionamiento". Por el momento nos es suficiente constatar la necesidad de detectar este tipo de matrices intrínsecamente "problemáticas".

3.3.6 Modificaciones en el metodo de Gauss.

Analicemos algunas técnicas que se pueden incorporar al algoritmo de eliminación gaussiana simple, para evitar algunos de los problemas reseñados.

Ciertamente, en algunos casos simples resulta suficiente intercambiar filas, como en el caso de que aparezca un pivote cero o bien para evitar ciertos errores de redondeo. El aumento del número de cifras significativas conduce también, por lo general, a una mayor precisión y a una disminución de la propagación de los errores de redondeo.

Pero otros problemas requieren un tratamiento más complejo[9] y la mejor opción a priori será la combinación de varias de las técnicas que pasamos a exponer.

Pivoteo parcial.

Hemos indicado en el epígrafe anterior cómo surgen soluciones absolutamente desafortunadas debido a los errores de redondeo, que aparecen y se propagan cuando la magnitud del pivote es pequeño en comparación con los otros elementos de la matriz.

Por ello, es conveniente determinar el coeficiente de mayor valor absoluto posible y proceder a una permutación de línea, de manera que éste sea el pivote.

Si este número se busca dentro de la misma columna, entre los coeficientes de las incógnitas x_i de subíndice mayor de la que se quiere eliminar, estamos ante la estrategia de pivoteo parcial.

Por tanto, se procede como en la eliminación de Gauss "simple" y al pasar de $A^{(r)} \cdot x = b^{(r)}$ a $A^{(r+1)} \cdot x = b^{(r+1)}$ se efectua la transformación fila F_{rt}, tal que

$$c_r = \left|a_{tr}^{(r)}\right| = \max_{r \leq k \leq n} \left|a_{kr}^{(r)}\right|. \tag{3.5}$$

Insistimos en el hecho de que llevar el número "más grande de la columna" a la posición diagonal tiene la ventaja de reducir el error de redondeo.

Ejemplo 8:

Si observamos el ejemplo 3, propuesto por Forsyte, del epígrafe 3.3.5, veremos que el resultado que se obtiene utilizando la estrategia de pivoteo parcial es notablemente próximo a la solución exacta.

En

$$\left[\begin{array}{cc|c} 1.00 \cdot 10^{-4} & 1.00 & 1.00 \\ 1.00 & 1.00 & 2.00 \end{array}\right]$$

efectuamos F_{12}, quedando

$$\left[\begin{array}{cc|c} 1.00 & 1.00 & 2.00 \\ 1.00 \cdot 10^{-4} & 1.00 & 1.00 \end{array}\right]$$

y luego $F_{21}\,(-10^{-4})$, resultando

$$\left[\begin{array}{cc|c} 1.00 & 1.00 & 2.00 \\ 0 & 1.00 & 1.00 \end{array}\right],$$

[9] El problema de la propagación del error en el proceso de eliminación gaussiana fué estudiado por completo y en profundidad por J.H. Wilkinson en 1965 (en "The Algebraic Eigenvalue Problem". Oxford, England U.K., Oxford University Press).

ya que en aritmética en punto flotante de dos dígitos se tiene

$$\begin{aligned} a_{22}^{(2)} &= \prec 1.00 \succ - \prec 1.00 \cdot 10^{-4} \succ \\ &= \prec 1.00 \succ - \prec 0.0001 \succ = \prec 1.00 \succ - \prec 0.00 \succ = 1.00, \\ b_2^{(2)} &= \prec 1.00 \succ - \prec 2.00 \cdot 10^{-4} \succ \\ &= \prec 1.00 \succ - \prec 0.0002 \succ = \prec 1.00 \succ - \prec 0.00 \succ = 1.00. \end{aligned}$$

La solución es ahora $x_2 = 1.00$ y $x_1 = 1.00$, con un error de una diezmilésima respecto del valor de la solución redondeada a cinco decimales, que es

$$x_1 = 1.00010 \quad \text{y} \quad x_2 = 0.99990.$$

□

Pivoteo total.

Desde luego para evitar los errores de redondeo producto de efectuar divisiones por números muy pequeños, así como superar la dificultad que supone encontrar pivotes nulos, puede buscarse aún un mejor pivote que el que proporciona la estrategia anterior.

Una más óptima selección de los pivotes se realiza con el pivoteo total, consistente en colocar, sucesivamente, mediante permutaciones de fila o columna, como pivote al elemento de mayor valor absoluto de la submatriz, que queda tras eliminar las filas y columnas donde se encuentren los pivotes anteriores.

Así, si nos encontramos en el r-ésimo paso tendremos que

$$\widehat{A}^{(r)} = \left[\begin{array}{ccccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & \cdots & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & \cdots & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{rr}^{(r)} & \cdots & a_{rn}^{(r)} & b_r^{(r)} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nr}^{(r)} & \cdots & a_{nn}^{(r)} & b_n^{(r)} \end{array}\right]$$

y el pivote $a_{rr}^{(r)}$ será el elemento de mayor valor absoluto de la submatriz

$$A^{*(r)} = \left[\begin{array}{ccc} a_{rr}^{(r)} & \cdots & a_{rn}^{(r)} \\ \vdots & \ddots & \vdots \\ a_{nr}^{(r)} & \cdots & a_{nn}^{(r)} \end{array}\right].$$

Por tanto, al pasar de $A^{(r)} \cdot x = b^{(r)}$ a $A^{(r+1)} \cdot x = b^{(r+1)}$ se efectuan las transformaciones fila F_{rt} y columna C_{rs}, tal que

$$|a_{ts}^{(r)}| = \max_{r \le k \le n} |a_{t_k s_k}^{(r)}|.$$

Aunque no es fácil contestar a la pregunta "¿cómo de bien controla el pivoteo la magnificación del error?"; sin embargo, en la práctica, el pivoteo total es raramente utilizado, debido a que:

1. La experiencia indica que la probabilidad de que ocurran problemas graves usando pivoteo parcial es muy pequeña. Además, es inusual que en el pivoteo parcial la transmisión de un error cometido a la solución, se magnifique más de dos veces respecto de lo que ocurriría utilizando pivoteo total.

2. Para la determinación, en cada caso, del pivote mediante pivoteo parcial hay que examinar $n-r+1$ coeficientes, mientras que en el pivoteo total se requiere $(n-r+1)^2$. El pivoteo total aproximadamente dobla el coste de la eliminación gaussiana, mientras que el coste del pivoteo parcial es casi inapreciable.

Por ello, el precio del pivoteo total es muy alto, para protegerse de situaciones muy infrecuentes; tanto, que a los analistas numéricos les ha costado años encontrar un ejemplo de sistema donde la magnificación del error transmitido, mediante pivoteo parcial, fuera desproporcionadamente grande.

Realmente es innecesario encontrar el mejor pivote, sólo es necesario evitar pivotes pésimos.

A pesar de ello, el pivoteo parcial no resuelve siempre todas las dificultades que en la práctica se pueden presentar.

Si los resultados siguen sin ser precisos se puede recurrir a la doble precisión, cuyo inconveniente es el considerable aumento del tiempo de cómputo.

Otra alternativa radica en el llamado escalamiento de la matriz del sistema.

Escalamiento.

Se ha observado empíricamente que cuando los elementos de la matriz varían considerablemente de tamaño, cosa que suele ocurrir en problemas de ingeniería en los que aparecen sistemas donde quedan involucradas magnitudes diferentes, los errores de redondeo son muy importantes. Para evitar este problema se utiliza el escalamiento de la matriz A, con lo que se consigue que el tamaño de los elementos de la misma varíen menos entre sí.

Se pretende, pues, que tanto por filas como por columnas los elementos de la matriz del sistema equivalente sean de tamaños muy parecidos.

La base del escalamiento consiste en que el sistema no se altera si se multiplican sus ecuaciones por números cualesquiera. La idea es encontrar unos números tales que el sistema equivalente resultante tenga una matriz con elementos muy parecidos en tamaño.

A veces este escalamiento por filas no es suficiente, siendo necesario escalar las columnas, multiplicando éstas por constantes adecuadas; en cuyo caso el sistema no será equivalente, pero se obtendrá la solución en términos de incógnitas proporcionales a las iniciales ($\lambda_i x_i$ tal que λ_i =cte.).

Así, si B denota el resultado del escalamiento de las filas y columnas de A, entonces:

$$B = D_1 \cdot A \cdot D_2,$$

donde D_1 y D_2 son matrices diagonales con los factores escalantes.

Como $A \cdot x = b$, tenemos que

$$D_1 \cdot A \cdot D_2 \cdot D_2^{-1} \cdot x = D_1 \cdot b,$$

entonces haciendo

$$B \cdot z = D_1 \cdot b \qquad \text{y} \qquad x = D_2 \cdot z \tag{3.6}$$

podemosresolver el sistema.

Este escalamiento es actualmente objeto de investigación. No se sabe bien cómo garantiza que el error de redondeo disminuya mediante tal escalamiento.

No existe, a priori, una estrategia para la selección de los factores escalantes que permita siempre hacer decrecer el efecto de la propagación del error de redondeo, basada únicamente en el conocimiento de A y b.

Lo más frecuente es que el escalamiento se restringa al de las filas de la matriz A.

Usualmente se intenta escoger los factores escalantes de forma que en la nueva matriz escalada $B = (b_{ij})$ el elemento de mayor valor absoluto de cada fila sea 1:

$$\max_{1 \le j \le n} |b_{ij}| = 1, \qquad i = 1, \ldots, n.$$

La forma más fácil es definir

$$s_i = \max_{1 \le j \le n} |a_{ij}|, \qquad \text{para} \quad i = 1, \ldots, n \tag{3.7}$$

y dividir cada elemento de la matriz A por el correspondiente s_i

$$b_{ij} = \frac{a_{ij}}{s_i}, \qquad \text{para } j = 1, \ldots\ldots, n.$$

Ejemplo 9:

Consideremos el sistema de matriz ampliada

$$\left[\begin{array}{cc|c} 2 & 100000 & 100000 \\ 1 & 1 & 2 \end{array}\right],$$

trabajando en aritmética de punto flotante con tres cifras significativas.

Si escalamos la matriz de forma que en cada ecuación el coeficiente máximo sea 1, tendremos

$$\left[\begin{array}{cc|c} 0.00002 & 1 & 1 \\ 1 & 1 & 2 \end{array}\right].$$

Ahora efectuamos la eliminación gaussiana con pivoteo parcial, y para ello realizaremos F_{12}:

$$\left[\begin{array}{cc|c} 1 & 1 & 2 \\ 0.00002 & 1 & 1 \end{array}\right]$$

y ahora $F_{21}\,(-0.00002)$:

$$\left[\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 0.999 & 0.999 \end{array}\right],$$

ya que

$$\begin{aligned} a_{22}^{(2)} &= \prec 1.000 \succ - \prec 0.00002 \succ \\ &= \prec 1.000 \succ - \prec 0.200 \cdot 10^{-4} \succ \\ &= \prec 10000 \cdot 10^{-4} \succ - \prec 0.200 \cdot 10^{-4} \succ = 0.999, \\ b_2^{(2)} &= \prec 1.00 \succ - \prec 2 \cdot 0.00002 \succ \\ &= \prec 1.00 \succ - \prec 2 \cdot 0.002 \cdot 10^{-2} \succ = 0.999, \end{aligned}$$

con lo que

$$x_1 = 1.00 \qquad \text{y} \qquad x_2 = 1.00,$$

obteniéndose una muy buena aproximación.

La respuesta correcta es

$$x_1 = 1.00002 \qquad \text{y} \qquad x_2 = 0.99998.$$

Si se hubiera resuelto sin escalamiento, pero sí con pivoteo parcial se habría obtenido

$$\left[\begin{array}{cc|c} 2 & 100000 & 100000 \\ 0 & -50000 & -50000 \end{array}\right],$$

con lo que

$$x_1 = 0.000 \qquad \text{y} \qquad x_2 = 1.00,$$

solución que obviamente es absolutamente inadecuada. □

Pero, normalmente, debido a la referida división de cada elemento de A por el correspondiente s_i se introduce un error de redondeo adicional en cada elemento de la nueva matriz. Por ello otras dos técnicas, menos inmediatas que las anteriores divisiones, son más ampliamente usadas. Pasamos segidamente a estudiarlas.

1. Escalamiento usando la base del computador:

   Sea β la base usada en la aritmética computacional, por ejemplo $\beta = 2$ en las máquinas binarias.

   Sea r_i el más pequeño entero para el que $\beta^{r_i} \geq s_i$, donde s_i viene determinado según se indicó en (3.7). Se define entonces la matriz escalada B mediante

   $$b_{ij} = \frac{a_{ij}}{\beta^{r_i}}, \qquad i, j = 1, \ldots\ldots, n, \tag{3.8}$$

   con lo que se consigue que en la definición de los b_{ij} no se vean involucrados errores de redondeo, sólo un cambio en el exponente en la forma punto flotante para a_{ij}, al estar dividiendo por potencias enteras de la base. Y además los elementos de B satisfacen

   $$\beta^{-1} < \max_{1 \leq j \leq n} |b_{ij}| \leq 1,$$

con lo que nos aproximamos suficientemente bien a

$$\max_{1\leq j\leq n} |b_{ij}| = 1, \qquad i = 1, \ldots, n.$$

2. Escalamiento implícito:

   El uso del escalamiento, por lo general cambiará la elección de los pivotes, cuando se utiliza la estrategia de pivoteo parcial en la eliminación gaussiana. Y sólo con tal cambio de pivotes los resultados de la eliminación gaussiana cambiarán.

   F. Bauer[10] demostró que no es preciso multiplicar previamente por un factor escalante, ya que en la práctica bastará elegir los pivotes de acuerdo con cierta estrategia, de manera que el único significado del escalamiento radique en la elección de los pivotes. Esto es lo que se conoce como escalamiento implícito.

   Para el escalamiento implícito, continuaremos usando la matriz A. Pero elegimos el pivote en el paso r-ésimo del algoritmo de la eliminación gaussiana mediante

$$c_r = \max_{r\leq i\leq n} \frac{|a_{ir}^{(r)}|}{s_i}, \tag{3.9}$$

   reemplazando la definición dada para el pivote en (3.5), usada para el pivoteo parcial. De forma que si c_r se obtiene para una fila $i\,(i \geq r)$ con $i \neq r$, entonces se intercambian las filas i y r, y proseguimos ordinariamente con el proceso de eliminación gaussiana. Y donde s_i viene dado por la expresión (3.7). Esta forma de escalamiento parece ser la forma más comunmente utilizada en los algoritmos que se publican en la actualidad.

Ejemplo 10:

Consideremos el sistema de matriz ampliada

$$\left[\begin{array}{rrr|r} 2.222 & 16.71 & 9.612 & 28.544 \\ 1.5611 & 5.1791 & 1.6852 & 8.4254 \\ 3.333 & 15920 & 10.333 & 15913 \end{array}\right],$$

trabajando seis cifras significativas.

El proceso es el siguiente:

$$\left[\begin{array}{rrr|r} 1.5611 & 5.1791 & 1.6852 & 8.4254 \\ 2.222 & 16.71 & 9.612 & 28.544 \\ 3.333 & 15920 & 10.333 & 15913 \end{array}\right],$$

$$\left[\begin{array}{rrr|r} 1.5611 & 5.1791 & 1.6852 & 8.4254 \\ 0 & 9.3383 & 7.21336 & 16.5516 \\ 0 & 15909 & -13.9309 & 15895 \end{array}\right],$$

[10] En FORSYTHE & MOLER 1967. Computer Solution of Linear Algebraic Systems. Englewood Cliffs, N.J. U.S.A., Prentice-Hall. Citado por Akitson, 1989, pg. 519.

$$\left[\begin{array}{rrr|r} 1.5611 & 5.1791 & 1.6852 & 8.4254 \\ 0 & 9.3383 & 7.21336 & 16.5516 \\ 0 & 0 & -12303 & -12303 \end{array}\right].$$

Por lo que la solución es

$$x_1 = 1.0000, \qquad x_2 = 1.0000, \qquad x_3 = 1.0000,$$

que coincide con la solución exacta.

Sin embargo, veamos que realizando simplemente Gauss, incluso utilizando pivoteo parcial, el resultado no es tan bueno.

Así, mediante eliminación simple de Gauss, se tendría:

$$\left[\begin{array}{rrr|r} 2.222 & 16.71 & 9.612 & 28.544 \\ 0 & -6.56076 & -5.06785 & -11.6286 \\ 0 & 15895 & -24.751 & 15870 \end{array}\right],$$

$$\left[\begin{array}{rrr|r} 2.222 & 16.71 & 9.612 & 28.544 \\ 0 & -6.56076 & -5.06785 & -11.6286 \\ 0 & 0 & -12302.7 & -12303 \end{array}\right]$$

y la solución sería

$$x_1 = 1.00002, \qquad x_2 = 0.999984, \qquad x_3 = 1.00002.$$

Si ahora utilizamos pivoteo parcial el resultado es:

$$\left[\begin{array}{rrr|r} 3.333 & 15920 & 10.333 & 15913 \\ 1.5611 & 5.1791 & 1.6852 & 8.4254 \\ 2.222 & 16.71 & 9.612 & 28.544 \end{array}\right],$$

$$\left[\begin{array}{rrr|r} 3.333 & 15920 & 10.333 & 15913 \\ 0 & -7451.4 & 6.52493 & -7444.85 \\ 0 & -10596.6 & 16.5006 & -10580 \end{array}\right],$$

$$\left[\begin{array}{rrr|r} 3.333 & 15920 & 10.333 & 15913 \\ 0 & -10596.6 & 16.5006 & -1058085 \\ 0 & -7451.4 & 6.52493 & -7444.85 \end{array}\right],$$

$$\left[\begin{array}{rrr|r} 3.333 & 15920 & 10.333 & 15913 \\ 0 & -10596.6 & 16.5006 & -1058085 \\ 0 & 0 & -5.07812 & -5.14285 \end{array}\right]$$

y la solución sería

$$x_1 = 1.01274, \qquad x_2 = 1, \qquad x_3 = 1.05010.$$

□

En el siguiente capítulo estudiaremos dos algoritmos en los que se incorporan las técnicas de pivoteo parcial y escalamiento implícito, para la resolución de $A \cdot x = b$.

Corrección de errores:

Desgraciadamente en algunas ocasiones la conjunción del pivoteo parcial y el escalamiento no son suficientes para garantizar resultados adecuados. Estos errores se pueden corregir con el siguiente procedimiento.

Consideremos el sistema de matriz ampliada

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right].$$

Supongamos que tiene una solución aproximada $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$. Al sustituirlos en el sistema original obtendremos realmente:

$$(b_1^*, b_2^*, \ldots, b_n^*).$$

Además se tiene que

$$x_1 = x_1^* + \Delta x_1, x_2 = x_2^* + \Delta x_2, \ldots, x_n = x_n^* + \Delta x_n, \tag{3.10}$$

donde los Δx_i representan los factores de corrección que se buscan.

Si se sustituyen estas expresiones en el sistema inicial tendremos

$$\begin{array}{rcl} a_{11}(x_1^* + \Delta x_1) + a_{12}(x_2^* + \Delta x_2) + \cdots + a_{1n}(x_n^* + \Delta x_n) & = & b_1, \\ a_{21}(x_1^* + \Delta x_1) + a_{22}(x_2^* + \Delta x_2) + \cdots + a_{2n}(x_n^* + \Delta x_n) & = & b_2, \\ \vdots & \vdots & \vdots \\ a_{n1}(x_1^* + \Delta x_1) + a_{n2}(x_2^* + \Delta x_2) + \cdots + a_{nn}(x_n^* + \Delta x_n) & = & b_n \end{array}$$

y teniendo en cuenta que $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ es una solución aproximada, quedará

$$\begin{array}{rcl} a_{11}\Delta x_1 + a_{12}\Delta x_2 + \cdots + a_{1n}\Delta x_n = b_1 - b_1^* & = & e_1, \\ a_{21}\Delta x_1 + a_{22}\Delta x_2 + \cdots + a_{2n}\Delta x_n = b_2 - b_2^* & = & e_2, \\ \vdots & \vdots & \vdots \\ a_{n1}\Delta x_1 + a_{n2}\Delta x_2 + \cdots + a_{nn}\Delta x_n = b_n - b_n^* & = & e_n, \end{array} \tag{3.11}$$

que es un sistema de ecuaciones lineales cuya solución son los factores de corrección, lo que nos permitirá obtener, tras su sustitución en (3.10), una mejor aproximación de la solución. Pudiéndose repetir el proceso, con mejoraciones sucesivas de la solución.

Se pueden incorporar las ecuaciones del error en los programas de la eliminación gaussiana, para lo que deberán contemplarse los siguientes pasos en el diseño del algoritmo:

1. Introducir los datos.
2. Resolver las ecuaciones mediante el método de Gauss (con estrategias adicionales si se desea, como pivoteo parcial, por ejemplo).
3. Sustituir los resultados en las ecuaciones originales para calcular el valor de los términos independientes.

4. Evaluar la diferencia entre los términos independientes calculados y los valores originales. Si la diferencia es cero, se ha obtenido la solución exacta, caso de que el sistema esté bien condicionado. Si existe una diferencia ir a 5.

5. Resolver las ecuaciones mediante el método de Gauss usando las diferencias calculadas en 4. como vector de los términos independientes.

6. Agregar estos resultados a la solución anterior y volver a 3. Es decir, tras la aplicación del método de Gauss se obtiene una primera solución aproximada $x^{*(1)}$; y a partir de ella se van obteniendo sucesivas mejores aproximaciones de la solución en la forma siguiente:

   (a) $A \cdot x^{*(n)} = b^{*(n)}$,

   (b) $b - b^{*(n)} = \epsilon^{(n)}$,

   (c) $A \cdot \Delta x^{*(n)} = \epsilon^{(n)}$,

   (d) $x^{*(n+1)} = x^{*(n)} + \Delta x^{*(n)}$.

3.3.7 Método de Gauss para cualquier tipo de sistema.

Desde un principio hemos considerado que trabajamos con un sistema cuya matriz es regular, pero esto en la práctica no siempre es posible. No obstante el método de Gauss sigue siendo válido, procediéndose exactamente igual que como hasta ahora se ha hecho.

La nulidad del determinante de la matriz del sistema se detectará cuando en uno de los pasos de la triangularización sea imposible encontar un pivote no nulo. Y en este punto, la matriz del sistema equivalente que se tenga indica claramente por su estructura el carácter del sistema. A tal fin, llamaremos primera incógnita en una ecuación a la primera incógnita, con coeficiente no nulo, leyendo de izquierda a derecha, en esa ecuación; y denominaremos primera columna para una fila a la columna que contiene el primer elemento no nulo en esa fila, leyendo de izquierda a derecha. Resultará entonces que:

1. Si la última columna (la de los términos independientes) es una primera columna para cierta fila, entonces no hay solución y recíprocamente.

2. Si la última columna no es una primera columna para alguna fila, entonces:

   2.1 Existe solución única si y sólo si cada incógnita es una primera incógnita para alguna fila.

   2.2 Existen infinitas soluciones si y sólo si existen incógnitas que no son primeras incógnitas.

Recordemos que las incógnitas que son primeras incógnitas en alguna fila, en el epígrafe **2.5** las denominamos incógnitas básicas, y que las que no son primeras incógnitas en ninguna fila las llamamos incógnitas libres. Como sabemos, las incógnitas básicas están determinadas en términos de los valores asignados a las libres.

3.4 Variante del método de Gauss. Método de Gauss-Jordan.

Esta variante del método de Gauss que pasamos a comentar es útil desde una perspectiva teórica, pero ineficiente para cálculos prácticos debido al considerable aumento del número de operaciones que se requieren, así como a las necesidades de almacenamiento en el computador.

En el proceso de eliminación de Gauss, en cada fila se han hecho ceros por debajo del pivote. Si hacemos también ceros por encima de él, nos encontramos con el proceso de eliminación de Gauss-Jordan.

Mediante el método de Gauss-Jordan se transforma la matriz del sistema en una matriz identidad, de manera que el sistema inicial se transforma en otro equivalente, de resolución trivial.

Básicamente el proceso es el seguidamente se indica.

1. Para la eliminación en la primera columna:

   1.1 Se elige un primer pivote $a_{11}^{(1)} \neq 0$, efectuando intercambios de filas si fuera preciso.

   1.2 Se realizan las transformaciones:

   1.2.1 $F_1\left(\frac{1}{a_{11}^{(1)}}\right)$.

   1.2.2 $F_{i1}\left(-a_{i1}^{(1)}\right)$, con $i = 2, \ldots, n$.

2. En general:

   2.1 Se elige el pivote de la j-ésima fila $a_{jj}^{(j)} \neq 0$, efectuando intercambios de filas F_{kj}, con $k > j$, si fuera preciso.

   2.2 Se realizan las transformaciones:

   2.2.1 $F_j\left(\frac{1}{a_{jj}^{(j)}}\right)$.

   2.2.2 $F_{ij}\left(-a_{ij}^{(j)}\right)$, con $i, j = 1, \ldots, n$ e $i \neq j$.

Con ello se consigue transformar la matriz ampliada del sistema inicial en la de otro equivalente de la forma

$$\widehat{A}^{(n+1)} = \left[\begin{array}{cccc|c} 1 & 0 & \cdots & 0 & b_1^{(n+1)} \\ 0 & 1 & \cdots & 0 & b_2^{(n+1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & b_n^{(n+1)} \end{array}\right],$$

cuya solución es obvia.

El grave inconveniente de este método radica en que el coste de operaciones elementales es considerablemente mayor. Sin embargo, podemos dar un tratamiento alternativo a esta cuestión, de la siguiente manera:

1. Para la eliminación en la primera columna:

1.1 Se elige un primer pivote $a_{11}^{(1)} \neq 0$, efectuando intercambios de filas si fuera preciso.

1.2 Se realizan las transformaciones:

$$F_{i1}\left(-a_{i1}/^{(1)}a_{11}^{(1)}\right), \text{ con } i = 2, \ldots, n.$$

2. En general:

2.1 Se elige el pivote de la j-ésima fila $a_{jj}^{(j)} \neq 0$, efectuando intercambios de filas F_{kj}, con $k > j$, si fuera preciso.

2.2 Se realizan las transformaciones:

$$F_{ij}\left(-a_{ij}^{(j)}/a_{jj}^{(j)}\right), \text{ con } i > j.$$

Con ello se ha triangularizado el sistema inicial, de forma que en la diagonal de la nueva matriz del sistema sólo hay "1". Básicamente es el método de Gauss.

3. Diagonalización de la matriz triangular:

3.1 Se toma la última fila y se realizan las transformaciones:

$$F_{in}\left(-a_{in}^{(n)}/a_{nn}^{(n)}\right), \text{ con } i = n-1, \ldots, 1.$$

3.2 Se van tomando sucesivamente las restantes filas, desde la $n-1$ hasta la 2, y se realizan las transformaciones:

$$F_{ij}\left(a_{ij}^{(n)}/a_{jj}^{(n)}\right), \text{ con } i = j-1, \ldots, 1 \text{ y } j = n-1, \ldots, 1.$$

Este procedimiento transforma el sistema inicial en el mismo obtenido mediante Gauss-Jordan y el coste en número de operaciones es el mismo que el que se requiere para la eliminación de Gauss y la resolución por sustitución hacia atrás.

4. Factorización en matrices triangulares

Si en la resolución de un sistema $A \cdot x = b$ se conociera una descomposición de A en el producto de otras dos matrices $A = L \cdot U$, con L triangular inferior y U triangular superior, la resolución del sistema se reduciría a la de dos triangulares:

$$L \cdot U \cdot x = b \Longleftrightarrow \begin{cases} L \cdot y = b, \\ U \cdot x = y. \end{cases}$$

Cuando se deben resolver varios sistemas de ecuaciones lineales, todos con la misma matriz A, aunque con distintos términos independientes, la resolución mediante esta descomposición tiende a ser más eficiente que la eliminación de Gauss.

Examinar cuándo es posible este tipo de descomposición, cómo construir su algoritmo y el coste en número de operaciones, es entre otras cuestiones el propósito de este epígrafe.

4.1 Factorización a partir de las transformaciones de Gauss.

Realmente el método de Gauss proporciona una descomposición de estas características.

Examinemos por separado los casos en que no se requiere efectuar permutaciones de fila y en los que sí es imprescindible.

4.1.1 Descomposicion LU sin intercambios.

Supongamos que hemos resuelto un sistema de matriz regular y además que todos los menores principales sucesivos de A son no nulos. Esto nos garantiza, como ya hemos visto, que el método de Gauss se puede culminar sin necesidad de efectuar permutaciones F_{ij} de filas; o lo que es lo mismo, que todos los pivotes que van apareciendo son no nulos.

Al estudiar este caso vimos cómo el sistema $A \cdot x = b$ se transformaba en otro equivalente triangular superior con matriz $A^{(n)}$ y verificándose las siguientes relaciones:

$$M \cdot A = A^{(n)},$$

$$M \cdot A \cdot x = M \cdot b, \qquad \text{de donde} \qquad A^{(n)} \cdot x = M \cdot b,$$

con $M = M^{(n-1)} \cdot M^{(n-2)} \cdot \ldots \cdot M^{(2)} \cdot M^{(1)}$, siendo

$$M^{(r)} = \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & m_{(r+1)r} & & \\ & & & \vdots & \ddots & \\ & & & m_{nr} & & 1 \end{bmatrix},$$

donde $m_{ir} = -a_{ir}^{(r)}/a_{rr}^{(r)}$, $i = r+1, \ldots, n$, y

$$\left(M^{(r)}\right)^{-1} = \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & -m_{(r+1)r} & & \\ & & & \vdots & \ddots & \\ & & & -m_{nr} & & 1 \end{bmatrix},$$

y, por tanto,

$$M^{-1} = \begin{bmatrix} 1 & & & & & & \\ -m_{21} & 1 & & & & & \\ -m_{31} & -m_{32} & \ddots & & & & \\ \vdots & & & 1 & & & \\ & & & -m_{(r+1)r} & & & \\ \vdots & & & \vdots & \ddots & & \\ -m_{n1} & -m_{n2} & \cdots & -m_{nr} & \cdots & 1 \end{bmatrix}$$

y siendo

$$A^{(n)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & \cdots & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & \cdots & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{rr}^{(r)} & \cdots & a_{rn}^{(r)} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \cdots & a_{nn}^{(n)} \end{bmatrix}.$$

Por consiguiente $M^{-1} \cdot A^{(n)} \cdot x = M^{-1} \cdot M \cdot b = b$. Luego $A = M^{-1} \cdot A^{(n)}$.

Así, pues, el método de Gauss sin permutación de filas proporciona a la vez M^{-1} (L en nuestra notación actual) y $A^{(n)}$ (U con la nomenclatura de ahora).

Es de notar que L tiene "1" en la diagonal y los elementos subdiagonales son los respectivos multiplicadores m_{ij}, cambiados de signo, utilizados en el proceso de eliminación de Gauss. Así mismo, U tiene en la diagonal los pivotes.

Si se realiza la descomposición por este método no es preciso mantener en memoria más que n^2 elementos, ya que los ceros que se van haciendo se pueden ocupar por los $-m_{ij}$, como ya se puso de manifiesto en el apartado **3.3.4** sobre almacenamiento en la computadora.

