Métodos Numéricos II
Segundo de Grado en Matemáticas, Curso 2021/22
Facultad de Ciencias de la Universidad de Málaga
Dpto. de Análisis Matemático, Estadística e Investigación Operativa, y Matemática Aplicada
Profs. Manuel J. Castro y Francisco J. Palma (Área de Conocimiento Mat. Aplicada)

TEMA 2

Resolución de Sistemas de Ecuaciones Lineales

Los objetivos de este tema son:

- motivar la necesidad de métodos numéricos para la resolución de sistemas de ecuaciones lineales;
- condicionamiento de un sistema lineal;
- métodos directos de resolución de sistemas de ecuaciones lineales: Gauss, Gauss-Jordan, factorización LU, factorización de Cholesky, factorización QR, etc.;
- métodos iterativos de resolución de sistemas de ecuaciones lineales: Jacobi, Gauss-Seidel, relajación, etc.

1. Introducción

Problema 1 *Dados $A \in \mathcal{M}_n(\mathbb{K})$ inversible y $B \in \mathbb{K}^n$, calcular $X \in \mathbb{K}^n$ tal que*

$$A\,X = B\,.$$

■

Teorema 2 *El Problema 1 posee una y sólo una solución.* ■

Las conocidas **fórmulas de Cramer** proporcionan un método directo de resolución del Problema 1. Sin embargo, dichas fórmulas necesitan del cálculo de $n+1$ determinantes distintos, cada uno de ellos de orden n; en consecuencia, el volumen total de cálculos a realizar se eleva a:

$$\begin{array}{cl} (n+1)! - n - 1 & \text{sumas,} \\ (n+1)!\,(n-1) & \text{multiplicaciones,} \\ n & \text{divisiones.} \end{array}$$

Para $n = 20$, solamente el número de multiplicaciones a efectuar es superior a 10^{21}, lo que hace que el método sea inviable, aún disponiendo del mejor de los ordenadores existentes.

Jamás se resuelve el Problema 1 calculando la **inversa** de la matriz A y multiplicándola posteriormente por B, ya que si ponemos

$$A^{-1} = (X_1|X_2|\cdots|X_n)$$

(cada X_i es la i-ésima columna de A^{-1}), entonces

$$A\,A^{-1} = I \iff A\,(X_1|X_2|\cdots|X_n) = (E_1|E_2|\cdots|E_n)\ ,$$

y por tanto

$$A\,X_i = E_i\,, \quad i = 1, 2, \ldots, n\,,$$

lo que demuestra que el cálculo de la matriz A^{-1}, columna por columna, es equivalente a la resolución de n sistemas de ecuaciones lineales en los que aparecen como segundos miembros los vectores E_i de la base canónica de $\mathbb{K}^n$. Se reemplazaría, pues, la resolución de un único sistema lineal por el de n sistemas lineales (aunque todos con la misma matriz A), más una posterior multiplicación matriz por vector.

Si en el Problema 1 la matriz A es **triangular,** la resolución del sistema lineal es inmediata; en efecto, el sistema triangular superior

$$\left\{\begin{array}{ccccccccccc} a_{1,1}\,x_1 & + & a_{1,2}\,x_2 & + & \cdots & + & a_{1,n-1}\,x_{n-1} & + & a_{1,n}\,x_n & = & b_1 \\ & & a_{2,2}\,x_2 & + & \cdots & + & a_{2,n-1}\,x_{n-1} & + & a_{2,n}\,x_n & = & b_2 \\ & & & & \ddots & & \vdots & & \vdots & & \vdots \\ & & & & & & a_{n-1,n-1}\,x_{n-1} & + & a_{n-1,n}\,x_n & = & b_{n-1} \\ & & & & & & & & a_{n,n}\,x_n & = & b_n \end{array}\right\},$$

tiene como solución

$$\left\{\begin{array}{rcll} x_n & = & {a_{n,n}}^{-1} & b_n \\ x_{n-1} & = & {a_{n-1,n-1}}^{-1} & (b_{n-1} \quad - a_{n-1,n}\,x_n) \\ \vdots & & \vdots & \qquad\qquad \ddots \\ x_2 & = & {a_{2,2}}^{-1} & (b_2 \quad - \cdots - a_{2,n-1}\,x_{n-1} - a_{2,n}\,x_n) \\ x_1 & = & {a_{1,1}}^{-1} & (b_1 \quad - a_{1,2}\,x_2 - \cdots - a_{1,n-1}\,x_{n-1} - a_{1,n}\,x_n) \end{array}\right\}.$$