Ejemplo 11:

Sea

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 2 & -1 & 3 & 9 \\ 1 & 2 & -1 & 2 \end{array}\right],$$

que tras las transformaciones $F_{21}(-2), F_{31}(-1), F_{32}(1/3)$, queda:

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 0 & -3 & 1 & -3 \\ 0 & 0 & -5/3 & -5 \end{array}\right].$$

Lo que es equivalente a haber escrito $L \cdot U \cdot x = b$ con

$$U = \begin{bmatrix} 1 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 0 & -5/3 \end{bmatrix}$$

y

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1/3 & 1 \end{bmatrix}.$$

Resolviendo $L \cdot y = b$:

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1/3 & 1 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 9 \\ 2 \end{bmatrix},$$

$$y_1 = 6,\ y_2 = -3,\ y_3 = -5.$$

Ahora resolvemos $U \cdot x = y$:

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 0 & -5/3 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -3 \\ -5 \end{bmatrix},$$

tendremos $x_3 = 3, x_2 = 2, x_1 = 1$.

Si consideramos

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -5/3 \end{bmatrix},$$

evidentemente invertible y de inversa

$$D^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1/3 & 0 \\ 0 & 0 & -3/5 \end{bmatrix},$$

podemos hacer

$$A = L \cdot U = L \cdot I \cdot U = L \cdot D \cdot D^{-1} \cdot U = L_0 \cdot U_0,$$

con

$$L_0 = L \cdot D \quad \text{y} \quad U_0 = D^{-1} \cdot U.$$

En nuestro caso

$$L_0 = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1/3 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -5/3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & -3 & 0 \\ 1 & 1 & -5/3 \end{bmatrix}$$

$$\text{y } U_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1/3 & 0 \\ 0 & 0 & -3/5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 \\ 0 & -3 & 1 \\ 0 & 0 & -5/3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1/3 \\ 0 & 0 & 1 \end{bmatrix}.$$

Por lo tanto L_0 tiene por diagonal los pivotes, mientras que U_0 tiene "1" en la diagonal.

Además L se puede obtener a partir de L_0 dividiendo cada columna de L_0 por el pivote de esa columna. Y U se puede obtener a partir de U_0 multiplicando cada fila de U_0 por el pivote de esa fila. □

Esto nos indica que la descomposición LU de A no es única. Sin embargo, si imponemos alguna condición, como que L sea triangular inferior con "1" en la diagonal, entonces la descomposición es única.

En el epígrafe **3.3.2** caracterizábamos el proceso de eliminación gaussiana. En particular vimos que A era triangularizableble por Gauss sin intercambios si y sólo si los menores principales sucesivos de A eran no nulos.

Dada la estrecha relación que estamos observando entre la factorización LU y la eliminación gaussiana no resultará inapropiado adelantar el siguiente resultado que seguidamente demostraremos.

TEOREMA 9 *La condición necesaria y suficiente para que una matriz regular A pueda descomponerse en producto LU de dos matrices, la primera triangular inferior con unos en la diagonal y la segunda triangular superior con elementos diagonales no nulos, es que todos sus menores principales sucesivos sean no nulos. Esta descomposición es única en las condiciones dadas para L y U.*

Demostración.

1. Supongamos que $A = L \cdot U$ en las condiciones enunciadas. Y consideremos particionadas dichas matrices en la forma

$$A = \left[\begin{array}{c|c} A_{r\times r} & \\ \hline & \end{array}\right] = \left[\begin{array}{c|c} L_{r\times r} & \theta \\ \hline & \end{array}\right] \cdot \left[\begin{array}{c|c} U_{r\times r} & \\ \hline \theta & \end{array}\right],$$

entonces $A_{r\times r} = L_{r\times r} \cdot U_{r\times r}$ y por tanto

$$|A_{r\times r}| = |L_{r\times r}| \cdot |U_{r\times r}| = |U_{rxr}|,$$

por lo que, dado que U es triangular superior,

$$|A_{r\times r}| = u_{11} \cdot u_{22} \cdot \cdots \cdot u_{rr}.$$

Por otra parte

$$|A| \neq 0 \text{ y } |A| = |L| \cdot |U| = |U| = u_{11} \cdot u_{22} \cdot \cdots \cdot u_{nn},$$

entonces $u_{ii} \neq 0$, para $i = 1, \ldots, n$.

Luego $|A_{r\times r}| = u_{11} \cdot u_{22} \cdot \cdots \cdot u_{rr} \neq 0$, cualquiera que sea r, con $1 \leq r \leq n$.

2. Recíprocamente si todos los menores principales sucesivos de A son no nulos, entonces el método de Gauss se puede aplicar sin intercambios de filas, como vimos en el epígrafe **3.3.2**. Y en este caso acabamos de ver que $A = L \cdot U$, con L y U verificando las condiciones del enunciado del teorema.

   Veamos la unicidad de la descomposición. Si $A = L \cdot U = L' \cdot U'$, en las condiciones indicadas, entonces $L'^{-1}LU = L'^{-1}L' \cdot U' = U'$, de donde $L'^{-1}LUU^{-1} = U'U^{-1}$, por lo que $L'^{-1}L = U'U^{-1}$, con lo que necesariamente, para que esta matriz sea a la vez triangular superior e inferior, se tiene que $L'^{-1}L$ ha de ser diagonal; es decir $L'^{-1}L = D$ diagonal, y por tanto $L = L'D$. Y como L y L' tienen sólo unos en la diagonal, entonces $D = I$, de donde $L = L'$ y $U = U'$.

■

Una consecuencia inmediata del proceso de demostración de este teorema es que

$$u_{11} = a_{11} \quad \text{y} \quad u_{rr} = \frac{|A_{r\times r}|}{|A_{(r-1)\times(r-1)}|}, \text{ para } r \geq 2.$$

Acabamos de ver la equivalencia entre la posibilidad de factorización LU y la posibilidad de triangularización por Gauss sin intercambios. Pero sabemos que toda matriz regular puede reducirse mediante Gauss a una triangular si se permiten los intercambios. La pregunta, pues, surge inmediatamente: si con intercambios puede completarse la eliminación de Gauss, ¿bajo qué condiciones podremos buscar una factorización de A similar a la anterior?

4.1.2 Descomposicion LU con intercambios.

Básicamente la idea es que si con intercambios puede completarse la eliminación de Gauss, será porque la matriz inicial A se puede modificar en otra matriz A', que sí admite una factorización sin permutación de filas $A' = L \cdot U$.

Ya vimos, en el epígrafe **3.3.2** sobre caracterización del proceso de eliminación gaussiana, que la condición necesaria y suficiente para que el proceso de eliminación de Gauss pudiese llevarse a término, es que la matriz A sea regular. En este caso, y aún cuando fuese necesario la realización de permutaciones de filas, se tendría la descomposición:

$$A^{(n)} = M^{(n-1)} \cdot \ldots \cdot P_{rk} \cdot M^{(r-1)} \cdot \ldots \cdot P_{hl} \cdot M^{(l-1)} \cdot \ldots \cdot M^{(1)} \cdot A,$$

donde las matrices P_{ij} representan las permutaciones de la fila i por la j, que fueron precisas realizar.

En lugar de ir haciendo las permutaciones por etapas, a medida que va apareciendo su necesidad, se puede hacer una permutación adecuada P, previa al comienzo del método, de manera que la matriz resultante $P \cdot A = A'$ se puede triangularizar sin intercambios; o lo que es lo mismo, que $A' = L \cdot U$. Resultado que podemos enunciar mediante el siguiente:

TEOREMA 10 *Para cada matriz regular A existe una matriz P (permutación de filas de A) tal que $A' = P \cdot A$ es factorizable en la forma LU, con unos en la diagonal de L y los elementos diagonales de U no nulos.*

Demostración. En efecto. El proceso de eliminación podemos expresarlo matricialmente en la forma siguiente:

$$M^{(n-1)} \cdot P_{i_{n-1}n-1} \cdot M^{(n-2)} \cdot P_{i_{n-2}n-2} \cdot \cdots \cdot M^{(2)} \cdot P_{i_2 2} \cdot M^{(1)} \cdot P_{i_1 1} \cdot A^{(1)} = A^{(n)}, \tag{4.1}$$

donde $P_{i_k k}$, con $k = 1, \ldots, n-1$, es la matriz de permutación correspondiente a la transformación elemental $F_{i_k k}$ con $i_k > k$, en el caso de que hubiese sido necesario intercambiar las filas i_k y k, o bien la identidad I, si tal transformación no se necesitó realizar; es decir:

$$P_{i_k k} = \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 0 & \cdots & 1 & \\ & & & \vdots & \ddots & \vdots & \\ & & & 1 & \cdots & 0 & \\ & & & & & & \ddots \\ & & & & & & & 1 \end{bmatrix},$$

o bien $P_{i_k k} = I$. Y donde, como recordaremos:

$$M^{(k-1)} = \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & m_{k(k-1)} & & & \\ & & & \vdots & \ddots & & \\ & & & m_{i_k k(k-1)} & & \ddots & \\ & & & \vdots & & & \ddots \\ & & & m_{n(k-1)} & & & 1 \end{bmatrix}.$$

Ahora bien, en el caso en que realmente se haya realizado una permutación de fila, al realizar el producto $P_{i_k k} \cdot M^{(k-1)}$, con $k = 2, \ldots, n-1$, se obtendrá la matriz $M^{(k-1)}$ en la que se han permutado las filas k e i_k. Si esta última matriz se multiplica por la derecha nuevamente por $P_{i_k k}$, el efecto conseguido será el de permutar las columnas k e i_k, con lo que se habrá obtenido una matriz

$$\widehat{M}^{(k-1)} = P_{i_k k} \cdot M^{(k-1)} \cdot P_{i_k k},$$

que difiere únicamente de la matriz $M^{(k-1)}$ en que los elementos m_{kk-1} y $m_{i_k k-1}$ se han intercambiado de sitio y, por tanto, se tendrá que:

$$P_{i_k k} \cdot M^{(k-1)} = \widehat{M}^{(k-1)} \cdot P_{i_k k}.$$

Por ello, (4.1) se transformará en:

$$\widehat{M}^{(n-1)} \cdot \widehat{M}^{(n-2)} \cdot \ldots \cdot \widehat{M}^{(2)} \cdot \widehat{M}^{(1)} \cdot P_{i_{n-1}n-1} \cdot P_{i_{n-2}n-2} \cdot \cdots \cdot P_{i_2 2} \cdot P_{i_1 1} \cdot A^{(1)} = A^{(n)}.$$

Las matrices $\widehat{M}^{(i)}$ son triangulares inferiores con unos en la diagonal, entonces su producto

$$\widehat{M}^{(n-1)} \cdot \widehat{M}^{(n-2)} \cdot \ldots \cdot \widehat{M}^{(2)} \cdot \widehat{M}^{(1)} = \widehat{M}$$

también lo será y el producto de las matrices de permutación

$$P_{i_{n-1}n-1} \cdot P_{i_{n-2}n-2} \cdot \cdots \cdot P_{i_2 2} \cdot P_{i_1 1} = P$$

será una matriz de permutación inicial de la matriz A, de manera que

$$\widehat{M} \cdot P \cdot A^{(1)} = A^{(n)},$$

de donde $P \cdot A^{(1)} = \widehat{M}^{-1} \cdot A^{(n)}$.

Llamando $\widehat{M}^{-1} = L$ y $A^{(n)} = U$ tendremos

$$A' = P \cdot A = L \cdot U.$$

■

Desde luego P es ortogonal ($P^{-1} = P^t$), con lo que $A = P^t \cdot L \cdot U$.

Por otra parte, es claro ya, que:

TEOREMA 11 *Una matriz A es regular si y sólo si se puede descomponer en la forma $A = P^t \cdot L \cdot U$, con las condiciones ya señaladas para cada una de las matrices factores.*

4.2 Cálculo directo de la descomposición LU.

Si los menores principales sucesivos de A son no nulos, entonces es posible la descomposición[1] LU. Pero el cálculo de los elementos l_{ij} y u_{ij} de L y U, respectivamente, puede hacerse sin necesidad de aplicar el método de Gauss.

Puesto que $A = L \cdot U$, tenemos

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix},$$

de donde

$$\begin{aligned} a_{11} &= l_{11}u_{11}, \\ a_{1j} &= l_{11}u_{1j}, && j \geq 2, \\ a_{i1} &= l_{i1}u_{11}, && i \geq 2, \\ a_{kk} &= l_{kk}u_{kk} + \textstyle\sum_{r=1}^{k-1} l_{kr}u_{rk}, && k \geq 2. \end{aligned}$$

$$a_{kj} = l_{kk}u_{kj} + \sum_{r=1}^{k-1} l_{kr}u_{rj},\ k \geq 2 \text{ y } j > k. \tag{4.2}$$

$$a_{ik} = l_{ik}u_{kk} + \textstyle\sum_{r=1}^{k-1} l_{ir}u_{rk}, \quad k \geq 2,\ i > k.$$

Se pueden fijar arbitrariamente los l_{kk} o los u_{kk}, para calcular después l_{11}, u_{1j}, l_{i1}. A continuación l_{22}, u_{2j}, l_{i2}, y así sucesivamente se resuelve el sistema anterior de forma recurrente, ya que sólo aparecen ecuaciones con una sóla incógnita. Por tanto el proceso de cálculo es:

[1] Como acabamos de ver, la descomposición es siempre posible si A es regular, para $A' = P \cdot A$.

1. 1.1 Se calcula la primera fila de U.

   1.2 Se calcula la primera columna de L.

2. 2.1 Cálculo de la segunda fila de U.

   2.2 Cálculo de la segunda columna de L.

3. Se continúa así sucesivamente hasta agotar las n filas de U y las n columnas de L.

Fijados los elementos diagonales de una de las matrices, la descomposición es única como ya sabíamos.

Si exigimos que todos los elementos diagonales de L sean iguales a 1, entonces el resultado de la descomposición LU es el dado por la eliminación gaussiana. El método "compacto" asociado, dado explícitamente por las fórmulas para l_{ij} y u_{ij}, se conoce con el nombre de *Método de Doolittle*. Y si elegimos que sean todos los elementos de la diagonal de U iguales a 1, el método "compacto" asociado se conoce como *Método de Crout*.

4.3 Coste en número de operaciones de la resolución LU.

Una vez descompuesta A, hay que resolver dos sistemas triangulares para resolver $A \cdot x = b$:

$$L \cdot U \cdot x = b \Longleftrightarrow \begin{cases} L \cdot y = b, \\ U \cdot x = y, \end{cases}$$

en el caso de que todos los menores principales sucesivos de A sean no nulos, situación que en la práctica no se conoce a priori. Es más, aún cuando fuera posible esta factorización, normalmente interesa una adecuada elección de los elementos diagonales de U; lo que equivale a decir que en el proceso de triangularización de Gauss es conveniente desarrollar una estrategia adecuada de pivoteo.

Por consiguiente, nos situamos en el caso de una matriz A regular cualquiera. En estas condiciones, siempre es posible: $A = P^t \cdot L \cdot U$.

Y puesto que la diferencia entre los métodos de *Doolittle* y de *Crout* radica exclusivamente en cómo fijemos los elementos de la diagonal principal de L, el coste en número de operaciones será el mismo en uno y otro caso. En lo que sigue consideraremos que estamos ante el método de *Doolittle*.

Los pasos una vez efectuada la descomposición son:

1. Determinar $b' = P \cdot b$.

   Es decir, permutar los elementos de b. Aquí no hay operaciones aritméticas.

2. Resolver $L \cdot y = b'$:

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} b'_1 \\ b'_2 \\ \vdots \\ b'_n \end{bmatrix}.$$

Para calcular y_k se necesitan $k-1$ productos (al sustituir los valores de $y_1, \ldots, y_{k-1}$ anteriormente calculados), y $k-1$ restas puesto que $l_{kk}y_k = b'_k - l_{k1}y_1 - \ldots - l_{kk-1}y_{k-1}$, y además no es necesaria la división porque estamos tomando los elementos diagonales de L iguales a 1. Luego, en total, para cada y_k son precisos $k-1$ productos y $k-1$ sumas. En total:

$$\sum_{k=1}^{n}(k-1) + \sum_{k=1}^{n}(k-1) = 2 \cdot \frac{n\,(n-1)}{2} = n^2 - n.$$

3. Resolver $U \cdot x = y$.

   Es igual que en el caso anterior $L \cdot y = b'$, pero es necesario en cada fila dividir porque los elementos diagonales no son 1. Por tanto se tendrán k productos/divisiones y $k-1$ sumas. En total:

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

Así, pues, el costo en operaciones en resolver el sistema, una vez efectuada la descomposición, es de $2n^2 - n$ operaciones elementales.

Y el coste, en operaciones elementales, del cálculo de la descomposición LU es el mismo que el de la eliminación de Gauss, por tanto $\frac{4n^3-3n^2-n}{6}$.

En definitiva, el costo total entre la descomposición y la resolución implica

$$2n^2 - n + \frac{4n^3 - 3n^2 - n}{6} = \frac{4n^3 + 9n^2 - 7n}{6}$$

operaciones que, como era de esperar, es el mismo coste que el de efectuar la resolución por Gauss.

4.4 Organización computacional.

Dada la íntima relación existente entre el proceso de eliminación de Gauss y la factorización LU, podríamos pensar en efectuar la elección de pivote mediante la estrategia de pivoteo parcial. En este supuesto, la factorización sería la de una matriz A' obtenida de A permutando filas de ella.

La forma más eficiente y más ampliamente utilizada para construir el algoritmo de Gauss consiste en desarrollar dos funciones independientes:

1. La primera efectúa una factorización triangular de la matriz **A** del sistema. Para evitar, en lo posible, graves repercusiones del error de redondeo, se utiliza la eliminación gaussiana con pivoteo parcial, combinada con escalamiento implícito. A esta primera función la llamaremos `factor`.

2. La otra función, que denominaremos `resuelve`, utiliza la matriz A ya factorizada por la función anterior para resolver el sistema $A \cdot x = b$.

Esta estrategia no hace otra cosa que plasmar de una forma eficiente las diversas consideraciones que a lo largo de todo el epígrafe 4 se han venido haciendo. Por ejemplo, resolver distintos sistemas con la misma matriz y distintos vectores de términos independientes[2], lo que es una situación frecuente en la realidad; evitar tanto pivotes nulos como muy pequeños, homogeneizar la magnitud de los elementos de la matriz, etc.

Hay que hacer observar que la función `resuelve` no es, al menos a priori, una función que se pueda utilizar directamente. Su adecuado uso exige que A esté factorizada y que se conozca la estrategia de pivoteo. Debe usarse, por tanto, sólo después de una llamada a `factor`.

Veamos sin más dilación los algoritmos.

Algoritmo p =`factor` (A, n, **pivote**)

1. Entrada:

   A es una matriz de dimensión nxn para ser factorizada en la forma LU mediante eliminación gaussiana con pivoteo parcial y equilibrio implícito en las filas.

   El último argumento **"pivote"** debe ser un vector de valores enteros para almacenar la estrategia de pivoteo utilizada.

   Salida:

   A la salida del algoritmo, la matriz triangular superior U estará almacenada en la parte triangular superior de la matriz A. La parte triangular inferior de A contendrá los multiplicadores que conforman la matriz L.

   El vector pivote contendrá un registro de todos los intercambios de filas.

   La función devuelve un código de salida "p" con el siguiente significado:

0	Éxito.
1	Peligro. Matriz cuasi-singular
2	Fracaso. Matriz singular.

2. $det := 1$
3. $s_i := \max_{1 \le j \le n} |a_{ij}|$, para $i = 1, \ldots, n$
4. Hacer para $k = 1, \ldots, n-1$

   4.1 $c_k := \max_{k \le i \le n} |\frac{a_{ik}}{s_i}|$

   4.2 Si i es el índice más pequeño para el cual el máximo es alcanzado, hacer pivote$_k := i$

   4.3 Si $c_k = 0$, hacer $p := 2$ e ir al paso 7

   4.4 Si pivote$_k \neq k$

   4.4.1 $det := -det$

[2]Comúnmente recibe el nombre de multisistemas.

4.4.2 Intercambiar las filas i y k-ésimas

4.4.3 Intercambiar s_k y s_i

4.5 Hacer para $i = k+1, \ldots, n$

4.5.1 $a_{ik} := m_i := \frac{a_{ik}}{a_{kk}}$

4.5.2 $a_{ij} := a_{ij} - m_i a_{kj}$, para $j = k+1, \ldots, n$

4.6 $det := a_{kk} \cdot det$

5. Si $|det| \leq 1.0e-10$, hacer $p := 1$ e ir al paso 7

6. $p := 0$

7. Salida del algoritmo.

Algoritmo p =`resuelve` $(A, n, b,$ **pivote**$)$

1. Entrada:

   Resuelve el sistema lineal de dimensión $n, A \cdot x = b$. Se asume que la matriz A original ha sido factorizada usando la función `factor` con el intercambio de filas registrado en pivote.

   Salida:

   Devuelve la solución del sistema, que se supone no singular, en b. La matriz A y el vector pivote permanecen inalterados. No se devuelve ningún código de retorno.

2. Hacer para $k = 1, \ldots, n-1$

   2.1 $i := \text{pivote}_k$

   2.2 Si $i \neq k$, intercambiar b_i y b_k

3. $b_i := b_i - \sum_{j=1}^{i-1} a_{ij} b_j$, para $i = 2, \ldots, n$

4. $b_n = \frac{b_n}{a_{nn}}$

5. $b_i := \frac{1}{a_{ii}} \left(b_i - \sum_{j=i+1}^{n} a_{ij} b_j \right)$, para $i = n-1, \ldots, 1$

6. Salida del algoritmo.

5. Caso de matrices especiales

5.1 Matrices simétricas definidas positivas: el método de Cholesky.

Cuando A es simétrica ($A^t = A$) es muy conveniente la factorización y se puede llevar a término sin la necesidad de pivoteo ni escalamiento, y con la mitad de consumo de operaciones elementales.

El proceso es conocido para matrices simétricas definidas positivas desde principios de este siglo. Sin embargo, el desarrollo de métodos numéricos computacionales eficientes, mediante factorización, para matrices indefinidas, ha tenido lugar sólo a partir de los años setenta.

5.1.1 La descomposición a partir de la factorización LU

Supongamos $A \cdot x = b$, siendo A regular y simétrica. Supongamos así mismo que es posible efectuar la eliminación gaussiana sin intercambios, obteniéndose la factorización $A = L \cdot U$, donde L es una matriz triangular inferior con unos en la diagonal y U una matriz triangular superior. Tendremos que

$$L \cdot U = A = A^t = U^t \cdot L^t,$$

entonces, por la unicidad de la factorización LU,

$$U = D \cdot L^t, \text{ donde } D = \text{diagonal } (u_{11}, u_{22}, \ldots, u_{nn}),$$

ya que L tiene unos en la diagonal.

Por ello, desde ahora podemos escribir

$$A = L \cdot D \cdot L^t,$$

que comúnmente se conoce como factorización LDL^t de A.

Si además $D = \text{diagonal } (u_{11}, u_{22}, \ldots, u_{nn})$, es tal que para $i = 1, \ldots, n$ se tiene que $u_{ii} > 0$, entonces tiene sentido considerar, para cada i, $u_{ii}^{1/2}$. Y la matriz diagonal integrada por estos $u_{ii}^{1/2}$,

$$D^{1/2} = \begin{bmatrix} u_{11}^{1/2} & & \\ & \ddots & \\ & & u_{nn}^{1/2} \end{bmatrix}$$

es tal que $D^{1/2} \cdot D^{1/2} = D$.

Por lo que $A = L \cdot D^{1/2} \cdot D^{1/2} \cdot L^t$, con lo que $M = D^{1/2} \cdot L^t$ es triangular superior con $u_{ii}^{1/2}$ en la diagonal y $M^t = L \cdot D^{1/2}$ es triangular inferior y con los mismos elementos diagonales que la anterior. Fijaremos los $u_{ii}^{1/2} > 0$.

Luego, en estas condiciones

$$A = M^t \cdot M.$$

Es de observar que M es regular, toda vez que sus elementos diagonales son no nulos.

La factorización $A = M^t \cdot M$ se llama factorización de Cholesky y M se llama factor de Cholesky.

La factorización de Cholesky es única salvo el signo de sus filas, ya que si $\widehat{M} = E \cdot M$, tal que $E = diag\,(\pm 1)$, también satisface $A = \widehat{M}^t \cdot \widehat{M}$, puesto que $A = \widehat{M}^t \cdot \widehat{M} = (E \cdot M)^t \cdot (E \cdot M) = M^t \cdot E^t \cdot E \cdot M = M^t \cdot I \cdot M = M^t \cdot M$.

No obstante fijaremos el signo de la descomposición, haciendo que los elementos diagonales de M sean positivos, con lo que la descomposición será única.

Hemos obtenido el siguiente resultado:

> *Si es posible efectuar la eliminación gaussiana sin intercambios de filas y todos los pivotes son positivos, entonces existe una matriz M^t triangular inferior con elementos diagonales positivos, tal que la matriz A se puede descomponer como el producto $A = M^t \cdot M$. Tomados los elementos diagonales de M positivos, la descomposición es única y recibe el nombre de "factorización de Cholesky".*

Desgraciadamente la existencia de la factorización LDL^t está basada en la condición (extremadamente fuerte, por otra parte) de que no se requieran intercambios en la factorización $A = L \cdot U$.

Ejemplo 12:

Sea $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, que en sí misma es una matriz permutación.

Si A pudiera escribirse como LDL^t, tendríamos

$$\begin{pmatrix} 1 & 0 \\ l_{21} & 1 \end{pmatrix} \cdot \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \cdot \begin{pmatrix} 1 & l_{21} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

con lo que $\begin{pmatrix} d_1 & l_{21}d_1 \\ l_{21}d_1 & l_{21}^2 d_1 + d_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$,

por tanto $d_1 = 0$ y por otra parte $d_1 l_{21} = 1$, con lo que obviamente tenemos una contradicción. Por consiguiente esta sencilla matriz simétrica no admite la factorización LDL^t. □

Si se realiza incluso un simple intercambio de filas, se deshace generalmente la simetría de la matriz original. Esta simetría se podrá conservar sólo si los mismos intercambios son aplicados tanto a las filas como a las columnas.

Si P es una matriz de permutación, este requisito se corresponde haciendo $P \cdot A \cdot P^t$ en cada paso. Dada una matriz simétrica A y una matriz de permutación P, $P \cdot A \cdot P^t$ recibe el nombre de permutación simétrica de A $((P \cdot A \cdot P^t)^t = P \cdot A^t \cdot P^t = P \cdot A \cdot P^t)$.

Ahora bien, ¿puede siempre encontrarse una permutación simétrica de A, tal que $P \cdot A \cdot P^t = L \cdot D \cdot L^t$?

En contraposición con la existencia garantizada de P tal que $P \cdot A = L \cdot U$ para la eliminación gaussiana, la respuesta es aquí, lamentablemente, no.

En el caso del ejemplo anterior, la única permutación posible es $P = A$, por lo que también la única permutación simétrica posible $P \cdot A \cdot P^t$ nos proporciona nuevamente $A = P \cdot A \cdot P^t$; y ya hemos visto que para A no es posible la factorización LDL^t.

Así pues:

La factorización LDL^t de una matriz simétrica cualquiera no existe necesariamente.

Sin embargo, vamos a demostrar que:

Cuando A es simétrica definida positiva la factorización LDL^t existe y puede ser computada de manera eficiente y estable.

Ya hemos dicho que A es simétrica si $A^t = A$, y decimos que es definida positiva si $x^t \cdot A \cdot x > 0$, cualquiera que sea $x \in \mathbb{R}^n$ no nulo.

Señalamos a continuación algunos resultados de interés, a los que necesitaremos recurrir, acerca de las matrices reales simétricas definidas positivas. Y aunque éstos podemos darlos por sobradamente conocidos, señalaremos sus demostraciones a pie de página, con la intención de que la exposición quede autocontenida en la mayor medida posible. En lo que sigue, como desde el inicio de este capítulo, supondremos que A es simétrica y además que es definida positiva.

1. *A es definida positiva si y sólo si los autovalores λ_i de A son todos positivos.*[1]
2. *A es invertible y su inversa A^{-1} también es definida positiva.*[2]
3. *Todas las submatrices principales de A, $A\begin{pmatrix} 1, & \cdots, & r \\ 1, & \cdots, & r \end{pmatrix}$ formadas por las primeras r filas y las primeras r columnas de A, son definidas positivas.*[3]

[1] Supongamos que A es definida positiva. Consideremos una base de autovectores ortonormales $x_1, \cdots, x_n$; por tanto $x_i^t x_j = \delta_{ij}$. Entonces considerando cualquier autovector de A, se tendrá que $Ax_i = \lambda_i x_i$ y por tanto $x_i^t A x_i = \lambda_i x_i^t x_i = \lambda_i > 0$.

Recíprocamente, si los autovalores son todos positivos, al considerar un vector x cualquiera, podemos expresarlo como combinación lineal de la base de autovectores ortonormales: $x = c_1 x_1 + \cdots + c_n x_n$. De este modo $x^t Ax = (c_1 x_1^t + \cdots + c_n x_n^t) A (c_1 x_1 + \cdots + c_n x_n) = c_1^2 x_1^t A x_1 + \cdots + c_n^2 x_n^t A x_n$, luego $x^t Ax = c_1^2 \lambda_1 + \cdots + c_n^2 \lambda_n > 0$.

[2] Al ser A definida positiva, todos sus autovalores serán positivos y por tanto su determinante también, luego A es regular. Por otra parte, dada la relación existente entre los autovalores de una matriz y los de su inversa, si los de A son positivos, entonces también lo son los de A^{-1} y, como consecuencia de 1, tendremos que A^{-1} es también definida positiva.

[3] En lugar de $A\begin{pmatrix} 1, & \cdots, & r \\ 1, & \cdots, & r \end{pmatrix}$ escribiremos A_r, a fin de simplificar la notación. Consideremos $y \in \mathbb{R}^r$ tal que $y \neq 0$. Entonces, si tomamos un vector u de $\mathbb{R}^n$, tal que sus componentes $u_1, \ldots, u_r$ sean respectivamente iguales a las componentes $y_1, \ldots, y_r$ de y, siendo las restantes componentes ceros, tendremos que $y^t A_r y = u^t A u > 0$, ya que A es definida positiva.

A continuación veremos un teorema que caracteriza la factorización de las matrices simétricas reales definidas positivas.

TEOREMA 12 *Sea A una matriz simétrica real. Entonces son equivalentes:*

1. *A es definida positiva.*
2. *Los menores principales de A son positivos, $|A_k| > 0$ para $k = 1, \ldots, n$.*
3. *Se puede aplicar el método simple de Gauss, o equivalentemente $A = LU$, y todos los pivotes son positivos $u_{ii} > 0$.*
4. *Existe una matriz M^t triangular inferior con elementos diagonales positivos tal que $A = M^t \cdot M$.*

Demostración.

1. De 1. se deduce 2.

   Como $|A| = \prod_{i=1}^{n} \lambda_i$, donde λ_i son los autovalores de A, y éstos son todos positivos por ser A definida positiva, se tiene que $|A| > 0$.