Notamos que al ser A triangular e inversible, su determinante (no nulo) es el producto de los elementos de su diagonal principal, por lo que todos los cálculos indicados tienen sentido (salvo errores de redondeo). El proceso presentado se llama de **resolución por remonte**; si la matriz A es triangular inferior, se empieza a resolver x_1 por la primera ecuación y el proceso se llama de **resolución por descenso.** En cualquier caso, el número total de operaciones a realizar es de:

$$\begin{array}{cl} \frac{1}{2}\,(n^2-n) & \text{sumas,} \\ \frac{1}{2}\,(n^2-n) & \text{multiplicaciones,} \\ n & \text{divisiones.} \end{array}$$

Los métodos de remonte y descenso se extienden de forma natural a las **matrices triangulares por bloques**; así, por ejemplo, la resolución del sistema lineal

$$\left(\begin{array}{c|c|c} A_{1,1} & A_{1,2} & A_{1,3} \\ \hline & A_{2,2} & A_{2,3} \\ \hline & & A_{3,3} \end{array}\right)\left(\begin{array}{c} X_1 \\ \hline X_2 \\ \hline X_3 \end{array}\right) = \left(\begin{array}{c} B_1 \\ \hline B_2 \\ \hline B_3 \end{array}\right),$$

se reduce a la resolución de los sistemas lineales sucesivos

$$\left\{\begin{array}{rcl} A_{3,3}\,X_3 & = & B_3 \\ A_{2,2}\,X_2 & = & B_2 - A_{2,3}\,X_3 \\ A_{1,1}\,X_1 & = & B_1 - A_{1,2}\,X_2 - A_{1,3}\,X_3 \end{array}\right\}.$$

2. Condicionamiento de un sistema lineal

Consideremos el sistema de ecuaciones lineales $A\,X = B$ (este **ejemplo** es debido a R. S. Wilson):

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \; ; \quad \text{solución:} \quad X = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} .$$

Consideremos ahora el sistema perturbado $A\,\bar{X} = \bar{B}$, donde los segundos miembros se modifican muy ligeramente ($\bar{B} = B + \delta B$) y la matriz se deja intacta (ponemos $\bar{X} = X + \delta X$):

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \bar{x}_4 \end{pmatrix} = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix} \; ; \quad \text{solución:} \quad \bar{X} = \begin{pmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{pmatrix} ;$$

como puede observarse, un error relativo menor que $\frac{1}{200}$ en los datos, implica un error relativo en los resultados que puede ser superior a 10, es decir, ha habido un factor mayor a 2000 en la amplificación de los errores.

Consideremos finalmente el sistema perturbado $\bar{A}\,\bar{X} = B$, donde ahora es la matriz la que varía muy ligeramente ($\bar{A} = A + \Delta A$) mientras que el segundo miembro se deja intacto (ponemos $\bar{X} = X + \Delta X$):

$$\begin{pmatrix} 10. & 7. & 8.1 & 7.2 \\ 7.08 & 5.04 & 6. & 5. \\ 8. & 5.98 & 9.89 & 9. \\ 6.99 & 4.99 & 9. & 9.98 \end{pmatrix} \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \bar{x}_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \; ; \quad \text{solución:} \quad \bar{X} = \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix} ;$$

aquí también, pequeñas variaciones en los datos han provocado una modificación total de los resultados.

Para terminar este ejemplo, digamos que todos los cálculos que se han realizado son exactos (no hay errores de redondeo), las modificaciones que se han introducido son aceptables desde un punto de vista práctico, la matriz A (y el segundo miembro B) tiene todos sus elementos naturales, es simétrica, su determinante vale 1 y su inversa es

$$A^{-1} = \begin{pmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{pmatrix} .$$

Definición 3 *Dadas $A \in \mathcal{M}_n(\mathbb{K})$ inversible y $\|\cdot\|$ norma matricial subordinada, se llama* **condicionamiento** *de A respecto de la norma dada a*

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

un sistema de ecuaciones lineales de matriz A se dice **bien** *o* **mal condicionado**, *según que el número $\mathrm{cond}(A)$ sea pequeño o grande respectivamente.* ■

Teorema 4 *Dados $A \in \mathcal{M}_n(\mathbb{K})$ inversible y $B, \delta B \in \mathbb{K}^n$, sean $X, \delta X \in \mathbb{K}^n$ tales que, llamando $\bar{B} = B + \delta B$ y $\bar{X} = X + \delta X$, se tiene que*

$$\begin{aligned} A\,X &= B\,, \\ A\,\bar{X} &= \bar{B}\,. \end{aligned}$$

Supuesto $B \neq 0$, entonces

$$\frac{\|\delta X\|}{\|X\|} \leq \mathrm{cond}(A)\,\frac{\|\delta B\|}{\|B\|}\,,$$

siendo esta desigualdad la mejor posible en el sentido de que existen $B \neq 0$ y $\delta B \neq 0$ para los que se alcanza la igualdad. ■