   Como A es definida positiva, entonces todas las submatrices principales A_k, con $k = 1, \ldots, n$, son definidas positivas. Y por consiguiente, para cada una de ellas vale el razonamiento anterior; es decir, sus autovalores serán positivos y consecuentemente sus determinantes también. En definitiva

   $$|A_k| > 0, \quad \text{para} \quad k = 1, \ldots, n.$$

2. De 2. se deduce 3.

   En el epígrafe **4.1.1** vimos la equivalencia entre la factorización LU y la no nulidad de los menores principales. Es más, como consecuencia del proceso de demostración del teorema 9, se dedujo que los elementos diagonales de U (pivotes de la eliminación de Gauss) verificaban que

   $$u_{11} = a_{11} \qquad \text{y} \qquad u_{kk} = \frac{|A_k|}{|A_{k-1}|}, \text{ para } k \geq 2.$$

   Por tanto, como todos los menores principales son positivos, se tiene que $u_{ii} > 0$ para $i = 1, \ldots, n$, y se puede realizar la eliminación de Gauss sin intercambios de filas.

3. De 3. se deduce 4.

   Este resultado no es otro que el obtenido al principio de este epígrafe.

4. De 4. se deduce 1.

   Sea x un vector cualquiera de $\mathbb{R}^n$ no nulo, entonces

   $$x^t A x = x^t M^t M x = (Mx)^t (Mx) > 0,$$

   ya que al ser M regular y $x \neq 0$, entonces $Mx \neq 0$.

■

5.1.2 Planteamiento directo del problema. Algoritmo de Cholesky

En el epígrafe anterior ha quedado inequívocamente determinado cuándo una descomposición $L \cdot D \cdot L^t$ es posible y más específicamente cuándo se puede y cuándo no realizar la factorización de Cholesky.

Hemos partido en nuestro análisis de la descomposición LU. Y al igual que en aquél caso, podemos considerar el problema directamente, tal y como se hizo en el epígrafe **4.2**. Bastará exigir que en las relaciones que allí se obtuvieron fueran:

- $l_{ij} = u_{ji}$ cualesquiera que sean i, j, con $1 \leq i, j \leq n$; es decir que $U = L^t$.
- L real y de elementos diagonales positivos.

Realicemos tal planteamiento y obtengamos las relaciones necesarias para plantear directamente el algoritmo de factorización de Cholesky, que como acabamos de ver sólo es válido para matrices simétricas definidas positivas.

Sea

$$M = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ \vdots & m_{22} & \cdots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & m_{nn} \end{bmatrix}$$

con elementos diagonales indeterminados.

$$A = M^t M = \begin{bmatrix} m_{11} & \cdots & \cdots & 0 \\ \vdots & m_{22} & & \vdots \\ \vdots & & \ddots & 0 \\ m_{1n} & \cdots & \cdots & m_{nn} \end{bmatrix} \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ \vdots & m_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & m_{nn} \end{bmatrix}.$$

Entonces, obtenemos las siguientes relaciones:

- $m_{11} = \sqrt{a_{11}}$.
- $a_{1j} = m_{11} m_{1j}$, luego $m_{1j} = a_{1j}/m_{11}$, para $j = 2, \ldots, n$.
- $a_{kk} = m_{kk}^2 + \sum_{r=1}^{k-1} m_{rk}^2$, por lo que $m_{kk} = \sqrt{a_{kk} - \sum_{r=1}^{k-1} m_{rk}^2}$, con $k = 2, \ldots, n-1$.
- $a_{kj} = m_{kj} m_{kk} + \sum_{r=1}^{k-1} m_{rk} m_{rj}$, de donde $m_{kj} = 1/m_{kk} \left[a_{kj} - \sum_{r=1}^{k-1} m_{rk} m_{rj} \right]$, con $j = k+1, \ldots, n$.
- $a_{nn} = m_{nn}^2 + \sum_{r=1}^{n-1} m_{rn}^2$, luego $m_{nn} = \sqrt{a_{nn} - \sum_{r=1}^{k-1} m_{rk}^2}$.

Obviamente como hay que dividir por los elementos m_{kk}, estos deberán ser no nulos. De hecho, se pone nuevamente de manifiesto que los m_{kk} sean positivos es condición necesaria y suficiente para que la descomposición sea posible.

En el caso de que la descomposición sea posible, acordamos tomar para los radicales el signo positivo con lo que la descomposición será única.

El conjunto de las relaciones anteriores se conoce como *algoritmo de Cholesky.*

5.1.3 Coste en número de operaciones de la resolución de $A \cdot x = b$

Dada la factorización de Cholesky, la solución x tal que

$$A \cdot x = M^t \cdot M \cdot x = b$$

puede resolverse mediante los dos sistemas triangulares

$$M^t \cdot y = b \qquad \text{y} \qquad M \cdot x = y.$$

Veamos ahora el coste en operaciones del método:

1. En el cálculo de los elementos de la fila k-ésima hacen falta:
   - Para calcular m_{kk} una extracción de raíz cuadrada, $k-1$ sumas y $k-1$ productos.
   - Para calcular $m_{kj}, k-1$ sumas y $k-1$ multiplicaciones y una división.
2. En total son[4]
   - n raíces,
   - $n(n-1)/2$ diviones,
   - sumas:

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

   - tantos productos como sumas.
3. Todo ello suma
$$(n^3+3n^2+116n)/6.$$
   - A lo que hay que añadir $2n^2$ operaciones de la resolución de sistemas triangulares.

Por tanto, para n grande, el coste es del orden de $n^3/6$ operaciones, lo que implica una sensible reducción (aproximadamente la mitad del coste en operaciones) respecto de la factorización LU para una matriz no simétrica.

5.1.4 Caso de una matriz no simétrica

Sea $A \cdot x = b$ tal que $|A| \neq 0$, pero no siendo ni definida positiva, ni simétrica.

En estas condiciones el método de Cholesky no puede aplicarse.

Ahora bien, de $A \cdot x = b$ se deduce $A^t \cdot A \cdot x = A^t \cdot b$. Haciendo

$$B = A^t \cdot A \qquad \text{y} \qquad h = A^t \cdot B$$

[4] Hay que tener en cuenta que una raíz cuadrada cuesta aproximadamente 20 operaciones elementales.

queda $B \cdot x = h$.

Por otro lado

1. $B^t = (A^t \cdot A)^t = A^t \cdot A = B$.
2. Para cada $x \neq 0$, $x^t B x = x^t A^t A x = (Ax)^t (Ax) = y^t y$, siendo $y = Ax$.

   Como $x \neq 0$ será $y \neq 0$, pues en caso contrario $Ax = 0$ tendría como solución $x = 0$ al ser A regular, lo que contradice el hecho de ser $x \neq 0$. Por consiguiente, para cada $x \neq 0$, se tiene que $x^t B x = y^t y > 0$.

Hemos pues obtenido a partir de A una matriz B simétrica y definida positiva y evidentemente regular, lo que nos permite aplicar el método de Cholesky al nuevo sistema $B \cdot x = h$ y de esta forma obtener x.

5.2 Matrices banda. Matrices tridiagonales

5.2.1 Matrices banda

Es difícil encontrar matrices que parezcan construidas al azar. Casi siempre hay una estructura visible, incluso a primera vista (a menudo una estructura de simetría y muchos términos nulos). En este caso cuando una matriz "rala"[5] contiene mucho menos de n^2 elementos (piezas de de información), los cálculos debieran realizarse de manera mucho más sencilla que cuando la matriz carece casi por completo de elementos nulos (matriz llena).

Concretamente nos vamos a interesar por el caso de las *"matrices banda"*. Por tal entendemos matrices cuadradas que tienen todos sus elementos iguales a cero con excepción de una banda centrada en la diagonal principal; es decir, sus elementos no nulos están situados en la diagonal principal y en algunas líneas paralelas a aquella, supradiagonales y subdiagonales, habiendo el mismo número de líneas con entradas no nulas tanto por encima como por debajo de la diagonal. El número de líneas con entradas no nulas se llama ancho de banda. En otros términos: A es una matriz a banda de ancho $2\omega + 1$ si cualesquiera que sean $i, j = 1, \ldots, n$, se verifica que $a_{ij} = 0$ si $|i - j| > \omega$.

Las matrices banda surgen fundamentalmente cuando se transforma un problema continuo (ecuaciones diferenciales, etc.) en otro discreto (sistema de ecuaciones lineales).

El problema continuo tendrá una infinidad de incógnitas y no podemos resolverlo exactamente. Por lo tanto, es necesario aproximarlo mediante un problema discreto (cuanto más incógnitas consideremos mayor será la precisión, ¡y también los costos!).

Consideremos un ejemplo de problema continuo y analicemos su discretización.

Ejemplo 13:

Sea la siguiente ecuación diferencial de segundo orden, tipo Euler:

$$x^2 y'' - 2xy' + 2y = 0, \tag{5.1}$$

[5] En inglés SPARSE.

con $1 \leq x \leq 2$.

Para eliminar la arbitrariedad en el planteamiento del problema, en el sentido de que admite una infinidad de soluciones, añadimos unas condiciones de frontera en los extremos del intervalo:

$$y(1) = 1 \qquad \text{e} \qquad y(2) = 4.$$

1. Se resuelve este problema en forma continua.

   Haciendo $x = e^t$, tendremos

   $$\frac{d^2y}{dt^2} - 3\frac{dy}{dt} + 2y = 0,$$

   con lo que $r^2 - 3r + 2 = 0$ de donde $r_1 = 1, r_2 = 2$, luego $y = c_1e^t + c_2e^{2t}$, de donde $y = c_1x + c_2x^2$ es la solución general.

   Aplicando las condiciones de contorno tenemos que

   $$\left.\begin{array}{ll} \text{si } x = 1, \text{ entonces} & c_1 + c_2 = 1 \\ \text{si } x = 2, \text{ entonces} & 2c_1 + 4c_2 = 4 \end{array}\right\} \text{ por lo que } c_2 = 1 \text{ y } c_1 = 0.$$

   Luego la solución exacta es $y = x^2$.

2. Pero nuestro objetivo es generar un problema discreto y por tanto con un número finito de incógnitas, es decir de álgebra lineal.

   Por ello no podemos aceptar más que una cantidad finita de información acerca de "y", digamos su valor en puntos igualmente espaciados del intervalo $(1, 2)$.

   Dividamos, pues, el intervalo en N partes iguales, de forma que el valor de cada parte sea igual h.

   $$\begin{array}{l} x_0 = 1, \\ x_1 = 1 + h, \\ \cdots \\ x_n = 1 + nh, \\ \cdots \\ x_N = 1 + Nh = 2. \end{array} \tag{5.2}$$

   Calcularemos valores $y_0, y_1, \cdots, y_N$ aproximados a la solución verdadera y en los puntos (5.2) que son tales que $x_{n+1} = x_n + h$ y donde

   $$\begin{array}{ll} y_0 = y(1) = 1, & y_N = y(2) = 4, \\ y_n = y(x_n), & y_{n+1} = y(x_{n+1}) = y(x_n + h). \end{array}$$

   ¿Cómo podemos reemplazar las derivadas por aproximaciones suyas? Como cada derivada es un límite de cocientes de diferencias, podemos aproximarla deteniéndonos en un paso finito y sin permitir que h tienda a cero. Por tanto, hacemos uso de las siguientes aproximaciones:

   $$y'(x_n) \simeq \frac{y(x_n + h) - y(x_n)}{h} = \frac{y_{n+1} - y_n}{h},$$

$$y''(x_n) \simeq \frac{y'(x_n+h)-y'(x_n)}{h} = \frac{\frac{y_{n+2}-y_{n+1}}{h} - \frac{y_{n+1}-y_n}{h}}{h}$$
$$= \frac{y_{n+2}-2y_{n+1}+y_n}{h^2}.$$

En cada punto $x_{n+1} = x_n + h$ sustituimos la ecuación diferencial (5.1) por una ecuación lineal haciendo uso de las aproximaciones anteriores, como sigue:

$$(1+nh)^2 \frac{y_{n+2} - 2y_{n+1} + y_n}{h^2} - 2(1+nh)\frac{y_{n+1} - y_n}{h} + 2y_n = 0,$$

de donde

$$\frac{(1+nh)^2}{h^2}(y_{n+2} - 2y_{n+1} + y_n) - \frac{2(1+nh)}{h}(y_{n+1} - y_n) + 2y_n = 0,$$

es decir

$$(\frac{1}{h} + n)^2(y_{n+2} - 2y_{n+1} + y_n) - 2(\frac{1}{h} + n)(y_{n+1} - y_n) + 2y_n = 0$$

y ello para cada $n = 0, \ldots, N-2$, con lo que se obtiene un sistema de ecuaciones lineales.

Para $N = 10$ tenemos que

$$(10+n)^2(y_{n+2} - 2y_{n+1} + y_n) - 2(10+n)(y_{n+1} - y_n) + 2y_n = 0,$$

luego

$$(122 + 22n + n^2)y_n - (220 + 42n + 2n^2)y_{n+1} + (10+n)^2 y_{n+2} = 0.$$

Variando $n = 0, \ldots 8$, obtenemos

$$\begin{array}{lccccccc}
n=0; & 122y_0 & -220y_1 & +100y_2 & & & & = 0 \\
n=1; & & 145y_1 & -264y_2 & +121y_3 & & & = 0 \\
n=2; & & & 170y_2 & -312y_3 & +144y_4 & & = 0 \\
\vdots & & & \cdots & & \cdots & & \vdots \\
n=8; & & & & & 362y_8 & -684y_9 & +324y_{10} = 0
\end{array}$$

Sistema con 11 incógnitas y 9 ecuaciones. Pero como $y_0 = 1$ e $y_{10} = 4$, queda realmente un sistema cuadrado 9×9, que en forma matricial es

$$\begin{bmatrix}
-220 & 100 & & & & & & & \\
145 & -264 & 121 & & & & & & \\
 & 170 & -312 & 144 & & & & & \\
\vdots & \vdots & \vdots & \vdots & \ddots & & & & \\
 & & & & & & 362 & -684
\end{bmatrix}
\begin{bmatrix} y_1 \\ \vdots \\ y_9 \end{bmatrix} =
\begin{bmatrix} -122 \\ 0 \\ \vdots \\ 0 \\ -1296 \end{bmatrix}$$

donde aparece como matriz del sistema una matriz a banda de anchura $2\omega + 1 = 3$. En este caso ($\omega = 1$) la matriz recibe el nombre de *matriz tridiagonal*. Se resuelve por eliminación y se obtiene

x_n	y_n solución del sistema	$y(x_n)$ valor real obtenido de la solución exacta
1	1	1
1.1	y_1	1.21
1.2	y_2	1.44
$\vdots$	$\vdots$	$\vdots$
1.9	y_9	3.61
2	4	4

Si hubiésemos tomado el paso $h = 1/100$ entonces tendríamos un sistema 99×99, consiguiéndose valores más aproximados para las y_n.

□

Como era de esperar los sistemas *con "matrices banda"* se pueden resolver por eliminación. Pero ahora gran parte del trabajo está hecho, ya que debido a la estructura especial de la matriz, el algoritmo de eliminación gaussiana se puede simplificar considerablemente y obtenerse las soluciones de forma muy eficiente.

El hecho de que la matriz tenga muchos ceros ha permitido desarrollar en los últimos años métodos especiales de almacenamiento de matrices muy grandes. De esta forma se han podido resolver sistemas de hasta cien mil incógnitas.

5.2.2 Matrices tridiagonales

Un caso particular de las anteriores matrices, sumamente frecuente (por ejemplo aparece siempre en la discretización de ecuaciones diferenciales de segundo orden, como acabamos de ver) lo constituye el de las matrices tridiagonales. Son éstas, matrices que salvo en la diagonal principal, la supradiagonal y la subdiagonal, todos los elementos son iguales a cero ($\omega = 1$).

Para estas matrices en lugar de utilizar una matriz completa $n \times n$ para almacenarla, se utiliza una matriz $n \times 3$, guardando en el computador sólo los elementos de la banda tridiagonal. Luego se aplica una variante del método de Gauss que lo resuelve con un coste proporcional a n en lugar de serlo a n^3.

Factorización de una matriz tridiagonal

Sea la matriz tridiagonal

$$\begin{bmatrix} a_1 & c_1 & & & & \\ b_2 & a_2 & c_2 & & & \\ & b_3 & a_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & c_{n-1} \\ & & & & b_n & a_n \end{bmatrix}.$$

Efectuamos su factorización por el método de *Crout* o bien por el de *Doolittle*, tal y como describimos a continuación. Si lo hacemos por el método de *Doolittle*, obtenemos tras $F_{21}(-b_2/a_1)$

$$\begin{bmatrix} a_1 & c_1 & & & \\ 0 & \alpha_2 & c_2 & & \\ & b_3 & a_3 & c_3 & \\ & & \ddots & \ddots & \ddots \end{bmatrix},$$

con $\alpha_2 = a_2 - b_2 c_1/a_1 = a_2 - m_2 c_1$, siendo $m_2 = b_2/a_1$.

Ahora realizamos $F_{32}(-b_3/\alpha_2)$ obteniendo

$$\begin{bmatrix} a_1 & c_1 & & & \\ 0 & \alpha_2 & c_2 & & \\ & 0 & \alpha_3 & c_3 & \\ & & \ddots & \ddots & \ddots \end{bmatrix},$$

siendo $\alpha_3 = a_3 - m_3 c_2$, con $m_3 = b_3/\alpha_2$.

Procediendo así sucesivamente tenemos que

$$U = \begin{bmatrix} \alpha_1 & c_1 & & & & \\ & \alpha_2 & c_2 & & & \\ & & \alpha_3 & c_3 & & \\ & & & \ddots & \ddots & \\ & & & & \ddots & c_{n-1} \\ & & & & & \alpha_n \end{bmatrix},$$

siendo $\alpha_1 = a_1, \alpha_{i+1} = a_{i+1} - m_{i+1} c_i$ y $m_{i+1} = b_{i+1}/\alpha_i$, con $i = 1, \ldots, n-1$.

Por tanto

$$L = \begin{bmatrix} 1 & & & & \\ m_2 & 1 & & & \\ & m_3 & 1 & & \\ & & \ddots & \ddots & \\ & & & m_n & 1 \end{bmatrix}.$$

De la misma forma, si lo hacemos por *Crout* obtendremos

$$U = \begin{bmatrix} 1 & \gamma_1 & & & & \\ & 1 & \gamma_2 & & & \\ & & 1 & \gamma_3 & & \\ & & & \ddots & \ddots & \\ & & & & \ddots & \gamma_{n-1} \\ & & & & & 1 \end{bmatrix} \quad y \quad L = \begin{bmatrix} \alpha_1 & & & & \\ b_2 & \alpha_2 & & & \\ & b_3 & \alpha_3 & & \\ & & \ddots & \ddots & \\ & & & b_n & \alpha_n \end{bmatrix},$$

siendo $\alpha_1 = a_1, \alpha_{i+1} = a_{i+1} - b_{i+1}\gamma_i$ y $\gamma_i = c_i/\alpha_i$, con $i = 1, \ldots, n-1$.

En cualquier caso, es necesario que en el proceso de factorización todos los pivotes sean no nulos.

Para resolver el sistema $A \cdot x = k$ hacemos $Ux = y$ y $Ly = k$. Para $Ly = k$, tendremos

$$\begin{aligned} y_1 &= k_1, \\ y_2 &= k_2 - m_2 y_1, \\ &\vdots \\ y_n &= k_n - m_n y_{n-1}. \end{aligned}$$

Y para $Ux = y$, se tiene que

$$\begin{aligned} x_n &= \tfrac{y_n}{\alpha_n}, \\ x_{n-1} &= \tfrac{y_{n-1} - c_{n-1} x_n}{\alpha_{n-1}}, \\ &\vdots \\ x_1 &= \tfrac{y_1 - c_1 x_2}{\alpha_1}. \end{aligned}$$

Resultados referidos al método de *Doolittle*.

De forma análoga se obtendría, para el método de *Crout*. De $Ly = k$, se tiene que

$$\begin{aligned} y_1 &= \tfrac{k_1}{\alpha_1}, \\ y_2 &= \tfrac{k_2 - b_2 y_1}{\alpha_2}, \\ &\vdots \\ y_n &= \tfrac{k_n - b_n y_{n-1}}{\alpha_n}. \end{aligned}$$

De $Ux = y$, se sigue que

$$\begin{aligned} x_n &= y_n, \\ x_{n-1} &= y_{n-1} - \gamma_{n-1} x_n, \\ &\vdots \\ x_1 &= y_1 - \gamma_1 x_2. \end{aligned}$$

Pero como hemos indicado anteriormente, estas descomposiciones no son siempre posibles, dada la necesidad de que los denominadores no se anulen. La pregunta inmediata será ¿bajo qué condiciones podrá realizarse?

Matrices tridiagonales de diagonal dominante

Vamos a ver que en determinadas condiciones podremos asegurar la viabilidad de la factorización. Examinamos las matrices tridiagonales de diagonal dominante.

Una matriz A es de *diagonal dominante* si verifica que

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n. \tag{5.3}$$

Y es de *diagonal estrictamente dominante* si se verifica (5.3) con desigualdad estricta. Cuando se verifica en general (5.3), pero en algún caso la desigualdad es estricta se dice que A es *de diagonal fuertemente dominante*.

TEOREMA 13 *Si A es tridiagonal y verifica que*[6]*:*

$$\begin{aligned} &|a_1| > |c_1| > 0, \\ &|a_k| \geq |b_k| + |c_k| \quad \textit{con } b_k c_k \neq 0,\ k = 2, \ldots, n-1, \\ &|a_n| > |b_n| > 0, \end{aligned} \tag{5.4}$$

entonces

1. $|\gamma_i| < 1$, $i = 1, \ldots, n-1$, *donde los γ_i son los obtenidos por el método de Crout.*
2. $|a_i| - |b_i| < |\alpha_i| < |a_i| + |b_i|$, *con $i = 2, \ldots, n$, donde los α_i son los obtenidos por el método de Crout.*[7]

[6] Es de observar que en las condiciones que se requieren, A es de diagonal fuertemente dominante, y además que $b_i, c_i \neq 0$, así como $a_i \neq 0$ para cada $i = 1, \ldots, n$.

[7] Un enunciado análogo se puede formular para el caso en que la descomposición se haya realizado por el método de Doolittle.

Demostración.

1. Lo haremos por inducción sobre i.

   - $\gamma_1 = c_1/a_1$, entonces $|\gamma_1| = (|c_1|/|a_1|) < 1$, ya que $|a_1| > |c_1| > 0$.
   - Suponemos que $|\gamma_i| < 1$.
   - Lo vemos para γ_{i+1}:

     $$|a_{i+1}| = |a_{i+1} - b_{i+1}\gamma_i + b_{i+1}\gamma_i| \leq |a_{i+1} - b_{i+1}\gamma_i| + |b_{i+1}\gamma_i|,$$

     pero $|\gamma_i| < 1$, por hipótesis de inducción, quedando

     $$|a_{i+1}| \leq |a_{i+1} - b_{i+1}\gamma_i| + |b_{i+1}|,$$

     de donde

     $$|a_{i+1}| - |b_{i+1}| < |a_{i+1} - b_{i+1}\gamma_i|$$

     y ya que por hipótesis $|a_k| - |b_k| \geq |c_k| > 0$, si $k = 2, \ldots, n-1$, tenemos que

     $$0 < |a_{i+1}| - |b_{i+1}| < |a_{i+1} - b_{i+1}\gamma_i| = |\alpha_{i+1}|. \tag{5.5}$$

     Por otra parte

     $$|\gamma_{i+1}| = |c_{i+1}|/|\alpha_{i+1}|, \text{ por definición de } \gamma_i. \tag{5.6}$$

     De (5.5), (5.6) y de la segunda hipótesis del teorema se tiene que

     $$|\gamma_{i+1}| < |c_{i+1}|/(|a_{i+1}| - |b_{i+1}|) \leq 1,$$

     con lo que $|\gamma_{i+1}| < 1$.

2. Veamos ahora la acotación de $|\alpha_k|$.

   - Por definición $|\alpha_i| = |a_i - b_i\gamma_{i-1}|$, luego $|\alpha_i| \leq |a_i| + |b_i\gamma_{i-1}|$. Y $|b_i\gamma_{i-1}| = |b_i||\gamma_{i-1}| < |b_i|$, ya que $|\gamma_{i-1}| < 1$, por **1**. Por lo tanto

     $$|\alpha_i| < |a_i| + |b_i|. \tag{5.7}$$

   - Por otra parte, de la definición de α_i se tiene que $a_i = \alpha_i + b_i\gamma_{i-1}$, entonces $|a_i| \leq |\alpha_i| + |b_i\gamma_{i-1}| < |\alpha_i| + |b_i|$, por la misma razón que antes. De donde resulta que

     $$|\alpha_i| > |a_i| - |b_i|. \tag{5.8}$$

     De (5.7) y (5.8) concluimos $|a_i| - |b_i| < |\alpha_i| < |a_i| + |b_i|$.

■

COROLARIO 14 *En las condiciones del teorema anterior, la descomposición para matrices tridiagonales es posible.*

Demostración. En efecto, ya indicamos que para que la descomposición fuese posible $\alpha_i \neq 0$, para $i = 1, \ldots, n$.

Ahora bien:

- $|\alpha_1| = |a_1| > |c_1| > 0$, por la primera de las hipótesis del teorema.
- Si $i \in \{1, \ldots, n\}$, $\alpha_i = a_i - b_i\gamma_{i-1}$, según su definición.

  En las condiciones del teorema, la segunda de sus tesis nos garantiza que $|a_i| - |b_i| < |\alpha_i| < |a_i| + |b_i|$, de donde $\alpha_i \neq 0$.

■

Concluimos que:

> *Si A es tridiagonal de diagonal fuertemente dominante y todos sus elementos "tridiagonales" son no nulos*[8]*, entonces A admite una descomposición en producto de dos matrices bidiagonales obtenidas por Crout o Doolittle.*

COROLARIO 15 *Toda matriz tridiagonal A que verifique*[9] *(5.4) es invertible.*

Demostración. Resulta inmediata de la descomposición factorial de A, un factor con determinante 1 y el otro con determinante $\prod_{i=1}^{n} \alpha_i \neq 0$. ■

El coste del método para la total resolución del sistema es de $8n - 7$ operaciones, resultado de las $3n - 3$ sumas, los $3n - 3$ productos y las $2n - 1$ divisiones necesarias.

[8] La no nulidad de los b_i y c_i en (5.4) no es esencial, porque si alguno es nulo el sistema se puede descomponer en dos o más, más simples, tridiagonales también.

[9] De diagonal fuertemente dominante con elementos tridiagonales no nulos.

6. Resolución de multisistemas. Inversa de una matriz

6.1 Multisistemas

Como hemos referido ya con anterioridad, con frecuencia se presentan sistemas de ecuaciones que tienen la misma matriz.

En el epígrafe **3.3.4** relacionamos esta situación con la búsqueda de un almacenamiento eficiente en la computadora. Vimos que el procedimiento más correcto era proceder en dos pasos:

1. Triangulación de la matriz A del sistema y almacenamiento en una misma matriz de la transformada de A y de los multiplicadores de la eliminación de Gauss.
2. Resolución de cada sistema, toda vez que la información sobre las transformaciones realizadas en el proceso de eliminación quedaron almacenadas en la parte triangular inferior de la nueva matriz.

Más adelante, en el epígrafe **4.4**, volvimos a este asunto proporcionando los algoritmos para la organización computacional de esta cuestión. Y pudimos comprobar cómo este procedimiento no es otro que el de realizar primero la factorización LU y después resolver cada sistema en cuestión.

Nuestra intención ahora es realizar algunas precisiones más sobre este asunto y ponerlo en conexión con el problema de la determinación de la inversa.

Cuando se presenta la necesidad de resolver varios sistemas que únicamente difieren en el término independiente (*multisistemas*):

$$\begin{aligned} Ax_1 &= b_1, \\ Ax_2 &= b_2, \\ &\cdots\cdots \\ Ax_h &= b_h, \end{aligned} \tag{6.1}$$

que también podemos escribir como

$$A \cdot [x_1, x_2, \ldots, x_h] = [b_1, b_2, \ldots, b_h],$$

otra alternativa consiste en determinar la inversa de A y luego proceder a realizar $x_i = A^{-1}b_i$, para cada $i = 1, \ldots, h$.

Como el producto matricial es muy rápido, el consumo real de tiempo se llevará por una vez en la determinación de A^{-1}.

El problema se ha trasladado, desde esta perspectiva, a la determinación de A^{-1}.

También podríamos haber formado la matriz "multipleampliada"

$$[A, b_1, b_2, \ldots, b_h] \tag{6.2}$$

que caracteriza a los sistemas (6.1) y mediante transformaciones de Gauss-Jordan pasar a

$$[I, b_1^{(n+1)}, b_2^{(n+1)}, \ldots, b_h^{(n+1)}], \tag{6.3}$$

con lo que $x_i = b_i^{(n+1)}$ para cada sistema $Ax_i = b_i$ de (6.1).

6.2 Cálculo de la inversa mediante Gauss-Jordan

Consideremos ahora el caso especial en que $h = n$ y tal que la matriz $[b_1, b_2, \ldots, b_n] = I$. Por lo tanto ahora tendremos que (6.2) queda en la forma $[A|I]$, que mediante las transformaciones de Gauss-Jordan se convertirá en $[I|I^{(n+1)}]$.

Ahora bien, las operaciones elementales realizadas sobre las filas de A que la convierten en la identidad, son las mismas que las efectuadas sobre las filas de I y que la transforman en $I^{(n+1)}$. Si llamamos $B = I^{(n+1)}$ y E representa el producto de las matrices elementales correspondientes a las sucesivas operaciones elementales, entonces tenemos que $E \cdot A = I$ y que $E \cdot I = B$, resultando que $B = A^{-1}$. En realidad lo que se ha hecho es calcular la inversa de A resolviendo el sistema $A \cdot X = I$.

En general es necesario la técnica de pivoteo, puesto que el esquema de inversión es, en esencia, una eliminación de Gauss. Afortunadamente la matriz inversa no se ve afectada por un cambio en el orden secuencial de las ecuaciones. La columna i-ésima de A^{-1} es la solución de $A \cdot x = e_i$. Y el orden secuencial de los elementos en x no se ve afectado por mezclar el orden de las ecuaciones, por lo que no se altera A^{-1} por el pivoteo.

6.3 Cálculo de la inversa mediante Gauss

El cálculo de la inversa de una matriz también puede hacerse, obviamente, por el método de Gauss. La ventaja aquí será que el número de operaciones elementales es menor que en el método de Gauss-Jordan, toda vez que sólo hay que triangularizar la matriz A.

Para hallar la matriz inversa de A, hay que resolver el multisistema de orden n, $A \cdot x_i = e_i$, donde e_i es el i-ésimo vector canónico.

Los n sistemas resueltos simultáneamente vienen dados por

$$\left[\begin{array}{ccccc|c|c|c|c} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} & 1 & 0 & & 0 \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} & 0 & 0 & & 1 \end{array}\right],$$

de forma que si escribimos $m_{ir} = -a_{ir}^{(r)}/a_{rr}^{(r)}$, con $r < i < n$, tenemos

$$a_{ij}^{(r+1)} = a_{ij}^{(r)} + m_{ir} a_{rj}^{(r)}, \qquad \text{con } r < i \leq n \text{ y } r < j \leq 2n. \tag{6.4}$$

Observamos que, dado r, los $a_{ij}^{(r)}$ son nulos a la izquierda de la diagonal de I, es decir si $r < j \leq r + n$. Así pues es innecesario transformarlos y por tanto en (6.4) es suficiente que el índice j vaya desde $r + 1$ hasta $r + n$.

Debemos resolver, finalizado el proceso de triangularización, los n sistemas. La columna j-ésima de A^{-1} debe satisfacer

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & & \\ & & & & a_{nn}^{(n)} \end{bmatrix} \cdot \begin{bmatrix} \alpha_{1j} \\ \alpha_{2j} \\ \vdots \\ \alpha_{nj} \end{bmatrix} = \begin{bmatrix} a_{1n+j}^{(1)} \\ \vdots \\ a_{jn+j}^{(j)} \\ a_{nn+j}^{(n)} \end{bmatrix}, \tag{6.5}$$

donde la columna j-ésima de A^{-1} es $[\alpha_{1j}, \ldots, \alpha_{nj}]$ y donde si los elementos de la matriz inicial se designan por a_{ij} para la matriz entera $n \times 2n$, $a_{ij} = \delta_{ij-n}$ para $j = n+1, \ldots, 2n$.

En (6.5): $a_{jn+j}^{(j)} = 1$ y $a_{kn+j}^{(k)} = 0$, para $k = 1, \ldots, j-1$. Además para una fila i se tiene que

$$\alpha_{ij} a_{ii}^{(i)} + \alpha_{i+1j}^{(i)} + \ldots + \alpha_{nj} a_{in}^{(i)} = a_{in+j},$$

con lo que

$$\alpha_{ij} = \frac{a_{in+j}^{(i)} - \sum_{e=i+1}^{n} a_{ie}^{(i)} \alpha_{ej}}{a_{ii}^{(i)}}.$$

Fijada la columna j se realiza la determinación de abajo hacia arriba, es decir desde α_{nj} hacia α_{1j}.

7. Error y Condicionamiento

En el epígrafe **3.3.5** se han abordado algunos problemas que pueden surgir en la resolución de un sistema de ecuaciones, como los errores de redondeo o las matrices mal condicionadas. Queremos ahora hacer un estudio más detenido del condicionamiento y perturbaciones de un sistema lineal.

Éste y los métodos iterativos requieren del uso de normas vectoriales y matriciales, así como de sucesiones y series matriciales, por ello pasamos a exponer someramente los aspectos más relevantes sobre dichas cuestiones, en relación al tema que nos ocupa.

Sobre algunos resultados relativos a espacios topológicos y a espacios normados pasaremos rápidamente, recordándolos, en tanto que sean necesarios, pero sin profundizar en este terreno, por no ser el específico de este trabajo.

7.1 Norma vectorial

Dado un espacio vectorial sobre el cuerpo[1] de los números reales $V(\mathbb{R})$, una aplicación $\| \, . \, \| \colon V \to \mathbb{R}$, tal que:

1. $\| x \| > 0$, si $x \neq 0$.
2. $\| \lambda x \| = |\lambda| \, \| x \|$, para todo $\lambda \in \mathbb{R}$ y para todo $x \in V$.
3. $\| x + y \| \leq \| x \| + \| y \|$, cualesquiera que sean $x, y \in V$.

recibe el nombre de norma vectorial sobre $V(\mathbb{R})$.

En particular la propiedad 2. implica, para $\lambda = 0$, que $\| 0 \| = 0$.

Se llama *espacio normado* a un par $(V, \| \, . \, \|)$, donde V es un espacio vectorial real y $\| \, . \, \|$ es una norma sobre V.

Todo espacio normado $(V, \| \, . \, \|)$ se puede considerar como un espacio métrico (V, d), donde d es la distancia inducida por la norma en el espacio vectorial, definida por:

$$\begin{aligned} V \times V &\to \mathbb{R} \\ (x, y) &\to d(x, y) = \| x - y \|, \end{aligned}$$

cuya comprobación resulta inmediata.

Esta distancia inducida por la norma verifica además que

$$d(x + z, y + z) = d(x, y)$$

y que

$$d(\lambda x, \lambda y) = |\lambda| d(x, y).$$

[1] El cuerpo podría ser también el de los complejos, pero para nuestras necesidades sólo hemos de considerar los espacios normados reales.

Dados $a \in V$ y $r \in \mathbb{R}^+$ podemos definir la bola abierta de centro a y radio r

$$B(a,r) = \{x \in V / d(a,x) < r\}.$$

Y la familia $\mathcal{B}$, de todas las bolas abiertas de V, es base de la topología métrica $\mathcal{T}$ en V.

Por consiguiente, una norma en V induce una topología en dicho espacio vectorial. Diremos que dos normas sobre V son equivalentes si inducen la misma topología.

Dadas dos topologías $\mathcal{T}_1$ y $\mathcal{T}_2$, se dice que la topología $\mathcal{T}_1$ es más fina que la $\mathcal{T}_2$, si todo abierto de $\mathcal{T}_2$ es también abierto de $\mathcal{T}_1$.

Dadas dos topologías $\mathcal{T}_1$ y $\mathcal{T}_2$, inducidas respectivamente por las normas $\| \cdot \|_1$ y $\| \cdot \|_2$, se demuestra que $\mathcal{T}_1$ es más fina que $\mathcal{T}_2$ si y sólo si existe $k \in \mathbb{R}^+$, tal que para todo $x \in V$ se verifica que $\| x \|_2 \leq k \| x \|_1$.

De aquí, resulta inmediato que dos normas $\| \cdot \|_1$ y $\| \cdot \|_2$ sobre V son equivalentes si y sólo si existen dos constantes positivas $\alpha, \beta \in \mathbb{R}^+$ tales que $\alpha \| x \|_1 \leq \| x \|_2 \leq \beta \| x \|_1$, cualquiera que sea $x \in V$.

Las normas más usuales en $\mathbb{R}^n$ son las p-normas

$$\| x \|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}, \quad \text{con } 1 \leq p < +\infty,$$

y la norma del supremo

$$\| x \|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

Para $p = 2$ se tiene la llamada norma euclídea. Y la distancia inducida por esta norma recibe el nombre de distancia euclídea, que coincide con la distancia ordinaria en los casos $n = 1, 2, 3$.

Haciendo uso de las desigualdades de Hölder y de Minkowski[2] podemos comprobar que las p-normas son efectivamente normas vectoriales y además que en espacios vectoriales de dimensión finita dos p-normas cualesquiera son equivalentes. Es más, sobre un espacio vectorial de dimensión finita todas las normas son equivalentes.

7.2 Norma matricial

7.2.1 Consideraciones Generales

Como el espacio vectorial real de las matrices reales rectangulares de orden $m \times n$, $\mathcal{M}_{m \times n}(\mathbb{R})$, es isomorfo a $\mathbb{R}^{mn}$, se puede definir una norma en $\mathcal{M}_{m \times n}(\mathbb{R})$, como una norma en el espacio vectorial $\mathbb{R}^{mn}$. Pero esta forma de definir una norma sobre un espacio vectorial de matrices, no tiene en cuenta que en el caso de las matrices cuadradas $\mathcal{M}_{n \times n}(\mathbb{R})$ (en adelante denotaremos este espacio por $\mathbb{R}^{n,n}$) hay definido un producto, por lo que estamos

[2] Con $p, q > 1$ y $1/p + 1/q = 1$, se tiene:
Desigualdad de Hölder: $\sum_1^n |x_i y_i| \leq \left(\sum_1^n |x_i|^p\right)^{1/p} \cdot \left(\sum_1^n |y_i|^q\right)^{1/q}$.
Desigualdad de Minkowski: $\left(\sum_1^n |x_i + y_i|^p\right)^{1/p} \leq \left(\sum_1^n |x_i|^p\right)^{1/p} + \left(\sum_1^n |y_i|^p\right)^{1/p}$.

ante una estructura de *álgebra*. Para contemplar esta situación, se le hace desempeñar un papel particular a las normas en el álgebra de las matrices cuadradas reales que, además de verificar las condiciones indicadas para las normas vectoriales, verifican:

$$\| A \cdot B \| \leq \| A \| \cdot \| B \| .$$

Las normas de esta forma se llaman multiplicativas. Nosotros incluiremos esta condición dentro de la definición siguiente.

Si consideramos el conjunto $\mathbb{R}^{n,n}$ de las matrices cuadradas reales[3] de orden n, $A_{n\times n} \in \mathbb{R}^{n,n}$, una aplicación

$$\| \cdot \| : \mathbb{R}^{n,n} \to \mathbb{R}$$

tal que

1. $\| A \| > 0$, si $A \neq \theta$,
2. $\| \lambda A \| = |\lambda| \, \| A \|$ para todo $\lambda \in \mathbb{R}$ y para todo $A \in \mathbb{R}^{n,n}$,
3. $\| A + B \| \leq \| A \| + \| B \|$ cualesquiera que sean $A, B \in \mathbb{R}^{n,n}$,
4. $\| A \cdot B \| \leq \| A \| \cdot \| B \|$, para cada $A, B \in \mathbb{R}^{n,n}$,

recibe el nombre de norma matricial sobre $\mathbb{R}^{n,n}$.

Dadas dos normas, una matricial $\| \cdot \|_M$ en $\mathbb{R}^{n,n}$ y otra vectorial $\| A \|_v$ en $\mathbb{R}^n$, se dice que ambas son compatibles si se verifica que

$$\| Ax \|_v \leq \| A \|_M \cdot \| x \|_v, \quad \text{cualesquiera que sean } A \in \mathbb{R}^{n,n} \text{ y } x \in \mathbb{R}^n.$$

A partir de una norma vectorial puede definirse una matricial[4], mediante

$$\| A \|_M = \sup_{\|x\|_v = 1} \| Ax \|_v = \max_{\|x\|_v = 1} \| Ax \|_v,$$

que se denomina norma matricial subordinada o inducida por la norma vectorial. Además esta norma matricial es compatible con la vectorial a partir de la cual se define.

Son especialmente usuales:

1. La norma columna $\| A \|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|$, inducida por la norma 1 vectorial.
2. La norma espectral $\| A \|_2 = \rho(A^t A)^{1/2}$, inducida por la norma vectorial euclídea. Teniendo en cuenta que $\rho(B) = \max_{1\leq i\leq n} |\lambda_i|$ y se denomina radio espectral de B, siendo λ_i los autovalores de B.
3. La norma fila $\| A \|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$, inducida por la norma vectorial del supremo.

[3] Análogamente se puede hacer para el caso complejo, aunque en nuestro caso carezca de interés.

[4] Ya que $\{x : \| x \|_v = 1\}$ es un compacto de $\mathbb{R}^n$, y $x \to \| Ax \|_v$ es continua sobre dicho compacto, por lo que el supremo se alcanza. Para ver una demostración completa y rigurosa de este resultado puede consultarse GASTINEL, N., 1975, pgs. 33-40.

Puede ésto comprobarse acotando $\| Ax \|_v$, con $x \in \mathbb{R}^n$ tal que $\| x \|_v = 1$, y viendo posteriormente que la cota se alcanza para alguno de esos vectores, que habrá que buscar en cada caso.

No toda norma vectorial y matricial son compatibles, por ejemplo $\| x \|_\infty$ y $\| A \|_1$ no son compatibles. Por otra parte, hay normas matriciales que no están inducidas por ninguna norma vectorial, ya que para que una norma matricial esté inducida por una vectorial, cualquiera que ésta sea, necesariamente se tendrá que $\| I \|_M = 1$ y, por ejemplo, la norma matricial euclídea verifica que

$$\| I \|_e = \left(\sum_{i,j=1}^{n} a_{ij}^2 \right)^{1/2} = \sqrt{n}.$$

Sin embargo, no sólo es cierto que dada una norma vectorial siempre existe una matricial compatible con ella (la inducida), sino que el recíproco también es cierto, pues dada una norma matricial existen infinitas normas vectoriales compatibles con ella. Bastaría definir, para $x \in \mathbb{R}^n$, $\| x \|_v = \| xu^t \|_M$, siendo $u \in \mathbb{R}^n$ un vector no nulo cualquiera que fijamos libremente.

7.2.2 Acotaciones de normas matriciales

Algunas acotaciones de las normas matriciales serán particularmente útiles.

Consideremos λ autovalor de A y x un autovector asociado, $Ax = \lambda x$. Dada la norma matricial $\| \cdot \|_M$ hay una norma vectorial compatible con ella, de la forma $\| y \|_v = \| yu^t \|_M$. Por ello $\| Ax \|_v \leq \| A \|_M \cdot \| x \|_v$. Y por otra parte $\| Ax \|_v = \| \lambda x \|_v = |\lambda| \, \| x \|_v$. Luego $|\lambda| \, \| x \|_v \leq \| A \|_M \cdot \| x \|_v$.

Y como x es un autovector, entonces $\| x \|_v \neq 0$.

Por tanto, $|\lambda| \leq \| A \|_M$, cualquiera que sea el autovalor λ de A. En particular tomando el autovalor máximo, tendremos que

$$\rho(A) \leq \| A \|_M .$$

Con lo que hemos obtenido el siguiente resultado.

TEOREMA 16 *Se verifica que $\rho(A) \leq \| A \|_M$, cualquiera que sea la norma matricial $\|.\|_M$.*

TEOREMA 17 *Para cada matriz A de $\mathbb{R}^{n,n}$ y para cada número real $\varepsilon > 0$ puede definirse una norma matricial, que depende de A y de ε, de forma que la norma de A difiera del radio espectral $\rho(A)$ precisamente en ε.*

Demostración. Sea J la forma canónica de Jordan de A, $R^{-1}AR = J$, con

$$J = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_k \end{bmatrix} \quad \text{donde } J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & 1 & \\ & & \ddots & \ddots \\ & & & \ddots & 1 \\ & & & & \lambda_i \end{bmatrix}$$

y tal que $J_i = \lambda_i I_i + N_i$, siendo

$$N_i = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & 0 \end{bmatrix} \text{ nilpotente.}$$

Para $\varepsilon \in \mathbb{R}, \varepsilon > 0$, definimos

$$\Omega_{i_\varepsilon} = \begin{bmatrix} 1 & & & & \\ & \varepsilon & & & \\ & & \varepsilon^2 & & \\ & & & \ddots & \\ & & & & \varepsilon^{m_i-1} \end{bmatrix},$$

donde m_i es la multiplicidad de λ_i. Se verifica que

$$\Omega_{i_\varepsilon}^{-1} N_i \Omega_{i_\varepsilon} =$$

$$\begin{bmatrix} 1 & & & & \\ & 1/\varepsilon & & & \\ & & 1/\varepsilon^2 & & \\ & & & \ddots & \\ & & & & 1/\varepsilon^{m_i-1} \end{bmatrix} \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & 0 \end{bmatrix} \begin{bmatrix} 1 & & & & \\ & \varepsilon & & & \\ & & \varepsilon^2 & & \\ & & & \ddots & \\ & & & & \varepsilon^{m_i-1} \end{bmatrix} =$$

$$\begin{bmatrix} 1 & & & & \\ & 1/\varepsilon & & & \\ & & 1/\varepsilon^2 & & \\ & & & \ddots & \\ & & & & 1/\varepsilon^{m_i-1} \end{bmatrix} \begin{bmatrix} 0 & \varepsilon & & & \\ & 0 & \varepsilon^2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \varepsilon^{m_i-1} \\ & & & & 0 \end{bmatrix} =$$

$$\begin{bmatrix} 0 & \varepsilon & & & \\ & 0 & \varepsilon & & \\ & & \ddots & \ddots & \\ & & & \ddots & \varepsilon \\ & & & & 0 \end{bmatrix} = \varepsilon N_i.$$

Por otro lado,

$$\Omega_{i_\varepsilon}^{-1} J_i \Omega_{i_\varepsilon} = \Omega_{i_\varepsilon}^{-1} (\lambda_i I_i + N_i) \Omega_{i_\varepsilon} = \lambda_i I_i + \varepsilon N_i =$$

$$\begin{bmatrix} \lambda_i & \varepsilon & & & \\ & \lambda_i & \varepsilon & & \\ & & \ddots & \ddots & \\ & & & \ddots & \varepsilon \\ & & & & \lambda_i \end{bmatrix} = H_i.$$

Tomando la norma fila[5], $||\Omega_{i_\varepsilon}^{-1} J_i \Omega_{i_\varepsilon}||_\infty = |\lambda_i| + \varepsilon$.

[5] Si la matriz es de orden mayor que 1, porque si es de orden 1 entonces $\| \Omega_{i_\varepsilon}^{-1} J_i \Omega_{i_\varepsilon} \| = |\lambda_i|$.

Construimos ahora

$$\Omega_\varepsilon = \begin{bmatrix} \Omega_{1_\varepsilon} & & & \\ & \Omega_{2_\varepsilon} & & \\ & & \ddots & \\ & & & \Omega_{k_\varepsilon} \end{bmatrix}$$

donde cada elemento de la diagonal de esta matriz celular tiene la misma dimensión que el correspondiente bloque de Jordan de la matriz J.

Haciendo ahora $\Omega_\varepsilon^{-1} J \Omega_\varepsilon = \begin{bmatrix} H_1 & & & \\ & H_2 & & \\ & & \ddots & \\ & & & H_k \end{bmatrix}$,

con lo que, teniendo en cuenta la norma fila de cada bloque, tenemos que

$$\| \Omega_\varepsilon^{-1} J \Omega_\varepsilon \|_\infty = \max|\lambda_i| + \varepsilon.$$

La igualdad $\| \Omega_\varepsilon^{-1} J \Omega_\varepsilon \|_\infty = \| \Omega_\varepsilon^{-1} R^{-1} A R \Omega_\varepsilon \|_\infty = \rho(A) + \varepsilon$, puede escribirse como $\| H^{-1} A H \|_\infty = \rho(A) + \varepsilon$, donde $H = R\Omega_\varepsilon$.

Por ello $\rho(A) \leq \| H^{-1} A H \|_\infty \leq \rho(A) + \varepsilon$. Llamando $\| A \|_{H\varepsilon} = \| H^{-1} A H \|_\infty$, tendremos que $\rho(A) \leq \| A \|_{H\varepsilon} = \rho(A) + \varepsilon$. ■

De la anterior desigualdad y del teorema 16, podemos deducir que

$$\rho(A) = \inf_{\|\cdot\|} \| A \|, \tag{7.1}$$

para cada matriz cuadrada A.

7.3 Sucesiones matriciales

7.3.1 Convergencia de sucesiones matriciales

Consideraremos una sucesión de matrices cuadradas[6]

$$A^{(1)}, A^{(2)}, \ldots, A^{(m)}, \ldots$$

que denotaremos por $\{A^{(m)}\}_1^\infty$, y tal que $A^{(m)} = (a_{ij}^{(m)})_{1 \leq i,j \leq n} \in \mathbb{R}^{n,n}$.

Decimos que $A \in \mathbb{R}^{n,n}$ es límite de la sucesión de matrices $\{A^{(m)}\}_1^\infty$, y lo denotamos por $\lim_{m \to \infty} A^{(m)} = A$, si y sólo si cualquiera que sea el número real $\varepsilon > 0$, existe un entero positivo m_0, tal que para cada $m \geq m_0$ se verifica que $\| A^{(m)} - A \| < \varepsilon$, para alguna norma $\| \cdot \|$ definida sobre $\mathbb{R}^{n,n}$. Es decir, si para alguna norma $\| \cdot \|$, se verifica que $\lim_{m \to \infty} \| A^{(m)} - A \| = 0$.

Naturalmente la definición es independiente de la norma tomada, pues todas las normas de $\mathbb{R}^{n,n}$ son equivalentes. Por lo que, si $\| \cdot \|$ y $\| \cdot \|^*$ son dos normas definidas en $\mathbb{R}^{n,n}$, se tendrá que

$$\lim_{m \to \infty} \| A^{(m)} - A \| = 0 \text{ si y sólo si } \lim_{m \to \infty} \| A^{(m)} - A \|^* = 0.$$

[6] Valdría también para matrices rectangulares $p \times q$.

Una sucesión de matrices que tenga límite se denomina *sucesión convergente.*

Es fácil comprobar además que

$$\lim_{m\to\infty} \| A^{(m)} \| = \| A \|,$$

con sólo considerar que $| \| A^{(m)} \| - \| A \| | \leq \| A^{(m)} - A \|$.

Si una sucesión de matrices $\{A^{(m)}\}_1^\infty$ converge a la matriz A, entonces existen los n^2 límites $\lim_{m\to\infty} a_{ij}^{(m)}$, con $1 \leq i, j \leq n$ y se verifica que

$$\lim_{m\to\infty} A^{(m)} = \left(\lim_{m\to\infty} a_{ij}^{(m)}\right)_{1\leq i,j\leq n} = A,$$

como es fácilmente demostrable a partir de la norma fila.

7.3.2 Sucesión de potencias de una matriz

Para la resolución aproximada de sistemas de ecuaciones lineales tiene especial interés el hecho de que la sucesión de potencias de una matriz tienda a la matriz nula θ. Por ello es muy importante el siguiente resultado.

TEOREMA 18 *Si $A \in \mathbb{R}^{n,n}$, la condición necesaria y suficiente para que $\lim_{m\to\infty} A^m = \theta$ es que $\rho(A) < 1$, donde A^m es la potencia m-ésima de A.*

Demostración.

1. La sucesión $A^{(m)}$ es ahora la sucesión A^m de potencias de A.

   Veamos la condición necesaria por reducción al absurdo.

   Supongamos que $\lim_{m\to\infty} A^m = \theta$, pero $\rho(A) \geq 1$. Entonces existirá un autovalor λ de la matriz A tal que $|\lambda| \geq 1$.

   Sea u un autovector asociado a λ, $Au = \lambda u$, entonces

   $$A^m u = \lambda^m u. \tag{7.2}$$

   Como dada una norma matricial siempre se puede elegir una vectorial compatible con ella, tendremos que

   $$\| A^m u \|_v \leq \| A^m \|_M \| u \|_v;$$

   tomando límites

   $$\lim_{m\to\infty} \| A^m u \|_v \leq \lim_{m\to\infty} \| A^m \|_M \| u \|_v = 0, \text{ al ser } \lim_{m\to\infty} A^m = \theta,$$

   por lo que

   $$\lim_{m\to\infty} \| A^m u \|_v = 0. \tag{7.3}$$

   Pero, por (7.2),

   $\lim_{m\to\infty} \| A^m u \|_v = \lim_{m\to\infty} \| \lambda^m u \|_v = \lim_{m\to\infty} |\lambda|^m \| u \|_v$.

Además $|\lambda| \geq 1$ y $\| u \|_v \neq 0$, por ser $u \neq 0$, luego

$$\lim_{m\to\infty} \| A^m u \|_v \neq 0,$$

lo que contradice (7.3). En consecuencia $|\lambda| < 1$ cualquiera que sea el autovalor λ de A, y por tanto $\rho(A) < 1$.

2. Recíprocamente[7], si $\rho(A) < 1$ entonces, en virtud del Teorema 17 y de (7.1), existirá una norma $\| \cdot \|_M$ tal que $\| A \|_M < 1$, ya que en caso contrario para toda norma matricial se tendría $\| A \| \geq 1$, y entonces $\inf_{\|\cdot\|} \| A \| \geq 1$, lo que contradice que $\rho(A) = \inf_{\|\cdot\|} \| A \| < 1$.

   En consecuencia existe $\| \cdot \|_M$ tal que $\| A \|_M < 1$. Y como

   $$\| A^m \|_M \leq \| A \|_M^m,$$

   entonces $\lim_{m\to\infty} \| A^m \| = 0$, de donde $\lim_{m\to\infty} A^m = \theta$.

■

Por último vamos a obtener otro importante resultado relativo a una especialísima sucesión de vectores, la obtenida de la forma que a continuación se detalla.

Sea $x^{(0)}$ un vector inicial arbitrario, y a partir de él vamos a obtener otros vectores, como se indica: $x^{(m+1)} = Ax^{(m)}$, con $m \geq 0$.

Veamos en qué condiciones esta sucesión de vectores converge al vector 0.

TEOREMA 19 *Dada la sucesió $x^{(m+1)} = Ax^{(m)}$, la condición necesaria y suficiente para que converja al vector cero, cualquiera que sea el vector inicial $x^{(0)}$, es que* $\lim_{m\to\infty} A^m = \theta$.

Demostración.

1. Si $\lim_{m\to\infty} A^m = \theta$, entonces se tendrá

   $$\begin{aligned} \lim_{m\to\infty} \| Ax^{(m)} \| &= \lim_{m\to\infty} \| A^m x^{(0)} \| \\ &\leq \lim_{m\to\infty} \| A^m \| \| x^{(0)} \| = 0, \end{aligned}$$

   de donde $\lim_{m\to\infty} Ax^{(m)} = 0$, por lo que $\lim_{m\to\infty} x^{(m+1)} = 0$.

2. Y recíprocamente, como $x^{(m)} = A^m x^{(0)} \to 0$, $m \to \infty$, para cada $x^{(0)}$, entonces $A^m \epsilon_i$ convergerá a 0, siendo ϵ_i el vector de $\mathbb{R}^n$ que tiene todas sus coordenadas nulas salvo la i-ésima que vale 1.

   Consecuentemente, cada columna de A^m, $A^m \epsilon_i$, converge a 0, y de aquí se sigue que $\lim_{m\to\infty} A^m = \theta$.

■

[7] Haremos uso del resultado anteriormente obtenido que indica que el radio espectral de una matriz coincide con el ínfimo de las normas matriciales de la misma. No obstante, puede realizarse una demostración directamente con sólo remitirnos a la forma canónica de Jordan de A y de ella a la de A^m, y teniendo en cuenta que en valor absoluto todos los autovalores son menores que 1, entonces cada célula de Jordan de J^m convergerá a θ, luego J^m también y A^m converge a θ.

7.4 Series de matrices

7.4.1 Consideraciones generales

Dada una sucesión de matrices

$$A^{(1)}, A^{(2)}, \ldots, A^{(m)}, \ldots$$

formamos una nueva sucesión, la sucesión de sumas parciales

$$S_1 = A^{(1)}, S_2 = A^{(1)} + A^{(2)}, \ldots, S_N = \sum_{m=1}^{N} A^{(m)}, \ldots$$

Esta nueva sucesión recibe el nombre de *serie de matrices.*

Si esta sucesión tiene límite, entonces la serie se dice convergente. Si la sucesión diverge, entonces la serie se dice divergente.

Si existe

$$\lim_{N\to\infty} S_N = A$$

escribiremos $A = \sum_{m=1}^{\infty} A^{(m)}$, y la matriz A recibe el nombre de *suma de la serie.*

Desde luego, si la serie $\sum_{m=1}^{\infty} A^{(m)}$ es convergente, entonces necesariamente $\lim_{m\to\infty} A^{(m)} = \theta$.

Para nuestro estudio, hay un tipo de series matriciales que es de especial interés y a las que dedicaremos algunas páginas, se trata las series de potencias de matrices.

7.4.2 Series de potencias de matrices

Sea X una matriz cuadrada real de orden n, y vamos a considerar la serie

$$\sum_{m=0}^{\infty} \frac{1}{z^{m+1}} X^m,$$

donde $z \in \mathbb{R}$. Seguidamente caracterizamos la convergencia de esta serie de potencias.

TEOREMA 20 *La condición necesaria y suficiente para que la serie*

$$\sum_{m=0}^{\infty} \frac{1}{z^{m+1}} X^m$$

converja, es que $\rho(X) < |z|$.

Demostración.

1. Si $\sum_{m=0}^{\infty} \frac{1}{z^{m+1}} X^m = \frac{1}{z}\left[I + \frac{1}{z}X + \frac{1}{z^2}X^2 + \ldots + \frac{1}{z^m}X^m + \ldots\right]$ es convergente, entonces necesariamente

$$\lim_{m\to\infty} \frac{1}{z^m} X^m = \theta.$$

Llamando $\frac{1}{z}X = B$, tenemos que $\lim_{m\to\infty} B^m = \theta$, lo que equivale a decir que $\rho(B) < 1$, de donde $|\lambda_i^{(B)}| < 1$ para cada i, y donde $\lambda_i^{(B)}$ son los autovalores de B.

Pero, si $\lambda_i^{(X)}$ son los autovalores de X, se tiene que

$$\lambda_i^{(B)} = \frac{1}{z}\lambda_i^{(X)},$$

ya que si $Xu = \lambda u$, entonces $(1/z)Xu = (\lambda/z)u$, luego $Bu = (\lambda/z)u$.

Por tanto

$$|\lambda_i^{(X)}| < |z|, \text{ cualquiera que sea } i = 1, \ldots, n.$$

Luego

$$\rho(X) < |z|.$$

2. Recíprocamente, si $\rho(X) < |z|$, entonces

$$|\lambda_i^{(X)}| < |z|, \tag{7.4}$$

cualquiera que sea el autovalor $\lambda_i^{(X)}$ de X.

Realizando el producto

$$\begin{aligned}(zI - X)\textstyle\sum_{m=0}^{k} \frac{1}{z^{m+1}}X^m &= (zI - X)(\tfrac{1}{z}I + \tfrac{1}{z^2}X + \ldots + \tfrac{1}{z^{k+1}}X^k)\\ &= I + \tfrac{1}{z}X + \tfrac{1}{z^2}X^2 + \ldots + \tfrac{1}{z^k}X^k - \tfrac{1}{z}X - \tfrac{1}{z^2}X^2 - \ldots - \tfrac{1}{z^{k+1}}X^{k+1}\\ &= I - \tfrac{1}{z^{k+1}}X^{k+1} = I - B^{k+1}.\end{aligned} \tag{7.5}$$

Ahora bien, los autovalores de $(zI - X)$ son $z - \lambda_i^{(X)}$, entonces por (7.4) se tiene que $z - \lambda_i^{(X)} \neq 0$, por lo que $|zI - X| = \Pi_{i=1}^{n}(z - \lambda_i^{(X)}) \neq 0$, de donde resulta que $(zI - X)$ es invertible.

Por consiguiente, de (7.5), obtendremos que

$$\sum_{m=0}^{k} \frac{1}{z^{m+1}}X^m = (zI - X)^{-1}(I - B^{k+1}).$$

Entonces

$$\begin{aligned}\sum_{m=0}^{\infty} \frac{1}{z^{m+1}}X^m &= \lim_{k\to\infty} \sum_{m=0}^{k} \frac{1}{z^{m+1}}X^m\\ &= \lim_{k\to\infty} \{(zI - X)^{-1}(I - B^{k+1})\}\\ &= (zI - X)^{-1}(I - \lim_{k\to\infty} B^{k+1}).\end{aligned} \tag{7.6}$$

Ahora bien,

$$\lambda_i^{(B)} = \frac{1}{z}\lambda_i^{(X)},$$

luego

$$|\lambda_i^{(B)}| = \frac{1}{|z|}|\lambda_i^{(X)}|.$$

Pero por nuestra hipótesis, (7.4),

$$|\lambda_i^{(B)}| = \frac{1}{|z|}|\lambda_i^{(X)}| < 1.$$

Luego $\rho(B) < 1$, lo que equivale a decir que $\lim_{k\to\infty} B^m = \theta$.

Quedando (7.6), por tanto, como

$$\sum_{m=0}^{\infty} \frac{1}{z^{m+1}} X^m = (zI - X)^{-1}.$$

Lo que nos dice que la serie es convergente y además el valor de su suma.

■

De camino hemos obtenido que la inversa de la matriz $(zI - X)^{-1}$ se puede obtener mediante un desarrollo en serie.

Es de notar que para $z = 1$ existe la inversa de $(I - X)$ si y sólo si $\rho(X) < 1$, ya que en este caso la serie $\sum_{m=0}^{\infty} X^m = (I - X)^{-1}$ es convergente.

Para calcular $(I + X)^{-1}$ no tenemos más que considerar que

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

siempre que esté garantizada la convergencia, es decir $\rho(X) < 1$.

Y para calcular $(A + B)^{-1}$, hacemos

$$(A + B) = A(I + A^{-1}B) = A(I + X), \text{ con } X = A^{-1}B$$

y naturalmente A regular. Si además $\rho(X) = \rho(A^{-1}B) < 1$, podremos calcular la inversa de $(A + B)$ mediante $(A + B)^{-1} = (I + X)^{-1}A^{-1}$.

Acotaciones

Algunas acotaciones son interesantes y de gran utilidad.

Si la norma $\| \cdot \|$ es tal que $\| I \| = 1$, siendo I la matriz identidad, y $\| X \| < 1$ y tenemos la serie $\sum_{m=0}^{\infty} X^m$, se verifica que

$$\| (I - X)^{-1} \| = \| \sum_{m=0}^{\infty} X^m \| = \| \lim_{N\to\infty} \sum_{m=0}^{N} X^m \| = \lim_{N\to\infty} \| \sum_{m=0}^{N} X^m \|$$

$$\leq \lim_{N\to\infty} \sum_{m=0}^{N} \| X \|^m = \sum_{m=0}^{\infty} \| X \|^m,$$

y esta última es una serie geométrica convergente, por ser $\| X \| < 1$.

Luego

$$\| (I - X)^{-1} \| \leq \sum_{m=0}^{\infty} \| X \|^m = \frac{1}{1 - \| X \|}.$$

Análogamente, si se trata ahora de acotar $\| (I+X)^{-1} \|$ en las mismas condiciones que antes, tendremos que

$$\| (I+X)^{-1} \| = \| (I-(-X))^{-1} \| \leq \frac{1}{1-\| -X \|} = \frac{1}{1-\| X \|}.$$

Para acotar $\| (A+B)^{-1} \|$, si A es regular, $\| I \| = 1$ y $\| X \| < 1$, con $X = A^{-1}B$, se tiene que

$$\| (A+B)^{-1} \| = \| (I+X)^{-1}A^{-1} \| \leq \| (I+X)^{-1} \| \| A^{-1} \| \leq \frac{\| A^{-1} \|}{1-\| X \|} =$$

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

por tanto

$$\| (A+B)^{-1} \| \leq \frac{\| A^{-1} \|}{1-\| A^{-1}B \|}. \tag{7.7}$$

7.5 Condicionamiento y perturbaciones

7.5.1 Número de condición de una matriz

En el epígrafe **3.3.5.** se analizaron los inconvenientes del método de Gauss. Especial atención merecieron los sistemas mal condicionados. Obtuvimos una aproximación intuitiva al problema y pusimos de relieve como la inestabilidad que se produce es inherente a estos tipos de sistemas, por lo que cualquier método que se utilice será peligroso.

Estamos ahora en condiciones de abordar con mayor profundidad este problema.

Para tener una estimación del comportamiento del sistema en los cálculos necesarios para resolverlo, se introduce el concepto de número de condición de una matriz A, en relación a la resolución del sistema $Ax = b$.

Aunque posteriormente, dentro de este mismo epígrafe, estudiemos detenida y sistemáticamente las perturbaciones en un sistema de ecuaciones lineales, vamos a considerar seguidamente el caso de que exista un error en el vector de términos independientes b, a fin de justificar el significado del **condicionamiento** de una matriz.

Sea $Ax = b$. Supongamos que hay un error en b, y que es b^* su aproximación; es decir, b^* es el resultado de una cierta perturbación $b^* = b + k$. Obtendremos entonces una aproximación x^* a la solución; es decir, se tendrá también una perturbación en la solución $x^* = x + h$, de forma que

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

y por tanto $Ah = k$, de donde $h = A^{-1}k$, suponiendo, como de costumbre que A es regular.

Consideremos una norma matricial compatible con alguna vectorial. Si no hay posibilidad de confusión omitiremos denotar con los subíndices v y M, respectivamente, a las normas vectorial y matricial. Desde luego

$$\| h \| = \| x^* - x \| \quad \text{y} \quad \| k \| = \| b^* - b \|$$

nos indican los errores cometidos en la solución y en el vector de términos independientes, respectivamente.

Veamos qué relación podemos encontrar entre ellos. Como $h = A^{-1}k$, entonces $\| h \| \leq \| A^{-1} \| \| k \|$. Por otro lado $b = Ax$, de donde $\| b \| \leq \| A \| \| x \|$. Por tanto:

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

de donde se tiene que:

$$\frac{\| x^* - x \|}{\| x \|} \leq \| A \| \| A^{-1} \| \frac{\| b^* - b \|}{\| b \|}.$$

Es decir, el error relativo cometido en la solución es $\| A \| \| A^{-1} \|$ veces, como máximo, el error relativo cometido en b. Por consiguiente, si el número $\| A \| \| A^{-1} \|$ está próximo a 1 el sistema estará bien condicionado (a una pequeña perturbación de b, le corresponde una pequeña perturbación de la solución). Cuanto más se aleje $\| A \| \| A^{-1} \|$ de 1 peor condicionado estará el sistema. Por tanto $\| A \| \| A^{-1} \|$ nos sirve para determinar el grado de condicionamiento del sistema.

Pasamos a definir el condicionamiento de una matriz A regular. Para A invertible[8] (situación de la que hemos partido en todo este estudio de los sistemas $Ax = b$), se define *número de condición*, *Cond (A)*, de la matriz A respecto de una cierta norma matricial dada $\| \, . \, \|_M$ a

$$Cond(A) = \| A \| \| A^{-1} \| .$$

Propiedades de *Cond A*

1. $Cond\ (A) \geq 1$, ya que $Cond\ (A) \geq \| AA^{-1} \| = \| I \| \geq \rho(I) = 1$.
2. $Cond\ (A^{-1}) = Cond\ (A)$, de forma obvia.
3. $Cond\ (\lambda A) = Cond\ (A)$, cualquiera que sea $\lambda \neq 0$, de forma también evidente.

El número de condición depende, obviamente, de la norma utilizada, pero como en los espacios vectoriales de dimensión finita todas las normas son equivalentes, se tendrá que

$$\alpha \| A \| \leq \| A \|^* \leq \beta \| A \|, \tag{7.8}$$

[8] El condicionamiento de A no invertible no se contempla, pues estamos considerando sistemas con matriz regular. No obstante, el condicionamiento de A singular puede ser definido como ∞.

con $\alpha, \beta > 0$, siendo $A \in \mathbb{R}^{n,n}$ una matriz cualquiera y cualesquiera que sean las normas $\| \ \|$ y $\| \ \|^*$.

También tendremos

$$\alpha \| A^{-1} \| \leq \| A^{-1} \|^* \leq \beta \| A^{-1} \| . \tag{7.9}$$

Entonces, multiplicando (7.8) y (7.9)

$$\alpha^2 Cond(A) \leq Cond^*(A) \leq \beta^2 Cond(A).$$

Si la norma utilizada es la espectral $\| A \|_2 = \rho(AA^t)^{1/2}$ se puede demostrar[9] que

$$Cond_2(A) = \frac{\sigma_1}{\sigma_n},$$

siendo σ_1 y σ_n, respectivamente, el máximo y el mínimo valor singular de A, (donde los valores singulares de A son $\sigma_i = +\sqrt{\lambda_i^{(AA^t)}}$).

En el caso particular de que A sea simétrica[10], evidentemente A^tA es simétrica y además resulta definida positiva, ya que $x^tA^tAx = (Ax)^tAx > 0$ cuando $x \neq 0$. Por tanto los autovalores de A^tA son positivos y además son los autovalores de A al cuadrado; o lo que es lo mismo, los autovalores de A son las raices cuadradas positivas de los autovalores de A^tA. Y por consiguiente:

$$\sigma_i^A = +\sqrt{\lambda_i^{(AA^t)}} = +\sqrt{(\lambda_i^A)^2} = \lambda_i^A,$$

es decir, coinciden los autovalores y los valores singulares de A. Y de aquí que

$$Cond_2(A) = \frac{\lambda_1}{\lambda_n},$$

donde λ_1 y λ_n son, respectivamente, los autovalores máximo y mínimo de A.

Si A es ortogonal, $A^t = A^{-1}$, se obtiene $Cond\ _2(A) = 1$. Las matrices con condicionamiento 1 se denominan *perfectamente condicionadas.*

Ahora bien, el condicionamiento de una matriz utiliza A^{-1}, con lo que es conveniente disponer de estimaciones que no hagan uso de $\| A^{-1} \|$. Así, teniendo en cuenta que de $Ay = d$ se tiene que $y = A^{-1}d$, tomando normas vectorial y matricial compatibles, resultará

$$\frac{\| y \|_v}{\| d \|_v} \leq \| A^{-1} \| .$$

Con lo que tomando distintos vectores "y" y sus correspondientes "d", resultan cotas inferiores de $\| A^{-1} \|$.

[9] Ver a este respecto GASCA, M. (1987): Cálculo numérico. Librería Central, Zaragoza.

[10] Vale, en general, para las matrices complejas normales: $A^*A = AA^*$, aunque no sea de nuestro interés aquí.

7.5.2 Perturbación en el término independiente

Anteriormente, al introducir el concepto de condicionamiento de una matriz, obtuvimos una cota superior del error relativo de la solución del sistema, cuando se ha producido una perturbación en el vector de términos independientes. Ahora vamos a proceder a completar aquella con una cota inferior.

Como antes, sea $Ax = b$ y sea una perturbación del término independiente $b + k$, obteniéndose otra en la solución $x + h$, quedando $A(x + h) = b + k$, y por tanto $h = A^{-1}k$.

Completemos la acotación del error, para lo que consideraremos, nuevamente, una norma matricial compatible con alguna vectorial. Como:

$$x = A^{-1}b, \text{ entonces } \| x \|_v \leq \| A^{-1} \|_M \| b \|_v, \tag{7.10}$$

$$k = Ah, \text{ entonces } \| k \|_v \leq \| A \|_M \| h \|_v . \tag{7.11}$$

Denotemos por $\frac{\|k\|}{\|b\|}$ la perturbación relativa o error relativo del término independiente y por $\frac{\|h\|}{\|x\|}$ el error relativo de la solución.

En consecuencia,

$$\begin{aligned} \frac{\|h\|}{\|x\|} = \frac{\|x^* - x\|}{\|x\|} &\geq \frac{\|h\|}{\|A^{-1}\| \, \|b\|} \\ &= \frac{\|A\| \, \|h\|}{\|A\| \, \|A^{-1}\| \, \|b\|} \geq \frac{\|k\|}{\|A\| \, \|A^{-1}\| \, \|b\|}, \end{aligned}$$

desigualdades resultantes de aplicar en cada desigualdad obtenida, respectivamente y de forma sucesiva, las desigualdades (7.10) y (7.11). La expresión anterior podemos expresarla en función de $Cond\,(A)$, en la forma siguiente

$$\frac{\| h \|}{\| x \|} = \frac{\| x^* - x \|}{\| x \|} \geq \frac{\| k \|}{Cond(A) \, \| b \|}.$$

■

Hemos obtenido el siguiente resultado.

TEOREMA 21 *Si dado un sistema $Ax = b$, se tiene una perturbación en el término independiente $b + k$, entonces el error relativo de la solución verifica la doble acotación:*

$$\frac{1}{Cond(A)} \frac{\| k \|}{\| b \|} \leq \frac{\| h \|}{\| x \|} \leq Cond(A) \frac{\| k \|}{\| b \|}.$$

En el caso particular de que $Cond(A) = 1$, entonces los errores relativos en el término independiente y en la solución coinciden

$$\frac{\| k \|}{\| b \|} = \frac{\| h \|}{\| x \|}$$

y estaremos ante un sistema bien condicionado.

Cuanto más próximo esté $Cond(A)$ a 1, mejor será. Cuando $Cond(A)$ es muy grande, significa que hay una amplia región de incertidumbre sobre la solución, por lo que los resultados obtenidos serán poco fiables.

7.5.3 Perturbación en la matriz del sistema

Si efectuamos una perturbación de la matriz del sistema, en la forma $A+H$, podemos escribir

$$(A+H) = A(I + A^{-1}H),$$

siempre que A sea invertible. Y como ya sabemos, del apartado **7.4.2**, $(A+H)$ será invertible si y sólo si $\rho(A^{-1}H) < 1$. Y en el caso en que $\| I \| = 1$ y $\| A^{-1}H \| < 1$ entonces se tiene la acotación (7.7) obtenida en el epígrafe citado, que es

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

Y desde luego, una condición suficiente para la invertibilidad o regularidad de $(A+H)$ es que $\| A^{-1}H \| \leq \| A^{-1} \| \| H \| < 1$, derivada del hecho de que para cualquier norma matricial se tiene $\rho(A^{-1}H) \leq \| A^{-1}H \|$.

Realizada la perturbación en la matriz del sistema tendremos también una perturbación en la solución del sistema, de acuerdo con

$$(A+H)(x+h) = b,$$

por lo que obtendremos $Hx + Ah + Hh = \theta$, de donde $(A+H)h = -Hx$ y por tanto $A(I + A^{-1}H)h = -Hx$.

Suponiendo $\| A^{-1}H \| \leq \| A^{-1} \| \| H \| < 1$ y, por tanto, que $(A+H)$ es invertible, tendremos que

$$h = -(I + A^{-1}H)^{-1}A^{-1}Hx.$$

Veamos una acotación del error

$$\| h \| \leq \| (I + A^{-1}H)^{-1} \| \| A^{-1} \| \| H \| \| x \|$$

suponiendo además que $\| I \| = 1$, podremos acotar $\| (I + A^{-1}H)^{-1} \|$, superiormente, por $\frac{1}{1-\|A^{-1}H\|}$, obteniéndose

$$\| h \| \leq \frac{1}{1 - \| A^{-1}H \|} \| A^{-1} \| \| H \| \| x \| \leq \frac{\| A^{-1} \| \| H \| \| x \|}{1 - \| A^{-1} \| \| H \|}.$$

Dividiendo por $\| x \|$,

$$0 \leq \frac{\| h \|}{\| x \|} \leq \frac{\| A^{-1} \| \| H \|}{1 - \| A^{-1} \| \| H \|} = \frac{\| A^{-1} \| \| A \| \frac{\|H\|}{\|A\|}}{1 - \| A^{-1} \| \| A \| \frac{\|H\|}{\|A\|}},$$

con lo que

$$0 \leq \frac{\| h \|}{\| x \|} \leq \frac{Cond(A)\frac{\|H\|}{\|A\|}}{1 - Cond(A)\frac{\|H\|}{\|A\|}},$$

donde $\frac{\|h\|}{\|x\|}$ y $\frac{\|H\|}{\|A\|}$ son, respectivamente, los errores relativos de la solución y de la matriz del sistema.

Hemos obtenido el siguiente resultado.

TEOREMA 22 *Si dado un sistema $Ax = b$, se tiene una perturbación en la matriz del sistema $A+H$, entonces el error relativo de la solución verifica la acotación:*

$$0 \leq \frac{\| h \|}{\| x \|} \leq \frac{Cond(A)\frac{\|H\|}{\|A\|}}{1 - Cond(A)\frac{\|H\|}{\|A\|}},$$

donde las normas verifican las condiciones señaladas anteriormente.

7.5.4 Perturbación total

Si se realizan simultáneamente perturbaciones en el término independiente y en la matriz del sistema, la solución quedará alterada de la siguiente forma

$$(A + H)(x + h) = b + k, \text{ luego } Hx + (A + H)h = k,$$

suponiendo, como de costumbre, la condición suficiente de invertibilidad

$$\| A^{-1}H \| \leq \| A^{-1} \| \| H \| < 1$$

de la matriz $(A + H)$, tendremos $(A + H)h = k - Hx$. Por tanto

$$h = (A + H)^{-1}(k - Hx) = (I + A^{-1}H)^{-1}A^{-1}(k - Hx).$$

Buscamos una acotación del error relativo suponiendo, como en ocasiones anteriores, que $\| A^{-1}H \| \leq \| A^{-1} \| \| H \| < 1$ y que $\| I \| = 1$.

$$\begin{aligned} \| h \| &\leq \| (I + A^{-1}H)^{-1} \| \| A^{-1} \| \| (k - Hx) \| \\ &\leq \| (I + A^{-1}H)^{-1} \| \| A^{-1} \| (\| k \| + \| H \| \| x \|) \\ &\leq \frac{1}{1-\|A^{-1}H\|} \| A^{-1} \| (\| k \| + \| H \| \| x \|), \end{aligned}$$

dividiendo por $\| x \|$ resulta

$$\begin{aligned} 0 \leq \frac{\|h\|}{\|x\|} &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}H\|}(\frac{\|k\|}{\|x\|} + \| H \|) \\ &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|H\|}(\frac{\|k\|}{\|x\|} + \| H \|) \\ &= \frac{\|A^{-1}\|\frac{1}{\|A\|}\|A\|}{1-\|A^{-1}\|\|A\|\frac{\|H\|}{\|A\|}}(\frac{\|k\|}{\|x\|} + \| H \|) \\ &= \frac{\|A^{-1}\|\|A\|}{1-\|A^{-1}\|\|A\|\frac{\|H\|}{\|A\|}}(\frac{\|k\|}{\|A\|\|x\|} + \frac{\|H\|}{\|A\|}) \\ &\leq \frac{Cond(A)}{1-Cond(A)\frac{\|H\|}{\|A\|}}(\frac{\|k\|}{\|b\|} + \frac{\|H\|}{\|A\|}), \end{aligned}$$

donde $\frac{\|h\|}{\|x\|}$, $\frac{\|H\|}{\|A\|}$ y $\frac{\|k\|}{\|b\|}$ son, respectivamente, los errores relativos de la solución, de la matriz del sistema y del término independiente.

Hemos obtenido por tanto el siguiente resultado:

TEOREMA 23 *Si dado un sistema $Ax = b$, se tiene una perturbación en la matriz del sistema $A+H$ y en el término independiente, entonces el error relativo de la solución verifica la acotación:*

$$0 \leq \frac{\| h \|}{\| x \|} \leq \frac{Cond(A)}{1 - Cond(A)\frac{\|H\|}{\|A\|}}(\frac{\| k \|}{\| b \|} + \frac{\| H \|}{\| A \|})$$

donde las normas verifican las condiciones señaladas anteriormente.

Puesto que con el número de condición obtenemos una "medida" del nivel de confianza que podemos depositar en los resultados obtenidos en la resolución de nuestro problema, si $Cond(A)$ crece no podemos tener confianza en la solución, por encontrarnos con una región de incertidumbre muy grande, donde no se sabe lo que ocurre, independientemente de que casualmente la solución fuese buena.

7.5.5 Error y correlación residuales

Otra interpretación del condicionamiento es la que seguidamente exponemos.

Si resolviendo $Ax = b$ se obtiene x^*, solución aproximada, a la diferencia $Ax^* - b = k$ se le llama *residuo* o *defecto*.

$$Ax^* = A(x + h) = b + k, \text{ por lo que } k = Ax^* - b,$$

que es la perturbación del término independiente, siendo $h = x^* - x$. Haciendo $Ax^* = b^*$ y, por tanto, $b^* - b = k$.

La primera de las acotaciones obtenidas

$$\frac{1}{Cond(A)} \frac{\| k \|}{\| b \|} \leq \frac{\| h \|}{\| x \|} \leq Cond(A) \frac{\| k \|}{\| b \|}$$

nos indica que si el condicionamiento $Cond(A)$ es grande, entonces un pequeño residuo k no equivale a un resultado satisfactorio.

El hecho de que la solución de un sistema sea más o menos aceptable no depende del error relativo residual (del error relativo del término independiente), en contra de lo que aparentemente (a primera vista) pudiera pensarse, sino que dependerá del condicionamiento de la matriz del sistema. Los siguientes ejemplos procuran aclarar esta cuestión.

Ejemplo 14:

Consideremos el sistema de matriz y término independiente, respectivamente,

$$A = \begin{bmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{bmatrix} \text{ y } b = \begin{bmatrix} 0.8642 \\ 0.1440 \end{bmatrix},$$

del que se ha obtenido como solución aproximada

$$x^* = x + h = \begin{bmatrix} 0.9911 \\ -0.4870 \end{bmatrix}.$$

La solución exacta es

$$x = \begin{bmatrix} 2 \\ -2 \end{bmatrix}.$$

Entonces, el residuo

$$k = Ax^* - b = \begin{bmatrix} 10^{-8} \\ 10^{-8} \end{bmatrix}$$

es muy pequeño y el error relativo residual también muy pequeño

$$\frac{\| k \|}{\| b \|} = 1.15713 \cdot 10^{-8}.$$

La solución aproximada aparentemente (respecto del error relativo residual) es satisfactoria.

Pero

$$A^{-1} = \begin{bmatrix} 0.1441 & -0.8161 \\ -0.2161 & 1.2969 \end{bmatrix} \cdot 10^8.$$

Por otra parte $\| A \|_\infty = 2.1617$ y $\| A^{-1} \|_\infty = 1.513 \cdot 10^8$, de donde

$$Cond(A) = 3.27065 \cdot 10^8,$$

lo que nos da un condicionamiento muy alto.

Examinando el error relativo de la solución, tenemos

$$\frac{\| h \|}{\| x \|} = \frac{\| x^* - x \|}{\| x \|} = 0.7565,$$

que resulta considerablemente grande.

Si ahora examinamos la acotación

$$\frac{1}{Cond(A)} \frac{\| k \|}{\| b \|} \leq \frac{\| h \|}{\| x \|} \leq Cond(A) \frac{\| k \|}{\| b \|}$$

tenemos

$$0 \leq \frac{\| h \|}{\| x \|} \leq 3.7846.$$

Así pues, un residuo pequeño no equivale a un resultado satisfactorio, toda vez que la matriz está mal condicionada. □

También nos puede ocurrir que el residuo sea grande y sin embargo la solución aproximada sea bastante aceptable.

Ejemplo 15:

Consideremos ahora este otro sistema de matriz y término independiente, respectivamente,

$$A = \begin{bmatrix} 1 & 1.001 \\ 1 & 1 \end{bmatrix} \text{ y } b = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

del que se ha obtenido como solución aproximada

$$x^* = x + h = \begin{bmatrix} -1001 \\ 1000 \end{bmatrix}.$$

La solución exacta es

$$x = \begin{bmatrix} -1000 \\ 1000 \end{bmatrix}.$$

Entonces, el residuo es

$$k = Ax^* - b = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

y el error relativo residual es muy grande

$$\frac{\| k \|}{\| b \|} = 1.$$

La solución aproximada aparentemente (respecto del error relativo residual) no es satisfactoria.

Pero

$$A^{-1} = \begin{bmatrix} -1000 & 1001 \\ 1000 & -1001 \end{bmatrix}.$$

Por otra parte $\| A \|_\infty = 2.001$ y $\| A^{-1} \|_\infty = 2.001 \cdot 10^3$, de donde

$$Cond(A) = 4.004 \cdot 10^3,$$

lo que nos da un condicionamiento muy alto.

Examinando el error relativo de la solución, tenemos

$$\frac{\| h \|}{\| x \|} = \frac{\| x^* - x \|}{\| x \|} = 0.001,$$

que resulta considerablemente pequeño; es decir, que la solución es bastante aceptable.

Si ahora examinamos la acotación

$$\frac{1}{Cond(A)} \frac{\| k \|}{\| b \|} \leq \frac{\| h \|}{\| x \|} \leq Cond(A) \frac{\| k \|}{\| b \|},$$

tenemos $0 \leq \frac{\|h\|}{\|x\|} \leq 4004$.

Así pues, un residuo grande no equivale a un resultado insatisfactorio, toda vez que la matriz está mal condicionada. □

Se puede mejorar la precisión de una solución obtenida en la resolución de un sistema $Ax = b$ si las cotas de error obtenidas no son satisfactorias.

Así, si x^* es la solución obtenida y $k = Ax^* - b$ es su residuo, tendremos $x = x^* - h$, pudiéndose calcular h resolviendo $Ah = k$. Para ganar precisión en el proceso el residuo k debería ser calculado con doble precisión.

El algoritmo sería:

1. Calcular $k = Ax^* - b$ con doble precisión.
2. Resolver $Ah = k$ y llamar h^* a la solución obtenida (aproximación de la real).
3. Llamar $x^{*'} = x^* - h^*$ y calcular una cota del error. Si no es satisfactoria: volver a 1.

Si, por ejemplo, se ha utilizado el método de Gauss, la resolución del sistema del paso 2 no requiere más que unos pocos cálculos, toda vez que la matriz A ya ha sido triangularizada.

Por otra parte, teniendo en cuenta la acotación del error cuando se perturba el término independiente, obtenida anteriormente,

$$\frac{1}{Cond(A)} \frac{\| k \|}{\| b \|} \leq \frac{\| h \|}{\| x \|} \leq Cond(A) \frac{\| k \|}{\| b \|},$$

que puede escribirse como

$$\frac{1}{Cond(A)}\frac{\| k \|}{\| b \|} \leq \frac{\| x^* - x \|}{\| x \|} \leq Cond(A)\frac{\| k \|}{\| b \|}.$$

Como $k^* = Ax^{*'} - b = A(x^* - h^*) - b = Ax^* - Ah^* - b$ y como $k = Ax^* - b$, tendremos que $k^* = k - Ah^*$, con lo que k^* puede interpretarse como el opuesto al residuo de la resolución del paso 2.

7.5.6 Obtención de la solución de un sistema modificado a partir del original

A veces A se debe cambiar mínimamente, bien porque cambie exclusivamente una variable, bien porque lo haga una ecuación. Se trata de valerse del sistema inicial con la matriz A, a fin de reducir el coste en número de operaciones al resolver el nuevo sistema con una matriz A' ligeramente modificada.

Si la columna c_i de A, se ha de cambiar por c'_i, ello puede hacerse mediante

$$A + (c'_i - c_i)e_i^t$$

ya que

$$(c'_i - c_i)e_i^t = \begin{bmatrix} a'_{1i} - a_{1i} \\ a'_{2i} - a_{2i} \\ \vdots \\ a'_{ni} - a_{ni} \end{bmatrix} [0, \ldots, 1, \ldots, 0] = \begin{bmatrix} 0 & \ldots & a'_{1i} - a_{1i} & \ldots & 0 \\ 0 & \ldots & a'_{2i} - a_{2i} & \ldots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \ldots & a'_{ni} - a_{ni} & \ldots & 0 \end{bmatrix}.$$

De la misma forma si se desea cambiar la fila f_i por la f'_i, se puede hacer como $A + e_i(f'_i - f_i)^t$, ya que

$$e_i(f'_i - f_i)^t = \begin{bmatrix} 0 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ a'_{i1} - a_{i1} & \ldots & a'_{in} - a_{in} \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 0 \end{bmatrix}.$$

De forma más general podría pensarse en unas alteraciones mayores y considerar la matriz $A + C \cdot F$ tal que C es una matriz de orden $n \times m$, con sólo algunas columnas no nulas, y F de orden $m \times n$ se trata de una matriz con sólo algunas filas no nulas. Y siempre $m \ll n$ (m mucho menor que n). Donde la alteración H de A se describe mediante el producto de dos matrices C y F.

El resultado que se obtendrá seguidamente describe la inversa de

$$A' = A + C \cdot F$$

y cómo obtener la solución del sistema modificado $A'x' = b$ a partir de la solución de $Ax = b$.

Tenemos que $A' = A + CF = A(I + A^{-1}CF)$. Si A' es regular será

$$\begin{aligned} A'^{-1} = (A + CF)^{-1} &= [A(I + A^{-1}CF)]^{-1} \\ &= (I + A^{-1}CF)^{-1}A^{-1} \\ &= (I + D)^{-1}A^{-1}, \end{aligned}$$

con $D = A^{-1}CF$. Y tendremos

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

$$= A^{-1} - A^{-1}CFA^{-1} + A^{-1}CFA^{-1}CFA^{-1} - A^{-1}CFA^{-1}CFA^{-1}CFA^{-1} + \ldots =$$

$$= A^{-1} - A^{-1}(CF - CFA^{-1}CF + CFA^{-1}CFA^{-1}CF - \ldots)A^{-1} =$$

$$= A^{-1} - A^{-1}C(I - FA^{-1}C + FA^{-1}CFA^{-1}C - \ldots)FA^{-1} =$$

$$= A^{-1} - A^{-1}C(I + FA^{-1}C)^{-1}FA^{-1}.$$

Por otra parte $A'x' = b$, entonces

$$\begin{aligned} x' = A'^{-1}b &= A'^{-1}Ax \\ &= (A^{-1} - A^{-1}C(I + FA^{-1}C)^{-1}FA^{-1})Ax \\ &= x - A^{-1}C(I + FA^{-1}C)^{-1}Fx. \end{aligned}$$

Luego se ha obtenido el siguiente resultado.

TEOREMA 24 *Si $K = I + FA^{-1}C$, que es una matriz $m \times m$, es regular y $\rho(A^{-1}CF) < 1$, entonces:*

1. *$A' = A + C \cdot F$ es regular y $A'^{-1} = A^{-1} - A^{-1}CK^{-1}FA^{-1}$.*
2. *$x' = x - A^{-1}CK^{-1}Fx$.*

Con lo que se ha resuelto el segundo sistema aprovechando la resolución del primero, con el considerable ahorro en número de operaciones.

A primera vista puede parecer que la aplicación de este resultado raras veces aparecerá en la práctica, pero esto no es así. Muy al contrario, modelos extremadamente grandes de sistemas complicados, como grandes redes de distribución eléctrica, grandes sistemas económicos, etc., consisten en un número pequeño de subsistemas muy grandes y casi independientes, con un reducido número de interconexiones. El resultado anterior modela esta situación, con A representando la situación de independencia de los subsistemas, C y F los ajustes necesarios para las interconexiones.

Por ejemplo, con sólo dos subsistemas, se podría abordar un sistema de 2000 ecuaciones con 2000 incógnitas, con una matriz A' celular

$$A' = \begin{bmatrix} B & \theta \\ \theta & R \end{bmatrix} + \begin{bmatrix} \theta & \theta & E & \theta \\ \theta & S & \theta & \theta \end{bmatrix}$$

siendo B y R 1000×1000, E y S 1000×10.

Supongamos que A es la primera de las dos matrices celulares y B y R son regulares. Se interpreta B como la representación de las relaciones dentro de un subsistema grande, R como la que representa las relaciones dentro de otro gran subsistema, y E y S como la representación de las interconexiones entre los dos subsistemas.

Haciendo $A' = A + CF$, donde

$$A = \begin{bmatrix} B & \theta \\ \theta & R \end{bmatrix}, \quad C = \begin{bmatrix} \theta & E \\ S & \theta \end{bmatrix} \quad \text{y} \quad R = \begin{bmatrix} \theta & I & \theta & \theta \\ \theta & \theta & I & \theta \end{bmatrix}.$$

En estas condiciones al hacer uso de los resultados del último teorema garantizamos un ahorro enorme en el cómputo.

Se han desarrollado varias técnicas especiales basadas esencialmente en este método, con una amplia variedad de áreas de aplicación.

Esta idea básica lleva, por ejemplo, al método de rompimiento en análisis de redes, a métodos de matrices de capacitancia para la solución numérica directa de ciertas ecuaciones diferenciales, a varios métodos modernos para la optimización de la programación no lineal restringida y no restringida, y a ciertas implementaciones del método del simplex en la programación lineal.

8. Métodos iterativos usuales

Hasta ahora hemos visto métodos para resolver sistemas de ecuaciones lineales, de forma directa. En el análisis numérico, esta forma de proceder es más la excepción que la regla. Y desde luego en el caso que nos ocupa, los métodos directos son útiles para valores de n no muy grandes. Para computadores pequeños el límite superior práctico es del orden de 250; para sistemas de orden superior las exigencias de memoria y la cantidad de operaciones hacen que los métodos directos sean lentos, costosos e incluso inviables.

Pero hay problemas, como por ejemplo la resolución aproximada de ecuaciones diferenciales, que dan lugar a sistemas con gran número de incógnitas. En problemas físicos o técnicos relacionados con la resolución de ecuaciones en derivadas parciales, pueden aparecer sistemas de orden 10^3 e incluso más. Y aunque es frecuente que el tipo de matrices que aparecen sean "ralas", y también que los elementos a_{ij} no nulos puedan obtenerse a través de una relación sencilla que los genere[1], el propio volumen de datos, incluso para los métodos directos menos costosos, hace que la acumulación de errores de redondeo los haga inadecuados.

Por eso tienen mucho interés los métodos iterativos, en los que se construye una sucesión de vectores $x^{(m)}$ tal que

$$\lim_{m\to\infty} x^{(m)} = x,$$

siendo x la solución del sistema $Ax = b$. Así que se llega a la solución mediante una sucesión de aproximaciones. Si el proceso funciona, dicha sucesión convergerá a la respuesta correcta, en el sentido de que cada término (resultado de cada iteración) de la sucesión será una mejor aproximación a la respuesta que el término precedente.

La forma más frecuente de construir la sucesión consiste en partir de un $x^{(0)}$ arbitrario y luego tomar

$$x^{(m+1)} = F(x^{(m)}), \text{ con } m = 0, 1, 2, \ldots .$$

Lógicamente el problema principal es saber bajo qué condiciones se puede tomar el vector inicial $x^{(0)}$ arbitrariamente, de manera que quede garantizada la convergencia. La segunda cuestión es qué función F tomar para efectuar las iteraciones.

Para este tipo de problemas la función vectorial F suele tomarse lineal; es decir, del tipo

$$F(x) = Bx + c,$$

[1] Recordemos que en el caso de matrices "ralas" existen técnicas de almacenamiento que reducen las necesidades de memoria.

siendo B una matriz cuadrada de orden n y c un vector de $\mathbb{R}^n$. Y veremos como el problema de la convergencia está íntimamente ligado con el de las características intrínsecas de la matriz del sistema.

Nosotros vamos a estudiar los métodos del tipo

$$x^{(m+1)} = Bx^{(m)} + c$$

y vamos a ver las elecciones más adecuadas de B y c.

Comenzaremos por analizar los algoritmos de los métodos iterativos clásicos, para pasar posteriormente en otros epígrafes al estudio de la construcción general y de la convergencia de los métodos lineales.

Existen dos técnicas iterativas empleadas comúnmente para resolver sistemas de ecuaciones de la forma $Ax = b$, son los métodos de Jacobi y de Gauss-Seidel. Si la matriz A es "rala", las técnicas iterativas son las que a menudo dan los mejores resultados con un menor esfuerzo. Sin embargo cada método puede no converger. Después de describirlos se examinarán algunas condiciones en las que los métodos siempre convergen. En lo que sigue, como de costumbre, se supondrá que A es regular.

8.1 Método de Jacobi

Supongamos que el sistema $Ax = b$ es tal que $a_{ii} \neq 0$, para cada i, cosa que siempre puede hacerse ya que al ser A regular siempre se puede transformar en otra A' con elementos diagonales no nulos.

8.1.1 Algoritmo

Reescribimos el sistema de forma que en la i-ésima ecuación x_i quede escrita en términos de las demás variables, lo que intuitivamente significa "despejar la diagonal".

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j \right), \qquad i = 1, 2, \ldots, n.$$

Se escoge de manera arbitraria[2] una aproximación inicial de la solución $x^{(0)}$. Las componentes de este vector se sustituyen en el lado derecho de las ecuaciones, con el fin de obtener una nueva aproximación $x^{(1)}$. Los valores obtenidos así se utilizan para calcular $x^{(2)}$. Se continúa de esta forma generando la sucesión de término general $x^{(m)}$. Es decir, las iteraciones se definen de la siguiente forma:

$$x_i^{(m+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(m)} \right), \quad i = 1, 2, \ldots, n, \quad m \geq 0.$$

[2] Debería escogerse un $x^{(0)}$ que garantizara la convergencia, ya que la misma puede depender del $x^{(0)}$ inicial. Pero en general las condiciones para efectuar una elección adecuada son extremadamente difíciles de comprobar.

Ejemplo 16:

Consideremos el sistema de matriz ampliada

$$\left[\begin{array}{ccc|c} 4.4 & -2.3 & 0.7 & -7.43 \\ 0.8 & 2.5 & 1.1 & 12.17 \\ -1.6 & 0.4 & -5.2 & 26.12 \end{array}\right].$$

Efectuamos los cálculos con cinco dígitos significativos.

1. Escribimos el sistema de manera que en la i-ésima ecuación, x_i quede escrita en términos de las restantes variables

$$x_1 = -\frac{7.43}{4.4} + \frac{2.3}{4.4}x_2 - \frac{0.7}{4.4}x_3 = -1.6886 + 0.52273x_2 - 0.15909x_3,$$

$$x_2 = \frac{12.17}{2.5} - \frac{0.8}{2.5}x_1 - \frac{1.1}{2.5}x_3 = 4.868 - 0.32x_1 - 0.44x_3,$$

$$x_3 = -\frac{26.12}{5.2} - \frac{1.6}{5.2}x_1 + \frac{0.4}{5.2}x_2 = -5.0231 - 0.30769x_1 + 0.076923x_2.$$

2. Se elige un valor inicial arbitrario como primera aproximación. Por ejemplo $x^{(0)} = (0, 0, 0)^t$.

3. Estos valores iniciales se sustituyen en las ecuaciones anteriores, al objeto de obtener una primera aproximación y obtenemos:

$$x_1^{(1)} = -1.6886,\ x_2^{(1)} = 4.868,\ x_3^{(1)} = -5.0231.$$

4. Los valores obtenidos en el paso 3. se emplean para calcular $x^{(2)}$ y así sucesivamente, previa sustitución en las ecuaciones anteriores, obtenemos:

$$x_1^{(2)} = -1.6886 + 0.52273(4.868) - 0.15909(-5.0231) = 1.6552,$$

$$x_2^{(2)} = 4.868 - 0.32(-1.6886) - 0.44(-5.0231) = 7.6185,$$

$$x_3^{(2)} = -5.0231 - 0.30769(-1.6886) + 0.076923(4.868) = -4.1291.$$

Repitiendo el proceso se obtiene:

Iteración	$x_1^{(m)}$	$x_2^{(m)}$	$x_3^{(m)}$
0	0.0	0.0	0.0
1	-1.6886	4.868	-5.0231
2	1.6552	7.6185	-4.1291
3	2.9507	6.1551	-4.9464
4	2.3158	6.1002	-5.4575
5	2.3684	6.5282	-5.2664
6	2.5617	6.4273	-5.2497
7	2.5063	6.3581	-5.3169
8	2.4808	6.4054	-5.3052
9	2.5037	6.4084	-5.2937
10	2.5034	6.3960	-5.3005
11	2.4980	6.3991	-5.3014
12	2.4998	6.4013	-5.2995
13	2.5006	6.3998	-5.2999
14	2.4999	6.3998	-5.3002
15	2.5000	6.4001	-5.3000.

Las sucesiones convergen hacia $x_1 = 2.5, x_2 = 6.4$ y $x_3 = -5.3$, que constituyen la solución como puede verificarse mediante sustitución directa.

8.1.2 Convergencia

Desde luego, para que el método sea efectivo, debemos dar condiciones bajo las que se asegure la convergencia de las aproximaciones $x^{(m)}$ a la solución correcta.

Consideramos los errores en cada iteración, dados por $e^{(m)} = x^{(m)} - x$, donde cada componente del vector error en la m-ésima iteración viene dada por

$$\begin{aligned} e_i^{(m+1)} &= \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j^{(m)}\right) - \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j\right) = \\ &= -\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{a_{ij}}{a_{ii}} e_j^{(m)}, \qquad i = 1, 2, \ldots, n \text{ y } m \geq 0. \end{aligned}$$

Podemos poner $e^{(m+1)} = Be^{(m)}$, $m \geq 0$, con

$$B = -\begin{bmatrix} 0 & a_{12}/a_{11} & a_{13}/a_{11} & \cdots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 0 & a_{23}/a_{22} & \cdots & a_{2n}/a_{22} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & a_{n3}/a_{nn} & \cdots & 0 \end{bmatrix}.$$

Por ello, se tendrá finalmente que

$$e^{(m+1)} = B^{m+1}e^{(0)}.$$

Por lo que si queremos que $x^{(m)}$ converja a x, necesariamente $e^{(m)}$ deberá converger a cero

$$\lim_{m\to\infty} e^{(m)} = 0,$$

para lo que es condición necesaria y suficiente (Teorema 19) que

$$\lim_{m\to\infty} B^m = \theta$$

o equivalentemente (Teorema 18) que

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

Como bien sabemos, es condición suficiente que para alguna norma matricial se verifique que $\| B \| < 1$.

8.2 Método de Gauss-Seidel

En el método de Jacobi, para calcular el valor mejorado $x_i^{(m+1)}$, se utilizan las aproximaciones $x_j^{(m)}$, con $j = 1, \ldots, n$, $j \neq i$. Sin embargo ya se han obtenido los valores mejorados anteriores a él, es decir $x_j^{(m+1)}$, recorriendo j los valores $j = 1, 2, \ldots, i-1$. Podríamos utilizar estos valores más próximos a los reales, en lugar de los correspondientes $x_j^{(m)}$ con $j = 1, 2, \ldots, i-1$.

El resultado de esta observación es el método de Gauss-Seidel.

8.2.1 Algoritmo

Como en el caso del método de Jacobi se parte de un valor inicial $x^{(0)}$ arbitrario, pero, en cada paso, una vez calculado un valor mejorado de una componente se utiliza dicho valor en el cálculo del valor mejorado de la siguiente.

Ejemplo 17:

En el caso del ejemplo anterior, partimos nuevamente de $x^{(0)} = (0,0,0)^t$. Entonces, al igual que antes, $x_1^{(1)} = -1.6886$.

Pero a continuación utilizamos ya este valor para calcular $x_2^{(1)}$:

$$x_2^{(1)} = 4.868 - 0.32x_1^{(1)} - 0.44x_3^{(0)} = 4.868 - 0.32(-1.6886) = 5.4084.$$

Y ahora para calcular $x_3^{(1)}$ se utilizan los valores mejorados de las dos componentes anteriores:

$$\begin{aligned} x_3^{(1)} &= -5.0231 - 0.30769x_1^{(1)} + 0.076923x_2^{(1)} = \\ &= -5.0231 - 0.30769(-1.6886) + 0.076923(5.4084) = -4.0875. \end{aligned}$$

Procediendo siempre de esta forma, haciendo uso siempre de las aproximaciones más recientes, se obtiene:

Iteración	$x_1^{(m)}$	$x_2^{(m)}$	$x_3^{(m)}$
0	0.0	0.0	0.0
1	-1.6886	5.4084	-4.0875
2	1.7888	6.0941	-5.1047
3	2.3091	6.3752	-5.2432
4	2.4780	6.3820	-5.2946
5	2.4898	6.4009	-5.2968
6	2.5000	6.3986	-5.3001
7	2.4993	6.4003	-5.2998
8	2.5002	6.3998	-5.3001
9	2.5000	6.4000	-5.3000
10	2.5000	6.4000	-5.3000 .

Hemos llegado nuevamente a la misma solución, pero más rápidamente. □

Este tema de la rapidez de la convergencia será tratado más tarde.

En general las iteraciones adoptan la forma

$$x_i^{(m+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(m)} \right), \; i = 1, 2, \ldots, n, \quad m \geq 0.$$

El hecho de utilizar cada componente mejorada inmediatamente en el uso de las siguientes, es también más conveniente desde el punto de vista del almacenamiento, puesto que el nuevo valor puede ser inmediatamente guardado en la localización del antiguo valor, lo que minimiza las necesidades de disponibilidad de memoria.

8.2.2 Convergencia

Como antes, analizaremos qué ocurre con el error

$$e^{(m+1)} = x^{(m+1)} - x,$$

con lo que se tiene

$$e_i^{(m+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} e_j^{(m+1)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} e_j^{(m)}, \quad i = 1, 2, \ldots, n,$$

que puede escribirse matricialmente como

$$M e^{(m+1)} = N e^{(m)},$$

con

$$M = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21}/a_{22} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & a_{n3}/a_{nn} & \cdots & 1 \end{bmatrix}$$

y

$$N = -\begin{bmatrix} 0 & a_{12}/a_{11} & a_{13}/a_{11} & \cdots & a_{1n}/a_{11} \\ 0 & 0 & a_{23}/a_{22} & \cdots & a_{2n}/a_{22} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & 0 \end{bmatrix},$$

o bien

$$e^{(m+1)}) = M^{-1} N\, e^{(m)} = B\, e^{(m)}.$$

Por lo tanto

$$e^{(m+1)} = B^{m+1}\, e^{(0)}.$$

Y al igual que antes el método convergerá independientemente del valor inicial $x^{(0)}$ tomado, si y sólo si $\rho(B) < 1$.

En general, aunque no siempre, el método de Gauss-Seidel es más eficiente que el de Jacobi, en el sentido de que cuando hay convergencia en ambos, es más rápida en el primer caso. No obstante puede ocurrir que converja el método de Gauss-Seidel pero no el de Jacobi y al revés. Aunque más adelante analizaremos el problema de la convergencia, veamos seguidamente un ejemplo para ilustrar lo que anteriormente hemos afirmado; en concreto se trata de un sistema que converge con el método de Jacobi y sin embargo diverge con el de Gauss-Seidel.

Ejemplo 18:

Consideremos un sistema cuya matriz es

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

Esta matriz la podemos descomponer en la forma $A = D - L - U$, donde

$$D = \begin{bmatrix} 1 & & \\ & 1 & \\ & & -3 \end{bmatrix}, \; L = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -1 & -2 & 0 \end{bmatrix} \text{ y } U = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Como veremos más adelante, en el epígrafe **9.4**, en el método de Jacobi se verifica que $B = D^{-1}(L+U)$, mientras que $B = (D-L)^{-1}U$ en el método de Gauss-Seidel. Por lo que el primero convergerá (independientemente del vector $x^{(0)}$), si y sólo si $\rho(D^{-1}(L+U)) < 1$, y el segundo lo hará si y sólo si $\rho((D-L)^{-1}U) < 1$.

Ahora bien,

$$D^{-1} = \begin{bmatrix} 1 & & \\ & 1 & \\ & & -1/3 \end{bmatrix}, \; D^{-1}(L+U) = \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & 0 \\ 1/3 & 2/3 & 0 \end{bmatrix}$$

y $$(D-L)^{-1}U = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & -1 \\ 0 & 0 & -1 \end{bmatrix}.$$

Se tiene que los autovalores para la matriz del método de Jacobi son

$$\lambda_1 \simeq 0.747, \;\; \lambda_2 \simeq -0.374 + 0.867i \;\text{ y }\; \lambda_3 \simeq -0.374 - 0.867i,$$

y así $\max_i |\lambda_i| \simeq 0.944$. Consecuentemente las iteraciones del método de Jacobi convergerán.

Sin embargo, los autovalores para la matriz del método de Gauss-Seidel son $\lambda_1 = \lambda_2 = 0$ y $\lambda_3 = -1$, de donde se deduce que las iteraciones de Gauss-Seidel, para un vector inicial $x^{(0)}$ arbitrario, divergen.

En cualquier caso resulta conveniente ya hacer algunas precisiones. De una parte, la convergencia la entendemos cualquiera que sea la aproximación inicial, por lo que estos métodos resultan autocorrectivos, en el sentido de que un error de cálculo individual no afectará al resultado final, ya que cualquier aproximación errónea puede considerarse como un nuevo vector inicial. Es conveniente que el valor inicial que se tome esté lo más próximo posible a la solución real, a fin de reducir el tiempo de cálculo y aumentar la rapidez de convergencia; resulta adecuado hacer, cuando sea posible, una estimación razonable de la solución.

Otro problema es ¿cuándo parar?, ¿cuándo dejar de efectuar iteraciones? Una opción sería detenerse después de un número predeterminado de pasos. Sin embargo, como no es posible saber de antemano cuántas iteraciones serán necesarias realizar para obtener una buena aproximación, parece poco recomendable tomar tal determinación. Otra opción más plausible es detenerse cuando el error relativo sea suficientemente pequeño, tan pequeño como nosotros previamente hayamos elegido. Toda vez que el error relativo no puede determinarse con exactitud, ya que no se conoce la solución exacta, se requerirá un tratamiento más detenido de esta cuestión que realizaremos en el epígrafe **9.5.2**, dedicado a la acotación del error, y que va en la línea de que el proceso se pare cuando $|x_i^{(m+1)} - x_i^{(m)}| < \varepsilon$, con ε previamente determinado por nosotros.

9. Construcción general de métodos iterativos lineales. Estudio de la convergencia

9.1 Consideraciones Generales

Sea el sistema de ecuaciones lineales $Ax = b$, tal que A es regular. Consideremos el método iterativo dado por la función vectorial lineal siguiente

$$x^{(m+1)} = Bx^{(m)} + c, \tag{9.1}$$

tal que $B \in \mathbb{R}^{n,n}$ y $c \in \mathbb{R}^n$.

Un método como el (9.1) se dice que es *convergente*, si cualquiera que sea el vector $x^{(0)} \in \mathbb{R}^{(n)}$ inicial, se tiene que

$$\lim_{m\to\infty} x^{(m)} = x = A^{-1}b.$$

Como se desea que

$$\lim_{m\to\infty} x^{(m)} = x,$$

tendríamos que

$$\lim_{m\to\infty} x^{(m+1)} = B \lim_{m\to\infty} x^{(m)} + c = x,$$

luego

$$x = Bx + c. \tag{9.2}$$

Equivalentemente, al ser $x = A^{-1}b$, resultaría

$$A^{-1}b = BA^{-1}b + c,$$

de donde

$$c = (I - B)A^{-1}b. \tag{9.3}$$

Un método del tipo (9.1) que verifique (9.2) o equivalentemente (9.3), se dice que es *consistente* con el sistema $Ax = b$.

Sin embargo, la consistencia no garantiza la convergencia de la sucesión $\{x^{(m)}\}$; es decir, no se puede garantizar que con cualquier $x^{(0)}$ inicial el método converja.

¿Cuándo se podrá garantizar la convergencia?, ¿cómo habrán de elegirse B y c?

9.2 Estudio general de la convergencia

La respuesta al problema general de la convergencia la proporciona el siguiente resultado.

TEOREMA 25 *El método* $x^{(m+1)} = Bx^{(m)} + c$ *converge, respecto al sistema* $Ax = b$, *si y sólo si se verifican:*

1. $c = (I - B)A^{-1}b$; *es decir, el método es consistente.*
2. $\rho(B) < 1$.

Demostración.

1. Si el método $x^{(m+1)} = Bx^{(m)} + c$ es convergente, entonces como acabamos de ver se tendrá que $c = (I - B)A^{-1}b$, con lo que se verifica 1.

   Por otra parte, si a $x^{(m+1)} = Bx^{(m)} + c$ le restamos $x = Bx + c$, tendremos

   $$x^{(m)} - x = B(x^{(m-1)} - x) = B^2(x^{(m-2)} - x) = \ldots = B^m(x^{(0)} - x).$$

   Y como $\lim_{m\to\infty} x^{(m)} = x$, cualquiera que sea el $x^{(0)}$ inicial, entonces $\lim_{m\to\infty}(x^{(m)} - x) = 0$, y por tanto

   $$\lim_{m\to\infty} B^m(x^{(0)} - x) = 0, \text{ cualquiera que sea } (x^{(0)} - x) \in \mathbb{R}^n,$$

   luego

   $$\lim_{m\to\infty} B^m = \theta$$

   o equivalentemente, según vimos en el epígrafe **7.3.2**,

   $$\rho(B) < 1,$$

   con lo que obtenemos la condición 2.

2. Recíprocamente, si $\rho(B) < 1$, y como $\rho(B) = \inf_{\|.\|_M} \|B\|$, según se vió en el epígrafe **7.2.2**, entonces existe $\| \, . \, \|_M$ tal que $\| \, B \, \|_M < 1$.

   Como dada una norma matricial existen infinitas normas vectoriales compatibles con ella, tomando una de éstas, se tendrá

   $$\| \, Bv \, \| \leq \| \, B \, \|_M \| \, v \, \|$$

   y tal que $\| \, B \, \|_M < 1$.

   Por otra parte, por 1., el método es consistente con el sistema, luego $x = Bx + c$. Si a $x^{(m)} = Bx^{(m-1)} + c$ le restamos $x = Bx + c$, tendremos

   $$x^{(m)} - x = B^m(x^{(0)} - x),$$

   luego

   $$\| \, x^{(m)} - x \, \| \leq \| \, B \, \|^m \| \, x^{(0)} - x \, \|, \text{ de donde } \lim_{m\to\infty} \| \, x^{(m)} - x \, \| = 0,$$

independientemente de quién sea $x^{(m)}$. En definitiva

$$\lim_{m\to\infty} x^{(m)} = x,$$

para todo $x^{(0)} \in \mathbb{R}^n$ y por tanto la sucesión de vectores $x^{(m)}$ converge.

Además, de $x = Bx + c$ se tiene que $(I - B)x = c$. Y como, por 1., $c = (I - B)A^{-1}b$, se tendrá que

$$(I - B)x = (I - B)A^{-1}b.$$

Pero $(I - B)$ es invertible, ya que $\rho(B) < 1$, luego 1 no es autovalor de B, por lo que existe $(I - B)^{-1}$. Por tanto

$$x = A^{-1}b;$$

es decir, la x hacia donde converge el método es la solución[1] de $Ax = b$.

■

Para que $\rho(B) < 1$, es suficiente, como ya sabemos, que para alguna norma ser verifique que $\| B \| < 1$.

9.3 Construcción de métodos iterativos

Sea $A = M - N$ tal que M es regular. Entonces el sistema $Ax = b$ adquiere la forma $(M - N)x = b$, de donde $Mx = Nx + b$, luego

$$x = M^{-1}Nx + M^{-1}b.$$

Ello sugiere estudiar el método

$$x^{(m+1)} = M^{-1}Nx^{(m)} + M^{-1}b. \tag{9.4}$$

En él $B = M^{-1}N$ y $c = M^{-1}b$, por lo que

$$I - B = I - M^{-1}N = M^{-1}M - M^{-1}N = M^{-1}(M - N) = M^{-1}A,$$

entonces

$$(I - B)A^{-1}b = M^{-1}AA^{-1}b = M^{-1}b = c,$$

concluyéndose, de este modo, que el método es consistente con el sistema.

■

Se ha obtenido el siguiente resultado:

Toda matriz cuadrada A se puede descomponer en $A = M - N$, tal que M es regular y además el método iterativo

$$x^{(m+1)} = M^{-1}Nx^{(m)} + M^{-1}b$$

es consistente con el sistema $Ax = b$.

[1] Se podría también haber demostrado haciendo uso del teorema del punto fijo, pues $\| B \|_M$ sería la constante de Lipschitz, con lo que se está en las condiciones del teorema, luego existirá una única solución de $x = Bx + c$ y tal que $\lim_{m\to\infty} x^{(m)} = x$, cualquiera que sea $x^{(0)} \in \mathbb{R}^n$.

La condición necesaria y suficiente de convergencia es que $\rho(M^{-1}N) < 1$, para lo que es suficiente que para alguna norma matricial se verifique que $\| M^{-1}N \| < 1$.

Veamos ahora que

TEOREMA 26 *Si un método iterativo es lineal; es decir, de la forma*

$$x^{(m+1)} = Bx^{(m)} + c,$$

para que sea convergente necesariamente ha de ser de la forma (9.4) con $A = M - N$, *y siendo* M *regular.*

Demostración. Si se tiene $Ax = b$ y el método $x^{(m+1)} = Bx^{(m)} + c$; o lo que es lo mismo, nos son dadas A, B, C y c, vamos a ver que existen M y N tales que:

$$B = M^{-1}N, \quad c = M^{-1}b \quad \text{y} \quad A = M - N.$$

Al ser, por hipótesis, el método convergente, se tendrá, en virtud del teorema anterior, que

$$c = (I - B)A^{-1}b \quad \text{y} \quad \rho(B) < 1,$$

por lo que 1 no es autovalor de B, y por tanto existe $(I - B)^{-1}$.

Tomemos $M = A(I - B)^{-1}$ y $N = A(I - B)^{-1}B$, se tiene entonces que

$$M - N = A(I - B)^{-1} - A(I - B)^{-1}B = A(I - B)^{-1}(I - B) = A$$

y $M^{-1} = (I - B)A^{-1}$, de donde

$$M^{-1}N = (I - B)A^{-1}A(I - B)^{-1}B = B$$

y

$$M^{-1}b = (I - B)A^{-1}b = c.$$

■

El método $x^{(m+1)} = M^{-1}Nx^{(m)} + M^{-1}b$ suele representarse en la forma

$$Mx^{(m+1)} = Nx^{(m)} + b, \tag{9.5}$$

que resulta más cómodo para los cálculos.

La matriz M interesa tomarla de forma que sea fácilmente invertible; por ejemplo, diagonal o triangular.

9.4 Los métodos usuales

Vamos, ahora, a analizar de nuevo los métodos usuales desde la perspectiva de este tratamiento general de los métodos iterativos lineales y comprobar cuáles son en cada caso las matrices M y N.

9.4.1 El método de Jacobi

Descomponemos A en la forma $A = D - L - U$, con

$$D = \begin{bmatrix} a_{11} & & & & \\ & a_{22} & & & \\ & & a_{33} & & \\ & & & \ddots & \\ & & & & a_{nn} \end{bmatrix}, \quad L = \begin{bmatrix} 0 & & & \\ -a_{21} & 0 & & \\ \vdots & \vdots & \ddots & \\ -a_{n1} & \cdots & -a_{nn-1} & 0 \end{bmatrix}$$

y

$$U = \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ & 0 & & \vdots \\ & & \ddots & -a_{n-1n} \\ & & & 0 \end{bmatrix}. \tag{9.6}$$

Suponemos, como de costumbre, que A es regular, por lo que se puede considerar que $a_{ii} \neq 0$ para todo $i = 1, \ldots, n$, ya que si hay algún elemento diagonal nulo bastaría hacer una permutación de filas tal que el nuevo elemento diagonal fuese distinto de cero. Naturalmente la necesidad, en su caso, de alterar una ecuación en nada influye en la solución del sistema.

Haciendo $M = D$ y $N = L + U$, tenemos que

$$A = M - N \text{ y que existe } M^{-1}.$$

Estas matrices M y N son las que en el método iterativo

$$Mx^{(m+1)} = Nx^{(m)} + b$$

dan lugar al método de Jacobi, obteniéndose

$$a_{ii}x_i^{m+1} = \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}x_j^{(m)} \right), \quad i = 1, 2, \ldots, n, \quad m \geq 0,$$

o equivalentemente

$$x_i^{m+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}x_j^{(m)} \right), \quad i = 1, 2, \ldots, n, \quad m \geq 0,$$

que fue la expresión obtenida con anterioridad.

Por tanto, si hacemos $M^{-1}N$ deberemos obtener la matriz B del método (9.1):

$$B = M^{-1}N = - \begin{bmatrix} 0 & a_{12}/a_{11} & a_{13}/a_{11} & \cdots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 0 & a_{23}/a_{22} & \cdots & a_{2n}/a_{22} \\ \vdots & \vdots & 0 & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & a_{n3}/a_{nn} & \cdots & 0 \end{bmatrix}$$

que coincide con el valor que se obtuvo en el epígrafe **8.1.2** cuando abordamos la convergencia del método de Jacobi.

En las condiciones dadas se verifica que $c = (I - B)A^{-1}b$. Luego sólo resta para la convergencia, como ya vimos, que $\rho(M^{-1}N) < 1$.

Pero $B = M^{-1}N$ es una matriz muy sencilla. Si $\| B \| < 1$ para alguna norma matricial, entonces tendremos garantizada la convergencia. Por ejemplo:

- Si $\sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}/a_{ii}| < 1$ cualquiera que sea $i = 1, \ldots, n$, entonces el máximo de las filas también será menor que 1, con lo que la norma fila será menor que 1,

$$\| B \|_\infty = \max_{1\leq i\leq n} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}/a_{ii}| < 1.$$

- O si $\sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}/a_{jj}| < 1$ cualquiera que sea $j = 1, \ldots, n$, entonces el máximo de las columnas también será menor que 1, y por ello la norma columna será menor que 1,

$$\| B \|_1 = \max_{1\leq j\leq n} \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}/a_{jj}| < 1.$$

Con ello se tendría garantizada que $\rho(B) < 1$ y como el método es consistente, se tiene garantizada la convergencia.

Si no fuera posible encontrar con facilidad una norma que nos garantizara la convergencia, como $|\lambda^{(B)}| < 1$ para cada autovalor $\lambda^{(B)}$ de B es necesario y suficiente para la convergencia, habría que estudiar las raíces de

$$|B - \lambda I| = 0,$$

con $B = M^{-1}N, M = D$ y $N = L + U$; es decir $B = D^{-1}(L + U)$. Por tanto

$$|D^{-1}(L + U) - \lambda I| = 0,$$

como $|D| \neq 0$ por ser todos los elementos $a_{ii} \neq 0$, tiene sentido $|D^{-1}| = 1/|D|$ y además que $|D|\,|D^{-1}(L + U) - \lambda I| = |D|0 = 0$. Por lo que

$$|DD^{-1}(L + U) - \lambda DI| = 0,$$

de donde $|(L + U) - \lambda D| = 0$.

Por tanto

$$|\lambda D - L - U| = \begin{vmatrix} \lambda a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & \lambda a_{22} & & & \\ a_{31} & & \lambda a_{33} & & \\ \vdots & & & \ddots & \\ a_{n1} & \cdots & \cdots & \cdots & \lambda a_{nn} \end{vmatrix} = 0.$$

La demostración de que los autovalores, raíces de la ecuación anterior, son de módulo menor que 1, puede ser, en general, difícil, por lo que poco se puede adelantar. Sin embargo, para ciertos casos particulares de la matriz A, que estudiaremos en el capítulo 10, puede asegurarse la convergencia.

9.4.2 El método de Gauss-Seidel

Consideremos nuevamente las matrices formuladas en (9.6) y creadas con elementos de la matriz del sistema A. Si ahora $M = D - L$ y $N = U$, tendremos entonces que $A = M - N = D - L - U$, con $M = D - L$ triangular inferior. Y suponiendo sin mayor dificultad, como antes, que $a_{ii} \neq 0$, cuando i recorre los valores $i = 1, \ldots, n$, entonces M también es invertible.

En este caso el método iterativo lineal (9.5)

$$Mx^{(m+1)} = Nx^{(m)} + b$$

se traduce en

$$\sum_{j=1}^{i} a_{ij} x_j^{(m+1)} = -\sum_{j=i+1}^{n} a_{ij} x_j^{(m)} + b_i, \quad i = 1, \ldots, n. \tag{9.7}$$

Partiendo de la primera ecuación se pueden ir calculando las siguientes, utilizando los resultados ya obtenidos, con lo que resulta

$$a_{ii} x_i^{(m+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(m)} \right), \quad i = 1, 2, \ldots, n, \quad m \geq 0,$$

o equivalentemente

$$x_i^{(m+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(m)} \right), \quad i = 1, 2, \ldots, n, \quad m \geq 0,$$

que fue la expresión que se obtuvo anteriormente, en el epígrafe **8.2.1**.

Análogamente a lo hecho con el método de Jacobi, la matriz B del método en su expresión (9.1) será $B = M^{-1}N$. En nuestro caso $B = (D - L)^{-1}U$.

Pero ahora la expresión de B es más compleja que en el caso del método de Jacobi, por ello encontrar una norma matricial tal que $\| B \| < 1$ resultará aún más complicado que en el caso anterior. Como la consistencia del método está garantizada, lo que habrá que estudiar serán los autovalores de B, a fin de averiguar si el radio espectral de esa matriz es menor que 1.

$$|B - \lambda I| = |(D - L)^{-1}U - \lambda I| = 0.$$

Al ser $M = (D - L)$ invertible, se tendrá que $|D - L| \neq 0$. Por tanto

$$\begin{aligned} |D - L|\,|(D - L)^{-1}U - \lambda I| &= |(D - L)(D - L)^{-1}U - \lambda(D - L)I| = \\ &= |U - \lambda(D - L)| = 0, \end{aligned}$$

luego

$$|\lambda D - \lambda L - U| = \begin{vmatrix} \lambda a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ \lambda a_{21} & \lambda a_{22} & & & \\ \vdots & & \lambda a_{33} & & \\ \vdots & & & \ddots & \\ \lambda a_{n1} & \cdots & \cdots & \cdots & \lambda a_{nn} \end{vmatrix} = 0.$$

Al igual que antes, la determinación de λ en la ecuación anterior no es fácil. Sin embargo para ciertas matrices A especiales, el problema de la convergencia de Gauss-Seidel está solucionado.

9.5 Consideraciones prácticas

9.5.1 Estabilidad de los métodos iterativos

Para ver la estabilidad del método

$$x^{(m+1)} = Bx^{(m)} + c,$$

veamos cómo le afectan las perturbaciones. Si al hacer los cálculos en lugar de $x^{(m)}$ obtenemos un valor perturbado del mismo $\overline{x}^{(m)}$, en vez de

$$x^{(m)} = Bx^{(m-1)} + c,$$

realmente lo que se ha hecho es

$$\overline{x}^{(m)} = B\overline{x}^{(m-1)} + c + \delta^{(m)},$$

donde $\delta^{(m)}$ es el vector perturbación en el paso m-ésimo. Si restamos las dos últimas expresiones, resultará

$$\overline{x}^{(m)} - x^{(m)} = B(\overline{x}^{(m-1)} - x^{(m-1)}) + \delta^{(m)}$$

y $\overline{x}^{(0)} - x^{(0)} = \delta^{(0)}$, de donde

$$\overline{x}^{(m)} - x^{(m)} = B^m\delta^{(0)} + \sum_{j=1}^{m} B^{m-j}\delta^j = \sum_{j=0}^{m} B^{m-j}\delta^{(j)}.$$

Por tanto, tomando normas matricial y vectorial compatibles, tendremos

$$\| \overline{x}^{(m)} - x^{(m)} \| \leq \sum_{j=0}^{m} \| B \|^{m-j} \| \delta^{(j)} \| .$$

Si el método es convergente $\rho(B) < 1$, entonces podemos tomar una norma matricial tal que $\| B \| < 1$, y de aquí que $\sum_{r=0}^{\infty} \| B \|^r$ es convergente, luego

$$\sum_{r=0}^{\infty} \| B \|^r \leq k < +\infty.$$

Por lo tanto, podemos poner

$$\begin{aligned} \| \overline{x}^{(m)} - x^{(m)} \| &\leq \textstyle\sum_{j=0}^{m} \| B \|^{m-j} \| \delta^{(j)} \| \\ &\leq \textstyle\sum_{r=0}^{\infty} \| B \|^r \ \sup_j \| \delta^{(j)} \| \\ &\leq k \sup_j \| \delta^{(j)} \| . \end{aligned}$$

Consecuentemente, errores pequeños de redondeo producirán pequeños errores en el resultado del paso m-ésimo; es decir, en este sentido el método es estable.

Cuanto más pequeña sea $\| B \|$ mejor será la acotación del error de perturbación.

9.5.2 Acotación del error

Como $x^{(m+1)} = Bx^{(m)} + c$ y $x = Bx + c$, entonces

$$x^{(m)} - x = B(x^{(m-1)} - x) = B^m(x^{(0)} - x),$$

luego $\| x^{(m)} - x \| \leq \| B \|^m \| x^{(0)} - x \|$, con normas compatibles. No obstante esta acotación del error en el paso m-ésimo presenta el inconveniente de tener x en $\| x^{(0)} - x \|$, no conociéndose x, por lo que hacemos

$$\begin{aligned} \| x^{(m-1)} - x \| &= \| x^{(m-1)} - x^{(m)} + x^{(m)} - x \| \\ &\leq \| -x^{(m)} + x^{(m-1)} \| + \| x^{(m)} - x \| \end{aligned}$$

y $\| x^{(m)} - x \| \leq \| B \| \| x^{(m-1)} - x \|$, por tanto

$$\| x^{(m-1)} - x \| \leq \| x^{(m)} - x^{(m-1)} \| + \| B \| \| x^{(m-1)} - x \|,$$

de donde $(1 - \| B \|) \| x^{(m-1)} - x \| \leq \| x^{(m)} - x^{(m-1)} \|$.

Suponemos que el método es convergente, luego $\| B \| < 1$; es decir

$$1 - \| B \| > 0,$$

por lo que

$$\| x^{(m-1)} - x \| \leq \frac{1}{1 - \| B \|} \| x^{(m)} - x^{(m-1)} \| .$$

Sustituyendo esta última expresión en $\| x^{(m)} - x \| \leq \| B \| \| x^{(m-1)} - x \|$, resulta

$$\| x^{(m)} - x \| \leq \frac{\| B \|}{1 - \| B \|} \| x^{(m)} - x^{(m-1)} \|,$$

que nos proporciona una acotación del error y puede servir como test de parada del método, cuando la norma es menor que 1. Como ordinariamente es difícil saber si la norma es menor que 1, vamos a realizar una estimación del valor de la misma.

¿Cuándo parar en la práctica?

Dado que $\| x^{(m)} - x \| \leq \| B \| \| x^{(m-1)} - x \|$, en cada paso el error decrece en un factor constante $c = \| B \| < 1$ estrechamente relacionado con $\rho(B)$. Podemos hacer una estimación del valor de la norma $\| B \|$ como sigue

$$(x^{(m)} - x) - (x^{(m-1)} - x) = e^{(m)} - e^{(m-1)} = x^{(m)} - x^{(m-1)}$$

y como $e^{(m)} = Be^{(m-1)}$, se tiene que

$$x^{(m)} - x^{(m-1)} = B(e^{(m-1)} - e^{(m-2)}) = B(x^{(m-1)} - x^{(m-2)}),$$

lo que nos indica que

$$c = \| B \| \simeq \frac{\| x^{(m)} - x^{(m-1)} \|}{\| x^{(m-1)} - x^{(m-2)} \|}$$

nos puede servir como estimación de la norma de B. Para mayor seguridad se puede tomar el mayor de tales cocientes para varios "m".

La acotación del error antes obtenida puede ser sustituida por esta aproximación, que resulta también una acotación del error, pero más práctica

$$\| x^{(m)} - x \| \leq \frac{c}{1-c} \| x^{(m)} - x^{(m-1)} \| . \tag{9.8}$$

Podemos decidir inicialmente que el método se pare cuando

$$\| x^{(m)} - x^{(m-1)} \| < \varepsilon,$$

siendo ε tan pequeño como deseemos. Y en ese momento podremos tener una acotación del error cometido, mediante (9.8).

9.5.3 ¿Cuándo usar métodos iterativos?

Nos hemos referido en múltiples ocasiones a cuándo resulta más práctico utilizar un método iterativo frente a otro directo. Veámos esta cuestión desde la óptica del número de iteraciones que serán necesarias para obtener una aproximación a la solución con un error deseado.

Supongamos que queremos hallar m cuando

$$\| x^{(m)} - x \| \leq \varepsilon \| x^{(0)} - x \|,$$

es decir, hacer tantas iteraciones como sean necesarias hasta que el error sea menor que un cierto error ε predeterminado por nosotros. La expresión anterior puede ser escrita como

$$\| e^{(m)} \| \leq \varepsilon \| e^{(0)} \| .$$

Como $\| x^{(m)} - x \| \leq \| B \|^m \| x^{(0)} - x \| \leq c^m \| x^{(0)} - x \|$, lo que debemos hallar es el menor valor de m tal que $c^m \leq \varepsilon$, por lo que

$$m \geq \frac{\ln \varepsilon}{\ln c} = m^*.$$

Por consiguiente, si A es de dimensión n, el número de operaciones por cada iteración es $2n^2 - n$, mientras que para la eliminación de Gauss es aproximadamente del orden de $2n^3/3$, por lo que el método iterativo será más interesante que el directo cuando

$$m^* 2n^2 < 2\frac{n^3}{3},$$

o lo que es lo mismo, cuando

$$m^* < \frac{n}{3}.$$

En líneas generales el coste de los métodos iterativos es similar al de los directos, pero a diferencia de estos últimos, los iterativos, una vez garantizada la convergencia, y para matrices que no estén mal condicionadas son "autocorrectivos", es decir, los errores de redondeo no causan estragos en la solución final; son métodos estables. Una vez garantizada la convergencia, los errores de redondeo no causan estragos en la solución final, pues el

método converge en las primeras iteraciones hacia una "región" dentro de la que se encuentra la solución exacta. Por otra parte, en los métodos directos la matriz debe estar integramente en la memoria RAM del ordenador, mientras que para los iterativos se pueden utilizar técnicas que no precisan que toda la matriz esté en memoria, leyendo del disco paulatinamente. Por ello, para sistemas muy grandes, los métodos iterativos son más recomendables. Si además estos sistemas muy grandes tienen una matriz "rala", entonces el coste de operaciones es mucho menor que $2n^2$, con lo que la conveniencia de los métodos iterativos se redobla.

10. Métodos iterativos en el caso de matrices especiales

En el capítulo **9** vimos que en general no era fácil la determinación de la convergencia de los métodos iterativos de Jacobi y de Gauss-Seidel.

De la misma forma avanzábamos que para ciertos casos particulares de matrices del sistema, era posible garantizar la convergencia de estos métodos.

Este apartado lo dedicaremos precisamente a analizar estas circunstancias.

10.0.4 Matrices diagonales estrictamente dominantes

Una matriz cuadrada $A \in \mathcal{M}_n(\mathbb{R})$ se dice que es de *diagonal estrictamente dominante* si, en cada fila, el valor absoluto del elemento diagonal es mayor que la suma de los valores absolutos de los elementos situados fuera de la diagonal. Es decir, si

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad \text{con } i = 1, \ldots, n.$$

Veamos que una matriz de estas características es invertible.

Si A no es invertible, el sistema $Ax = 0$ tendrá solución x distinta de la trivial. Para ese x, sea $|x_p| = \max_{1 \leq i \leq n} |x_i|$.

Al hacer Ax, en la fila p se tendrá

$$\sum_{j=1}^{n} a_{pj} x_j = 0, \text{ de donde } a_{pp} x_p = -\sum_{\substack{j=1 \\ j \neq p}}^{n} a_{pj} x_j \text{ y por tanto}$$

$|a_{pp}||x_p| \leq \sum_{\substack{j=1 \\ j \neq p}}^{n} |a_{pj}||x_j|$, luego

$$|a_{pp}| \leq \sum_{\substack{j=1 \\ j \neq p}}^{n} |a_{pj}| \frac{|x_j|}{|x_p|} \leq \sum_{\substack{j=1 \\ j \neq p}}^{n} |a_{pj}|, \tag{10.1}$$

ya que $\frac{|x_j|}{|x_p|} < 1$. Pero (10.1) contradice que A es de diagonal estrictamente dominante. Consecuentemente A es regular. ■

Hemos obtenido el siguiente resultado.

TEOREMA 27 *Si A es de diagonal estrictamente dominante, entonces A es invertible.*

Vamos a estudiar la convergencia de los métodos tratados en los capítulos anteriores, para el caso en que A sea de diagonal estrictamente dominante.

- Método de Jacobi

  Sea $M^{-1}N = B_J$ la matriz del método de Jacobi. Por la definición de M y N para este método, tendremos que

  $$B_J = D^{-1}(L+U),$$

  por tanto, como ya vimos

  $$B_J = -\begin{bmatrix} 0 & a_{12}/a_{11} & a_{13}/a_{11} & \cdots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 0 & a_{23}/a_{22} & \cdots & a_{2n}/a_{22} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & a_{n3}/a_{nn} & \cdots & 0 \end{bmatrix}.$$

  De ser A de diagonal estrictamente dominante se tiene que

  $$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \text{ para } i = 1, \ldots, n, \text{ por tanto } \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1,$$

  luego el máximo de todos los i también será menor que 1; es decir,

  $$\| B_J \|_\infty = \max_{1\leq i\leq n} \left\{ \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} \right\} < 1,$$

  por lo que $\rho(B) < 1$ y el método converge.

- Método de Gauss-Seidel

  La matriz $B_G = M^{-1}N$ del método en este caso es, dadas las formas de M y N en esta ocasión,

  $$B_G = (D-L)^{-1}U.$$

  Desgraciadamente B_G no adquiere una forma tan sencilla en su determinación como B_J, por lo que utilizaremos otro procedimiento.

  Consideremos la norma matricial fila inducida por la norma vectorial del supremo, que es

  $$\| B_G \|_\infty = \max_{\|x\|_\infty = 1} \| B_G\, x \|_\infty .$$

  Ya que A es de diagonal estrictamente dominante, repitiendo el razonamiento efectuado anteriormente, en el caso del método de Jacobi, tendremos que

  $$c = \max_{i\leq i\leq n} \left\{ \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} \right\} < 1.$$

  Supongamos que $\| B_G \|_\infty \leq c < 1$, entonces sería $\rho(B_G) < 1$, con lo que el método convergerá.

La cuestión es, pues, demostrar que$\| B_G \|_\infty \leq c$. Para ello, sea x un vector cualquiera tal que $\| x \|_\infty = 1$, y llamemos $y = B_G\, x$. En estas condiciones, se trata de demostrar que

$$\| y \|_\infty \leq \max_{1\leq i\leq n} \left\{ \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} \right\} = c.$$

Para ello veamos los valores absolutos de las componentes de y.

Recordemos que el método de Gauss-Seidel

$$x^{(m+1)} = M^{-1}Nx^{(m)} + M^{-1}b,$$

siendo $B_G = M^{-1}N$, adquiría la forma

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

$i = 1, 2, \ldots, n, \quad m \geq 0$; por lo tanto, para $y = B_G\, x$, sus componentes adquirirán la expresión

$$y_i = \frac{1}{a_{ii}} \left(-\sum_{j=1}^{i-1} a_{ij}y_j - \sum_{j=i+1}^{n} a_{ij}x_j \right), \quad i = 1, 2, \ldots, n,$$

toda vez que, para $j = 1, \ldots, i-1$, las componentes del vector producto de B_G y x, son las calculadas antes y por tanto y_j; mientras que para $j = i+1, \ldots, n$, son las de x, luego las x_j.

Conociendo ya la expresión de las componentes de y, veamos la acotación de sus valores absolutos, procediendo por inducción.

$$|y_1| \leq \frac{1}{|a_{11}|} \sum_{j=2}^{n} |a_{1j}||x_j|$$

y como $\| x \|_\infty = 1$, entonces $|x_j| \leq 1$, y $\sum_{j=2}^{n} \frac{|a_{1j}|}{|a_{11}|} < 1$ por ser A de diagonal estrictamente dominante, finalmente quedará

$$|y_1| \leq \sum_{j=2}^{n} \frac{|a_{1j}|}{|a_{11}|} < 1.$$

Supongamos que es cierto hasta $i = k-1$ que

$$|y_i| \leq c = \max_{1\leq i\leq n} \left\{ \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} \right\} < 1.$$

Veamos que también lo es para $i = k$.

$$|y_k| \leq \frac{1}{|a_{kk}|} \left(\sum_{j=1}^{k-1} |a_{kj}||y_j| + \sum_{j=k+1}^{n} |a_{kj}||x_j| \right).$$

Ahora bien, en la anterior expresión se tiene que $|y_j| < 1$ por hipótesis de inducción, y $|x_j| \leq 1$ ya que la norma del supremo de x es 1. Por tanto tendremos que

$$|y_k| < \frac{1}{|a_{kk}|} \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}| < 1$$

y menor que 1, por ser A de diagonal estrictamente dominante.

Concluimos entonces que cualquiera que sea $i = 1, \ldots, n$, se verifica que $|y_i| < 1$ y, por tanto, que $\|y\|_\infty = \| B_G\, x \|_\infty < 1$, cualquiera que sea x tal que $\| x \|_\infty = 1$; por lo que se tendrá que

$$\max_{\|x\|_\infty = 1} \| B_G x \|_\infty < 1.$$

De esta manera $\| B_G \|_\infty < 1$ y $\rho(B_G) < 1$, lo que significa que el método converge. ■

Hemos obtenido el siguiente resultado.

TEOREMA 28 *Si la matriz A del sistema $Ax = b$ es de diagonal estrictamente dominante, entonces los métodos de Jacobi y Gauss-Seidel son convergentes.*

10.1 Matrices simétricas definidas positivas

TEOREMA 29 *Sea A una matriz simétrica real, y sea $A = M - N$ una descomposición de A con M invertible, tal que $M^t + M - A$ es definida positiva. En estas condiciones, el método $x^{(m+1)} = M^{-1}Nx^{(m)} + M^{-1}b$ converge si y sólo si A es definida positiva*[1].

Demostración.

1. Supongamos que A es definida positiva.

   Sean λ un autovalor cualquiera de $M^{-1}N$ y $u \neq 0$ un autovector asociado; es decir $M^{-1}Nu = \lambda u$, o equivalentemente $Nu = \lambda Mu$.

   (Hay que hacer notar que tanto λ como u pueden estar sobre el cuerpo de los complejos, ya que λ no es autovalor de A, simétrica real, sino de $M^{-1}N$).

   Tendremos $Au = (M - N)u = Mu - Nu = Mu - \lambda Mu$ y, por tanto,

   $$Au = (1 - \lambda)Mu.$$

   Como A es definida positiva, se tendrá que $v^*Av > 0$, siendo $v \in \mathbb{C}^n$ (ya que los vectores v pueden ser también complejos). De aquí que $Au \neq 0$, porque en caso contrario $u^*Au = 0$, contradiciendo que A es definida positiva.

[1] El resultado sigue siendo cierto con matrices A hemíticas definidas positivas.

Por tanto, $Au = (1-\lambda)Mu \neq 0$ y entonces $u^*Au = (1-\lambda)u^*Mu \neq 0$. Luego

$$\frac{1}{1-\lambda} = \frac{u^*Mu}{u^*Au}, \tag{10.2}$$

siendo $u^*Au > 0$ y M y A reales.

Tomando conjugados complejos en (10.2) tendremos

$$\left(\frac{u^*Mu}{u^*Au}\right) = \frac{\overline{u}^*\overline{M}\overline{u}}{u^*Au} = \frac{u^tM\overline{u}}{u^*Au} = \frac{(\overline{u}^tM^tu)^t}{u^*Au} = \frac{(u^*M^tu)^t}{u^*Au} = \frac{u^*M^tu}{u^*Au}. \tag{10.3}$$

Sumando (10.2) y (10.3), se tendrá que

$$2Re\frac{1}{1-\lambda} = \frac{u^*(M+M^t)u}{u^*Au}.$$

Como $M = A + N$, quedará

$$2Re\frac{1}{1-\lambda} = \frac{u^*(A+N+M^t)u}{u^*Au} = 1 + \frac{u^*(N+M^t)u}{u^*Au}.$$

Ahora bien, si $\lambda = \alpha + \beta i$, se tiene que

$$2Re\frac{1}{1-\lambda} = \frac{2(1-\alpha)}{(1-\alpha)^2+\beta^2}\ .$$

Además, por hipótesis $M^t + M - A = M^t + N$ es definida positiva, luego $u^*(M^t+N)u > 0$ y como $u^*Au > 0$, finalmente tendremos que

$$\frac{2(1-\alpha)}{(1-\alpha)^2+\beta^2} = 1 + \frac{u^*(N+M^t)u}{u^*Au} > 1.$$

Operando queda $1 > \alpha^2+\beta^2 = |\lambda|^2$, por lo que $|\lambda| < 1$ y ello para todo autovalor λ de $M^{-1}N$; entonces $\rho(M^{-1}N) < 1$ y por consiguiente el método converge.

2. Recíprocamente, si el método es convergente, y por lo tanto

$$\rho(M^{-1}N) < 1,$$

vamos a ver que entonces A es definida positiva.

Sea $v^{(0)} \in \mathbb{R}^n$ cualquiera, tal que $v^{(0)} \neq 0$. Definimos la sucesión de vectores

$$v^{(m+1)} = M^{-1}Nv^{(m)}.$$

Llamando $B = M^{-1}N$, tendremos que $v^{(m+1)} = Bv^{(m)} = B^{m+1}v^{(0)}$.

Como $\rho(B) < 1$, entonces B^{m+1} converge a θ.

Consecuentemente, de estas dos últimas situaciones, se tiene que

$$\lim_{m\to\infty} v^{(m)} = 0.$$

Para ver que A es definida positiva, hemos de ver que $v^{(0)t}Av^{(0)} > 0$. Y para comprobar ésto último vamos a formar la sucesión $\{v^{(m)t}Av^{(m)}\}$,

y vamos a demostrar que esta sucesión es monótona decreciente y que su límite es cero.

Veamos que el límite es cero.

Por una parte [2],

$$|v^{(m)t}Av^{(m)}| \leq \| v^{(m)t} \|_1 \| Av^{(m)} \|_1 \leq \| v^{(m)t} \|_1 \| A \|_1 \| v^{(m)} \|_1,$$

por lo que $|v^{(m)t}Av^{(m)}| \leq \| A \|_1 \| v^{(m)} \|_1^2$.

Y como, por otra parte, $\lim_{m\to\infty} v^{(m)} = 0$, entonces

$$\lim_{m\to\infty} \| v^{(m)} \|_1^2 = 0.$$

Luego $\lim_{m\to\infty} |v^{(m)t}Av^{(m)}| = 0$, de donde $\lim_{m\to\infty} v^{(m)t}Av^{(m)} = 0$.

Veamos que la sucesión $\{v^{(m)t}Av^{(m)}\}$ es monótona decreciente.

Como $v^{(m+1)} = M^{-1}Vv^{(m)}$, tendremos que

$$\begin{aligned} & v^{(m)t}Av^{(m)} - v^{(m+1)t}Av^{(m+1)} = \\ & = v^{(m)t}Av^{(m)} - (M^{-1}Nv^{(m)})^t AM^{-1}Nv^{(m)} \\ & = v^{(m)t}(A - N^tM^{-1t}AM^{-1}N)v^{(m)}. \end{aligned} \tag{10.4}$$

Ahora bien, $N = M - A$ por lo que

$$\begin{aligned} A - N^tM^{-1t}AM^{-1}N &= A - (M^{-1}(M-A))^tAM^{-1}(M-A) \\ &= -(M^{-1}M - M^{-1}A)^tAM^{-1}(M-A) \\ &= A - (I - A^tM^{-1t})AM^{-1}(M-A) \\ &= A - AM^{-1}(M-A) + AM^{-1t}AM^{-1}(M-A) \\ &= A - AM^{-1}M + AM^{-1}A + AM^{-1t}AM^{-1}M - AM^{-1t}AM^{-1}A \\ &= AM^{-1}A + AM^{-1t}A - AM^{-1t}AM^{-1}A \\ &= A(M^{-1t}M^t)M^{-1}A + AM^{-1t}MM^{-1}A - AM^{-1t}AM^{-1}A \\ &= AM^{-1t}(M^t + M - A)M^{-1}A = AM^{-1t}(M^t + N)M^{-1}A. \end{aligned}$$

Por ello (10.4) queda en la forma

$$\begin{aligned} v^{(m)t}Av^{(m)} - v^{(m+1)t}Av^{(m+1)} &= v^{(m)t}AM^{-1t}(M^t + N)M^{-1}Av^{(m)} \\ &= y^t(M^t + N)y, \end{aligned}$$

llamando $y = M^{-1}Av^{(m)}$.

Y para $y \neq 0$, lo que equivale a decir que $v^{(m)} \neq 0$, se tiene que

$$y^t(M^t + N)y > 0,$$

[2] La desigualdad $|v^{(m)t}Av^{(m)}| \leq \| v^{(m)t} \|_1 \| Av^{(m)} \|_1$ es inmediata a partir de las propiedades de los valores absolutos de los números reales.

ya que, por hipótesis $(M^t + N)$ es definida positiva.

De $v^{(m)t}Av^{(m)} - v^{(m+1)t}Av^{(m+1)} = y^t(M^t + N)y > 0$ se concluye que la sucesión es estrictamente decreciente, con lo que se finaliza la demostración.

■

TEOREMA 30 *Sea A simétrica real con elementos diagonales a_{ii} positivos. Entonces el método de Gauss-Seidel converge si y sólo si A es definida positiva.*

Demostración.

En el método de Gauss-Seidel: $M = D-L, N = U$ y $A = M-N = D-L-U$, por lo que $M^t+N = D^t - L^t+U$. Pero como $A^t = A$, se tendrá que $L^t = U$ y $D^t = D$, resultando

$$M^t + N = D - L^t + L^t = D.$$

Como D = diagonal (a_{ii}) y $a_{ii} > 0$, entonces D es definida positiva. Por consiguiente $M^t + N = M^t + M - A$ es definida positiva. Entonces, por el teorema anterior, el método converge si y sólo si A es definida positiva. ■

TEOREMA 31 *Sea A simétrica real. Y la matriz que resulta sustituyendo los a_{ij}, con $i \neq j$, por sus opuestos, es definida positiva; entonces el método de Jacobi converge si y sólo si A es definida positiva.*

Demostración.

En este caso $M = D$, $N = L + U$ y $A = M - N$, por lo que

$$M^t + N = D^t + L + U = D + L + U \quad \text{y} \quad A = D - L - U.$$

Por hipótesis, $M^t + N$ es definida positiva y por tanto el método de Jacobi converge si y sólo si A es definida positiva. ■

10.2 Matrices tridiagonales

El estudio de la convergencia de los métodos de Jacobi, Gauss-Seidel y *Successive overrelaxation SOR*[3], para este tipo de matrices, fue desarrollado en la década de los cincuenta por Young, Frankel y Varga.

Vamos a ver que para matrices tridiagonales los métodos de Jacobi y Gauss-Seidel son simultáneamente convergentes o divergentes.

Consideremos la matriz tridiagonal

$$A = \begin{bmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & a_n & b_n \end{bmatrix}$$

[3]En castellano suele denominársele método de Sobrerrelajación, aunque hay autores que lo denominan simplemente método de Relajación. Lo más extendido, no obstante, es denominarlo por su abreviatura en inglés, SOR, que será la nomenclatura que utilizaremos aquí. Nosotros realizaremos su estudio en el epígrafe 11.2

y para $\mu \neq 0$, llamamos

$$A(\mu) = \begin{bmatrix} b_1 & c_1 & & & \\ \mu a_2 & b_2 & c_2/\mu & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c_{n-1}/\mu \\ & & & \mu a_n & b_n \end{bmatrix}.$$

Considerando

$Q(\mu) = \text{diag}(\mu, \mu^2, \ldots, \mu^n)$, entonces

$$Q(\mu)^{-1} = \text{diag}(1/\mu, 1/\mu^2, \ldots, 1/\mu^n),$$

por lo que

$$Q(\mu)AQ(\mu)^{-1} = A(\mu).$$

Luego A y $A(\mu)$ son semejantes, luego tienen el mismo determinante

$$|A(\mu)| = |A|.$$

■

Este resultado previo nos será de utilidad para demostrar el que adelantábamos al inicio de este epígrafe y que podemos enunciar como sigue.

TEOREMA 32 *Los radios espectrales de las matrices de los métodos de Jacobi y Gauss-Seidel para matrices tridiagonales verifican que*

$$\rho(B_G) = \rho(B_J)^2$$

y, por tanto, ambos métodos son simultáneamente convergentes o divergentes.

Demostración.

$$B_J = D^{-1}(L+U) \quad \text{y} \quad B_G = (D-L)^{-1}U$$

son las matrices de los métodos de Jacobi y Gauss-Seidel, respectivamente.

Como ya se puso de manifiesto en la sección **9.4**, los valores propios de B_J son los ceros del polinomio

$$P_{B_J} = |\lambda D - L - U| = \begin{bmatrix} \lambda b_1 & c_1 & & & \\ a_2 & \lambda b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & a_n & \lambda b_n \end{bmatrix}$$

y los de B_G son los ceros del polinomio

$$P_{B_G} = |\lambda D - \lambda L - U| = \begin{bmatrix} \lambda b_1 & c_1 & & & \\ \lambda a_2 & \lambda b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & \lambda a_n & \lambda b_n \end{bmatrix}$$

como ya se indicó también en el epígrafe **9.4.2**.

Para todo $\alpha \neq 0$, se tiene que

$$P_{B_G}(\alpha^2) = |\alpha^2 D - \alpha^2 L - U| = \begin{bmatrix} \alpha^2 b_1 & c_1 & & & \\ \alpha^2 a_2 & \alpha^2 b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & \alpha^2 a_n & \alpha^2 b_n \end{bmatrix} =$$

(la siguiente igualdad podríamos verla por inducción; o también multiplicando las filas $n-1, \ldots, 1$, respectivamente por $\alpha, \alpha^2, \ldots, \alpha^{n-1}$ y dividiendo las columnas $n-1, \ldots, 1$, respectivamente por $\alpha, \ldots, \alpha^{n-1}$)

$$\begin{bmatrix} \alpha^2 b_1 & \alpha c_1 & & & \\ \alpha a_2 & \alpha^2 b_2 & \alpha c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \alpha c_{n-1} \\ & & & \alpha a_n & \alpha^2 b_n \end{bmatrix} = |\alpha^2 D - \alpha L - \alpha U|. \tag{10.5}$$

Por continuidad, esta igualdad se verificará también para $\alpha = 0$, ya que $P_{B_G}(\alpha^2)$ es una función continua de α, puesto que se trata de un polinomio.

Por tanto, si α es autovalor de B_J, entonces será raíz de $|\lambda D - L - U| = 0$, por lo que $|\alpha D - L - U| = 0$.

Luego $|\alpha^2 D - \alpha L - \alpha U| = |\alpha(\alpha D - L - U)| = \alpha^n |\alpha D - L - U| = 0$.

Por otra parte, por (10.5),

$$|\alpha^2 D - \alpha L - \alpha U| = |\alpha^2 D - \alpha^2 L - U|,$$

luego si α es autovalor de B_J, α^2 es raíz de $|\lambda D - \lambda L - U| = 0$ y

$$\alpha^2 \text{ es autovalor de } B_G. \tag{10.6}$$

Y si $\beta \neq 0$ es valor propio de B_G, entonces $|\beta D - \beta L - U| = 0$. Y como

$$\begin{aligned} 0 = |(\sqrt{\beta})^2 D - (\sqrt{\beta})^2 L - U| &= |(\sqrt{\beta})^2 D - \sqrt{\beta} L - \sqrt{\beta} U| \\ &= |\sqrt{\beta}(\sqrt{\beta} D - L - U)| \\ &= (\sqrt{\beta})^n |\sqrt{\beta} D - L - U)|, \end{aligned}$$

entonces $|\sqrt{\beta} D - L - U)| = 0$, por lo que $\pm\sqrt{\beta}$ es raíz de $|\lambda D - L - U| = 0$ y, por lo tanto,

$$\pm\sqrt{\beta} \text{ es autovalor de } B_J. \tag{10.7}$$

Si[4] $\mu = \rho(B_G)$, $\mu \neq 0$, entonces μ es el autovalor de B_G de módulo máximo, entonces, por (10.7), $\sqrt{\mu}$ será autovalor de B_J, y por (10.6), será el de módulo máximo, de lo que se deduce que $\rho(B_J) = \sqrt{\mu}$.

Luego $\rho(B_G) = \rho(B_J)^2$. ■

[4] Naturalmente si $\mu = 0$ se tiene que $\rho(B_G) = 0$ y entonces necesariamente $\rho(B_J) = 0$, con lo que el teorema sigue verificándose.

11. Aceleración de la convergencia

11.1 Métodos de relajación

En el método de Jacobi el paso de $x^{(m)}$ a $x^{(m+1)}$ no se puede subdividir en etapas intermedias, en el sentido de que hasta que no se ha calculado por completo $x^{(m+1)}$ el vector que se utiliza en los cálculos es $x^{(m)}$.

Sin embargo, en el caso del método de Gauss-Seidel cada paso se puede subdividir en tantas etapas intermedias como componentes tengan los vectores de la sucesión. En cada una de ellas se modifica una componente de la solución aproximada y el vector que se utiliza en los cálculos es distinto en una componente al pasar de una etapa a otra, como bien sabemos.

Los métodos que son de esta forma; es decir, que en cada paso se altera únicamente una componente del vector de aproximación anterior, se denominan *métodos de relajación*, también se dice que el medio de pasar de un vector aproximación a otro es una *relajación*.

Vimos en el caso de Gauss-Seidel que éste tenía dos ventajas respecto del de Jacobi: por una parte aceleraba generalmente el proceso de convergencia, y por otra necesitaba menor capacidad de almacenamiento disponible.

Vamos a ver cómo proceder para acelerar la convergencia de los métodos iterativos lineales.

11.2 Método SOR

11.2.1 Construcción

El método de Gauss-Seidel es un método de relajación, pero no es el único.

En la descomposición $A = D - L - U$, se puede escribir

$$D = \frac{1}{\omega} D - \frac{1-\omega}{\omega} D \text{ tal que } \omega \neq 0,$$

por lo que

$$A = \frac{1}{\omega} D - \frac{1-\omega}{\omega} D - L - U.$$

Escribiendo

$$M = \frac{1}{\omega} D - L \quad \text{y} \quad N = \frac{1-\omega}{\omega} D + U$$

se tendrá que $A = M - N$ y que M es invertible con sólo exigir que $a_{ii} \neq 0$ cualquiera que sea $i = 1, \ldots, n$, cosa que sabemos que es siempre posible.

Por consiguiente, el método (9.5) $Mx^{(m+1)} = Nx^{(m)} + b$ queda en la forma siguiente:

$$\left(\frac{1}{\omega} D - L\right) x^{(m+1)} = \left(\frac{1-\omega}{\omega} D + U\right) x^{(m)} + b.$$

De aquí que:

$$\begin{aligned}
\tfrac{1}{\omega}a_{11}x_1^{(m+1)} &= \tfrac{1-\omega}{\omega}a_{11}x_1^{(m)} - a_{12}x_2^{(m)} \ldots - a_{1n}x_n^{(m)} + b_1,\\
a_{21}x_1^{(m+1)} + \tfrac{1}{\omega}a_{22}x_2^{(m+1)} &= \tfrac{1-\omega}{\omega}a_{22}x_2^{(m)} \ldots - a_{2n}x_n^{(m)} + b_2,\\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots\\
a_{n1}x_1^{(m+1)} + a_{n2}x_2^{(m+1)} + \ldots + \tfrac{1}{\omega}a_{nn}x_n^{(m+1)} &= \tfrac{1-\omega}{\omega}a_{nn}x_n^{(m)} + b_n.
\end{aligned}$$

En definitiva, para $i = 1, \ldots, n$ se tiene que

$$\frac{1}{\omega}a_{ii}x_i^{(m+1)} + \sum_{j=1}^{i-1} a_{ij}x_j^{(m+1)} = \frac{1-\omega}{\omega}a_{ii}x_j^{(m)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(m)} + b_i. \qquad (11.1)$$

Para calcular $x_i^{(m+1)}$ procederemos en dos etapas:

1) Calculando un valor auxiliar $\overline{x}_i^{(m+1)}$.

2) A partir de éste, calcular $x_i^{(m+1)}$ en la forma siguiente:

hacemos

$$a_{ii}\overline{x}_i^{(m+1)} = \frac{1}{\omega}a_{ii}x_i^{(m+1)} - \frac{1-\omega}{\omega}a_{ii}x_i^{(m)},$$

con lo que

$$a_{ii}\overline{x}_i^{(m+1)} = \frac{1}{\omega}a_{ii}\left(x_i^{(m+1)} - x_i^{(m)} + \omega x_i^{(m)}\right);$$

multiplicando por ω y dividiendo por a_{ii} (que es siempre distinto de 0), queda

$$\omega\overline{x}_i^{(m+1)} = x_i^{(m+1)} - x_i^{(m)} + \omega x_i^{(m)},$$

entonces

$$x_i^{(m+1)} = \omega\overline{x}_i^{(m+1)} + (1-\omega)x_i^{(m)}. \qquad (11.2)$$

Por tanto, (11.1) queda como

$$a_{ii}\overline{x}_i^{(m+1)} + \sum_{j=1}^{i-1} a_{ij}x_j^{(m+1)} = -\sum_{j=i+1}^{n} a_{ij}x_j^{(m)} + b_i. \qquad (11.3)$$

Así de (11.3) se obtiene $\overline{x}_i^{(m+1)}$ y luego de (11.2) se obtiene $x_i^{(m+1)}$.

Para cada valor de ω obtenemos un método distinto. En el caso de que $\omega = 1$ se tiene el método de Gauss-Seidel.

Dependiendo del valor del *peso* ω, la velocidad de convergencia será mayor o menor[1]. El método para una elección óptima de ω recibe el nombre de *método de sobrerrelajación* o simplemente *SOR*.

Sin embargo, el cálculo del valor óptimo de ω es difícil y frecuentemente sólo se obtiene de forma aproximada, basado en varios intentos y observando el comportamiento de la convergencia. En algunos problemas puede merecer la pena el esfuerzo empleado en la estimación de un ω óptimo, dada la notable rapidez de convergencia que se obtiene.

[1] Aunque no hayamos dado una definición precisa de "velocidad de convergencia", intuitivamente podemos entender la misma como una medida que refleja el número de iteraciones necesarias para obtener un grado de precisión preestablecido.

11.2.2 Convergencia

El método de relajación lo hemos escrito, en el epígrafe anterior, como

$$\left(\frac{1}{\omega}D - L\right) x^{(m+1)} = \left(\frac{1-\omega}{\omega}D + U\right) x^{(m)} + b.$$

Multiplicando por ω ($\omega \neq 0$) se tiene que

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

Multiplicando por D^{-1}, tendremos que

$$\left(I - \omega D^{-1}L\right) x^{(m+1)} = \left((1-\omega)I + \omega D^{-1}U\right) x^{(m)} + \omega D^{-1}b$$

y $(I - \omega D^{-1}L) = D^{-1}(D - \omega L)$ es invertible por ser producto de dos matrices que lo son; ya que D obviamente lo es y $(D - \omega L)$, por la forma de las matrices indicada en (9.6), es una matriz triangular inferior con elementos diagonales a_{ii} no nulos (por la regularidad de A), por lo que existe $(D - \omega L)^{-1}$. Luego

$$\begin{aligned} x^{(m+1)} = \; & (I - \omega D^{-1}L)^{-1} ((1-\omega)I + \omega D^{-1}U)\, x^{(m)} + \\ & + \omega\, (I - \omega D^{-1}L)^{-1} D^{-1} b. \end{aligned}$$

Llamando

$$B_\omega = \left(I - \omega D^{-1}L\right)^{-1} \left((1-\omega)I + \omega D^{-1}U\right)$$

y $\hat{b} = \omega(I - \omega D^{-1}L)^{-1}D^{-1}b$,

tendremos que

$$x^{(m+1)} = B_\omega\, x^{(m)} + \hat{b}, \tag{11.4}$$

que es la forma ordinaria de un método iterativo lineal. Dada la consistencia del método, la condición necesaria y suficiente de convergencia ya sabemos que es $\rho(B_\omega) < 1$.

Lo que nos interesa es poner en relación este hecho con el rango de valores que pueda tomar ω para que el método converja.

TEOREMA 33 *Si $B_\omega = (I - \omega D^{-1}L)^{-1}((1-\omega)I + \omega D^{-1}U)$ es la matriz del método de relajación, entonces:*

$$\rho(B_\omega) \geq |\omega - 1|.$$

Demostración.

Los autovalores de B_ω son las soluciones de $|\lambda I - B_\omega| = 0$. Por otra parte D^{-1} es diagonal, por serlo D, por lo que $D^{-1}L$ es triangular inferior con ceros en la diagonal, entonces $|I - \omega D^{-1}L| = 1$.

Por tanto:

$$|B_\omega| = |(I - \omega D^{-1}L)^{-1}|\, |(1-\omega)I + \omega D^{-1}U| = |(1-\omega)I + \omega D^{-1}U|.$$

Además, $D^{-1}U$ es triangular superior con ceros en la diagonal. Entonces, $((1-\omega)I+\omega D^{-1}U)$ es triangular superior con $(1-\omega)$ en la diagonal. Luego

$$|B_\omega| = (1-\omega)^n. \tag{11.5}$$

Por otra parte, sabemos que el determinante de una matriz es igual al producto de sus autovalores, luego

$$|B_\omega| = \prod_{i=n}^{n} \lambda_i, \text{ tal que } \lambda_i \text{ es autovalor de } B_\omega. \tag{11.6}$$

De (11.5) y (11.6) se tiene que

$$\prod_{i=n}^{n} |\lambda_i| = |(1-\omega)^n|.$$

Entonces

$$\rho(B_\omega) \geq |1-\omega|.$$

■

Una consecuencia inmediata de este teorema es que:

El método de relajación sólo puede converger si $0 < \omega < 2$.

En efecto, ya que el método converge si y sólo si $\rho(B_\omega) < 1$, se tendrá que el método diverge si $|\omega - 1| \geq 1$, por lo que el método diverge para $\omega \geq 2$ y para $\omega \leq 0$, y por lo tanto sólo será posible la convergencia si $0 < \omega < 2$.

■

Apéndice

Código C para las rutinas usuales

FUNCIONES PARA LA ELIMINACION DE GAUSS

```
/*
    Funciones
      factor()        factoriza la matriz A mediante eliminación
                      gaussiana con pivoteo parcial y equilibrio
                      implícito

      resuelve()      resuelve el sistema Ax = b, donde A se supone
                      factorizada por la función anterior, por
                      sustitución hacia atrás

      tridiagonal()   resuelve el sistema Ax = b, donde A se supone
                      tridiagonal

      factorcond()    factoriza una matriz A mediante una llamada a
                      factor() y hace una estimación del número de
                      condición de A

      cholesky()      factoriza la matriz simétrica definida positiva
                      A en la forma LL^t

      reschol()       resuelve el sistema Ax = b donde A ha sido
                      factorizada por cholesky()

      jacobi()        resuelve Ax = b por iteraciones del método de
                      Jacobi
      seidelsr()      resuelve Ax = b por iteraciones del método de
                      Gauss-Seidel son sobre-relajación (para ω = 1
                      se obtiene el método de Gauss-Seidel)

      códigos de salida  -1  Memoria insuficiente
                          0  Éxito
                          1  Peligro. Matriz casi-singular
                          2  Fracaso. Matriz singular
*/
```

```
#include <stdlib.h>
#include <stdio.h>
#include <math.h>

#define A(i,j) A[n*i+j]

 int factor (A,n,pivote)
   double *A;
   int n;
   int *pivote;

 {
   double abs A, tmp, det;
   int i,j,k;
   double *s;
   double *c;

   /* Asigna memoria dinámica a los vectores s y c */

   s=calloc(n,sizeof(double));
   if (s==NULL) return(-1);

   s=calloc(n,sizeof(double));
   if(c==NULL) return(-1);

   /* Calcula el máximo de cada fila y lo almacena en el vector s */

   det=1.;
   for (i=0; i<n; i++)
     {
       s[i]=0;
       for (j=0; j<n; j++)
         {
           absA=fabs(A(i,j));
           if (absA > s[i])   s[i]=absA;
         }
     }
   /* Hace elimininación de Gauss con pivoteo parcial y equilibrio
   implícito */

   for (k=0; k<n-1; k++)
     {
       c[k]=0;
```

```
        pivote[k]=k;
        for (i=k; i<n; i++)
          {
            absA=fabs (A(i,k)/s[i]);
            if (absA>c[k])
              {
                c[k]=absA;
                pivote[k]=i;
              }
          }

        if (c[k]==0)
          {
            free(s);
            free(c);
            return(2);
          }

        if (pivote[k] !=k)
          {
            det=-det;
            for (j=0; j<n;j++)
              {
                tmp=A(k,j);
                A(k,j)=A(pivote[k],j);
                A(pivote[k],j)=tmp;
              }

            tmp=s[k];
            s[k]=s[pivote[k]];
            s[pivote[k]]=tmp;
          }

        for (i=k+1; i<n; i++)
          {
            A(i,k)=A(i,k) /A(k,k);
            for (j=k+1; j<n; j++)
              A(i,j)=A(i,j)-A(i,k)*A(k,j);
          }

        det=A(k,k)*det;
      }

    free(s);
    free(c);
    if (fabs(det)< 1.0e-10) return(1);
    return(0);
}
```

```
int resuelve (A,n,b,pivote)
     double *A;
     int n;
     double *b;
     int *pivote;
{
  int k,i,j;
  double tmp;

  */Construye la estrategia de pivoteo para el vector b */

  for (k=0,k<n-1; k++)
    {
      i=pivote[k];
      if (i!=k)
        {
          tmp=b[k];
          b[k]=b[i];
          b[i]=tmp;
        }
    }

  /* Construye la eliminación de Gauss para b */

  for (i=1; i<n; i++)
    {
      for (j=0,  tmp=0; j<i; j++)
        tmp=tmp+b[j]*A(i,j);
      b[i]=b[i]-tmp;
    }

  /* Calcula la solución por sustitución hacia atrás */

  b[n-1]=b[n-1]/A((n-1),n-1);
  for (i=n-2; i>=0; i--)
    {
      tmp=0.0;
      for (j=i+1; j<n; j++)
        tmp+=A(i,j)*b[j];

      b[i]=(b[i]-tmp)/A(i,i);
    }

  return (0);
}
```

```
int factorcond(A,n,pivote,cond)
     double *A;
     int n;
     int *pivote;
     double *cond;

{
  int i,j, result;
  double tmp;
  double *y;

  /* Asigna memoria dinámica al vector y */

  y=calloc (n,sizeof(double));
  if (y==NULL) return (-1);

  /* Calcula la norma-infinito de A para cond */

  for (i=0, *cond=0; i<n; i++)
    {
      for (j=0, tmp=0; j<n; j++)
        tmp=tmp+fabs(A(i,j));

      if (tmp > *cond) *cond=tmp;
    }

  /* Factoriza A */

  result=factor (A,n,pivote);
  if (result == -1 || result==2)
    {
      free(y);
      return(result);
    }

  /* Calcula la estimación de la norma-infinito de A^-1 para cond */

  y[0]=1;
  for (i=1; i<n; i++)
    {
      for (j=0, tmp=0; j<i; j++)
        tmp=tmp+y[j]*A(i,j);
      if (tmp <0)  y[i]=1.-tmp;
      else         y[i]=-1.-tmp
    }

  y[n-1]=y[n-1] / A((n-1),(n-1));
```

```
  for (i=n-2; i>=0; i--)
    {
      for (j=i+1, tmp=0; j<n; j++)
        tmp=tmp+y[j]*A(i,j);
      y[i]=(y[i]-tmp)/A(i,j);
    }

  for (i=0, tmp=0; i<n; i++)
    if (fabs(y[i]) > tmp) tmp=fabs(y[i]);

  *cond=*cond *tmp;
  free(y);
  return(result);
}



int tridiagonal (A,n,b)
     double A[ ][3];
     int n;
     double b[ ];

{
  int i;

  for (i=1; i<n;i++)
    {
      if (A[i-1][1]==0.0)
        return (-1);
      A[i][0]=A[i][0]/A[i-1][1];
      A[i][1]=A[i][1]-A[i][0]*A[i-1][2];
      b[i]=b[i]-A[i][0]*b[i-1];
    }

  b[n-1]=b[n-1] / A[n-1][1];

  for (i=n-2; i>=0; i--)
    b[i]=(b[i]-A[i][2]*b[i+1])/A[i][1];

  return(0);
}
```

```
int cholesky(A,n)
     double *A;
     int n;
{
  double aux;
  int i, j, k;

  for (i=0; i<n-1; i++)
    for (j=i+1; j<n; j++)
      A(i,j)=0.0;

  for (i=0; i<n; i++)
    {
      aux=0.0;
      for (k=0; k<i; k++)
        aux+=A(i,k)*A(i,k);
      aux=A(i,i)-aux;
      if (aux < 0.0) return (-1);

      A(i,i)=sqrt(aux);
      for (j=i+1; j<n; j++)
        {
          aux=0.0;
          for (k=0; k<i; k++)
            aux+=A(j,k)*A(i,k);
          A(j,i)=(A(j,i)-aux)/A(i,i);
        }
    }
  return(0);
}



int reschol(A,n,b)
     double *A;
     int n;
     double b[];
{
  double aux;
  int i, k;

  for (i=0; i<n; i++)
    {
      aux=0.0;
      for (k=0; k<i; k++)
        aux+=A(i,k)*b[k];
      b[i]=(b[i]-aux)/A(i,i);
    }
```

```
  for (i=n-1; i>=0; i--)
    {
      aux=0.0;
      for (k=i+1; k<n; k++)
        aux+=A(k,i)*b[k];

      b[i]=(b[i]-aux)/A(i,i);
    }

  return (0);
}



int jacobi (A, b, n, xm, iter, tol)
    double *A;
    double b[];
    int n;
    double xm[];
    int iter;
    double tol;
{
  double *ym;
  double nym, ndif, suma;
  int i, j, m;

  ym=calloc(n, sizeof(double));
  if (ym==NULL) return (-1);

  for (m=1; m<=iter; m++)
    {
      for (i=0, ndif=nym=0; i<n; i++)
        {
          for (j=0, suma=0; j<n; j++)
            if (j!=i)
              suma=suma+A(i,j)*xm[j];

          ym[i]=(b[i]-suma)/A(i,i);

          if (fabs(ym[i]-xm[i])> ndif) ndif=fabs(ym[i]-xm[i]);
          if (fabs(ym[i])>nym) nym=fabs(ym[i]);
        }

      for (i=0; i<n; i++)
        xm[i]=ym[i];
```

```
    if (ndif/nym < tol)
      {
        free(ym);
        return(m);
      }
    }
  free(ym);
  return (m-1);
}


int seidelsr (A, b, n, xm, iter, tol, omega)
    double *A;
    double b[ ];
    int n;
    double xm[ ];
    int iter;
    double tol;
    double omega;
{
  double nym, ndif, suma, z, xiant;
  int i, j, m;

  for (m=1; m<=iter; m++)
    {
      for (i=0, ndif=nym=0; i<n; i++)
        {
          for (j=0, suma=0; j<n; j++)
            if (j!=i)
              suma=suma+A(i,j)*xm[j];

          xiant=xm[i];
          z=(b[i]-suma)/A(i,i);
          xm[i]=omega*z+(1-omega)*xm[i];

          if (fabs(xm[i]-xiant)>ndif) ndif=fabs(xm[i]-xiant);
          if (fabs(xm[i]>nym)  nym=fabs(xm[i]);
        }

      if (ndif/nym < tol)
        return(m);
    }

  return (m-1);
}
```

Bibliografía

[1] ABAFFY, J. - SPEDICATO, E. (1989). *ABS Projection algorithms: mathematical techniques for linear and non linear equations.* Chichester, England (UK) , Ellis Horwood Limited.

[2] ANTON, H. (1990). *Introducción al Algebra Lineal.* México, Limusa.

[3] ATKINSON, K. E. (1989). *An introduction to numerical analysis.* Nueva York (USA), John Wiley & Sons.

[4] BAKHALOV, N. (1980). *Métodos Numéricos.* Madrid, Paraninfo.

[5] BARNETT, S. (1990). *Matrices: Methods and Applications.* Oxford (UK), Clarendon Press.

[6] BELLMAN, R. (1965). *Introducción al Análisis Matricial.* Barcelona, Reverté.

[7] BURDEN, R. & FAIRES, D. (1985). *Análisis Numérico.* México (México), Grupo Editorial Iberoamericana.

[8] CARBO, R. - HERNANDEZ, J.A. (1983). *Introducción a la Teoría de Matrices.* Madrid, Alhambra.

[9] CHAPRA, S.C. - CANALE, R.P. (1987). *Métodos Numéricos para ingenieros.* México, McGraw-Hill.

[10] DEMIDOVICH, B.P. - MARON, I.A. (1988). *Cálculo Numérico Fundamental.* Madrid, Paraninfo.

[11] DUFF, I.S. - ERISMAN, A.M. and REID, J.K. (1989). *Direct Methods for Sparse Matrices.* Oxford (UK) , Oxford Science Publications.

[12] FRÖBERG, C.E. (1977). *Introducción al Análisis Numérico.* Barcelona, Vicens-Vives

[13] GANTMACHER, S. (1960). *The Theory of Matrices.* New York (USA), Chelsea Publishing Company.

[14] GASCA, M. (1976). *Cálculo Numérico.* Madrid, UNED.

[15] GASCA, M. (1987). *Cálculo Numérico: Resolución de ecuaciones y sistemas.* Zaragoza, Librería Central.

[16] GASTINEL, N. (1975). *Análisis Numérico Lineal.* Barcelona, Reverté.

[17] GERALD, C.F. (1991). *Análisis Numérico.* México, Alfaomega.

[18] GILL, P.E., MURRAY, W. & WOIGHT, M.H. (1991). *Numerical Linear Algebra and Optimization.* San Diego, California (USA), Addison-Wesley Publishing Company.

[19] GROSSMAN, S.I. (1991). *Algebra Lineal con aplicaciones.* México, McGraw-Hill.

[20] HADJIDIMOS, A. (editor). (1988). *Iterative methods for the solution of linear systems.* Amsterdams (The Netherlans), North-Holland.

[21] HULTQUIST, P. F. (1988). *Numerical Methods for engineers and Computer scientists.* San Diego, California (USA), The Benjamin / Cummings Publishing Company.

[22] ISAACSON, E.-KELLER, H.B. (1966). *Analysis of Numerical Methods.* New York (USA), John Wiley & Sons.

[23] KINCAID, D.R. - HAYES, L.J. (editors). (1990). *Iterative Methods for Large Linear Systems.* San Diego, California (USA), Academic Press.

[24] MICHAVILLA,F. - GAVETE, L. (1985). *Programación y Cálculo Numérico.* Barcelona, Reverté.

[25] NAKAMURA, S. (1992). *Métodos Numéricos aplicados con software.* México, Prentice-Hall Hispanoamericana.

[26] NOBLE, B. - DANIEL, J.W. (1989). *Algebra Lineal Aplicada.* México, Prentice-Hall Hispanoamericana.

[27] RALSTON, A. - RABINOWITZ, P. (1978). *A first Course in Numerical Analysis.* Auckland (USA), McGraw-Hill Book Company.

[28] RICE, J.R. (1985). *Numerical Methods, Software, and Analysis.* Auckland (USA), MacGraw-Hill Book Company.

[29] RORRES, C. - ANTON, H. (1979). *Aplicaciones de Algebra lineal.* México, Limusa.

[30] SCHEID, F. (1972). *Análisis Numérico.* México, McGraw-Hill.

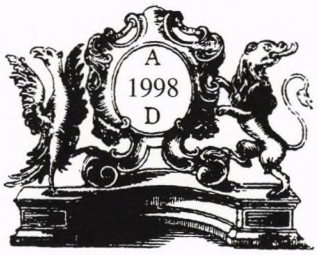
[31] SCHEID, F. - DI CONSTANZO, R.E. (1991). *Métodos Numéricos.* México, McGraw-Hill.

[32] SCHENDEL, U. (1989). *Sparse Matrices: numerical aspects with applications for scientists and engineers.* Chichester, England (UK) , Ellis Horwood Limited.

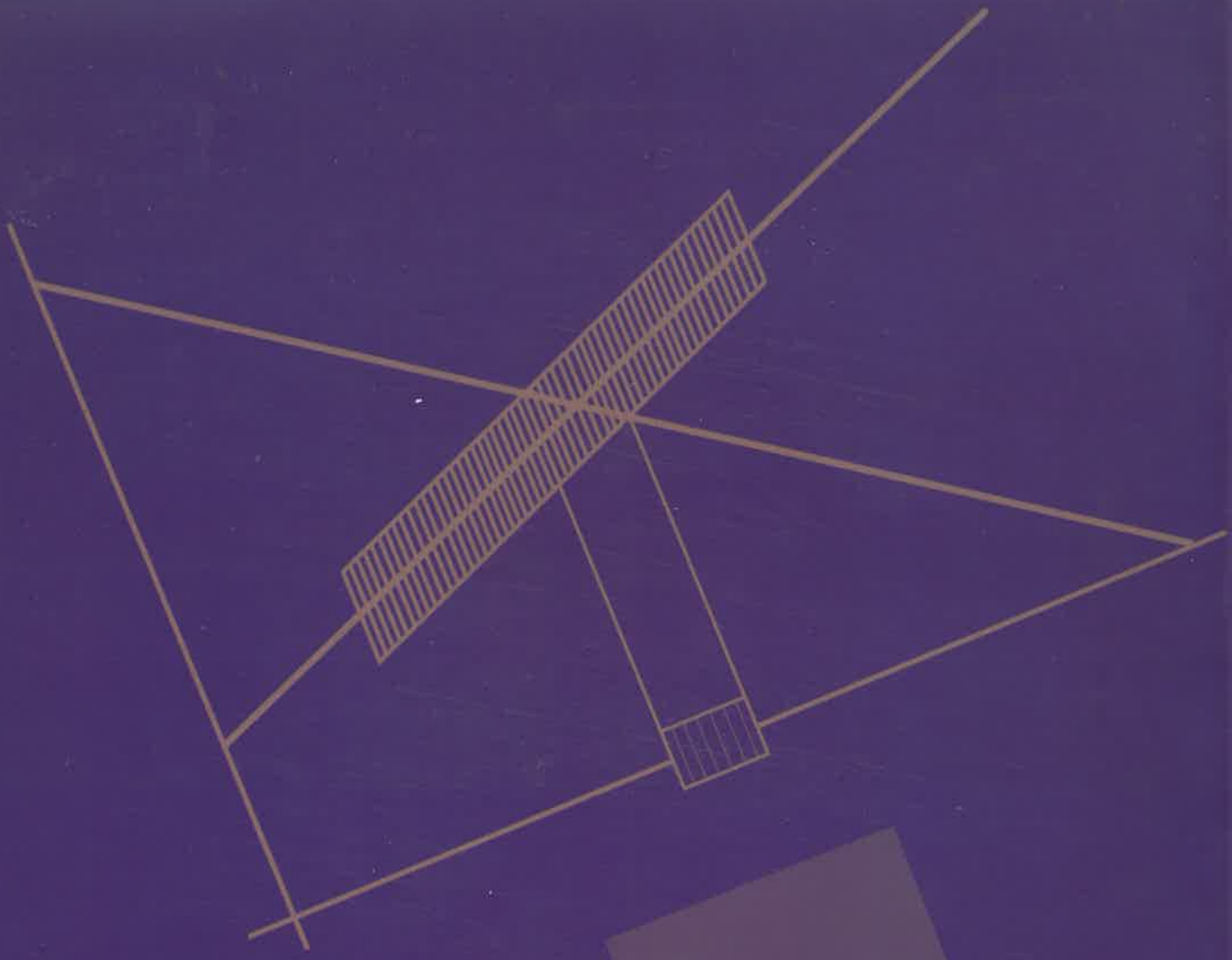
[33] STRANG, G. (1986). *Algebra Lineal y sus aplicaciones.* Delaware (USA), Addison-Wesley Iberoamericana.

[34] TORREGROSA, J.R. - JORDAN, C. (1987). *Algebra Lineal y sus aplicaciones.* Madrid, McGraw-Hill.

[35] WILKINSON, J.H. (1965). *The Algebraic Eigenvalue Problem.* Oxford (England), Oxford University Press.



Se terminó de componer este libro
el día 24 de octubre,
festividad tradicional del arcángel Rafael,
compañero de Tobías
y uno de los que asisten delante de Dios.



SERVICIO·DE·PUBLICACIONES

UNIVERSIDAD·DE·CADIZ