Teorema 5 *Dados $A, \Delta A \in \mathcal{M}_n(\mathbb{K})$, A inversible, y $B \in \mathbb{K}^n$, sean $X, \Delta X \in \mathbb{K}^n$ tales que, llamando $\bar{A} = A + \Delta A$ y $\bar{X} = X + \Delta X$, se tiene que*

$$\begin{aligned} A\,X &= B\,, \\ \bar{A}\,\bar{X} &= B \end{aligned}$$

(no es preciso que $\bar{A}$ sea inversible, basta con que el sistema perturbado tenga al menos una solución). Supuesto $B \neq 0$, entonces

$$\frac{\|\Delta X\|}{\|\bar{X}\|} \leq \mathrm{cond}(A)\,\frac{\|\Delta A\|}{\|A\|}\,,$$

siendo esta desigualdad la mejor posible en el sentido de que existen $B \neq 0$ y $\Delta A \neq 0$ para los que se alcanza la igualdad. ■

Corolario 6 *Si a las hipótesis del Teorema 5 le añadimos que $\|\Delta A\| < \|A^{-1}\|^{-1}$ (lo que garantiza tiene una única solución), se verifica entonces que*

$$\frac{\|\Delta X\|}{\|X\|} \leq \mathrm{cond}(A)\,\frac{\|\Delta A\|}{\|A\|}\,\frac{1}{1 - \|A^{-1}\|\,\|\Delta A\|}\,.$$

■

Lo más **habitual** es utilizar las normas vectoriales $\|\cdot\|_p$, $p = 1, 2, \infty$. Como el condicionamiento de una matriz está asociado a la elección de una norma matricial subordinada, notamos $\mathrm{cond}_p(A)$, $p = 1, 2, \infty$, cuando hacemos referencia a alguna de las normas subordinadas anteriores.

Teorema 7 *Dada $A \in \mathcal{M}_n(\mathbb{K})$ inversible, se verifica:*

1.

$$\begin{aligned} \mathrm{cond}(A) &\geq 1\,; \\ \mathrm{cond}(A) &= \mathrm{cond}(A^{-1})\,; \\ \mathrm{cond}(A) &= \mathrm{cond}(\alpha\, A)\,, \quad \textit{para todo} \quad \alpha \in \mathbb{K} - \{0\}\,; \end{aligned}$$

2.

$$\mathrm{cond}_2(A) = \frac{\mathrm{máx}\{\mu \,:\, \mu \textit{ es valor singular de } A\}}{\mathrm{mín}\{\mu \,:\, \mu \textit{ es valor singular de } A\}}\,;$$

3. si A es normal, entonces

$$\mathrm{cond}_2(A) = \frac{\mathrm{máx}\{|\lambda| \,:\, \lambda \in \mathrm{sp}(A)\}}{\mathrm{mín}\{|\lambda| \,:\, \lambda \in \mathrm{sp}(A)\}}\,;$$

4. si A es ortogonal-unitaria, entonces

$$\mathrm{cond}_2(A) = 1\,;$$

5. $\mathrm{cond}_2(\cdot)$ es invariante por transformación ortogonal-unitaria, es decir,

$$\mathrm{cond}_2(A) = \mathrm{cond}_2(U\,A) = \mathrm{cond}_2(A\,U) = \mathrm{cond}_2(U^{-1}\,A\,U)\,,$$

para toda $U \in \mathcal{M}_n(\mathbb{K})$ ortogonal-unitaria.

■

Consecuencias del Teorema 7:

- con vistas al condicionamiento, los mejores sistemas a resolver son aquellos de matriz ortogonal-unitaria;
- el condicionamiento de un sistema no se mejora multiplicando toda la matriz por una constante, pero sí se puede mejorar multiplicando todas o algunas filas o columnas por distintas constantes (problema del **equilibrado** de una matriz) o por otra matriz arbitraria (problema del **precondicionado** de una matriz).

En el ejemplo de R. S. Wilson se tiene que $\mathrm{cond}_2(A) \approx 2984$.

3. Método de Gauss

El **método de Gauss** es un método directo de resolución del Problema 1 en el que se pueden diferenciar tres etapas:

1. determinación de una matriz $M \in \mathcal{M}_n(\mathbb{K})$ inversible tal que $M\,A$ sea triangular superior;
2. cálculo simultáneo del vector $M\,B$;
3. resolución mediante un proceso de remonte del sistema lineal

$$M\,A\,X = M\,B\,.$$

Evidentemente la matriz M no se llega a calcular nunca explícitamente para después multiplicarla por A y B; en la práctica se calcula directamente la matriz $M\,A$ y el vector $M\,B$ mediante el proceso iterativo finito de eliminación que se describe a continuación.

Descripción de la primera iteración del proceso de eliminación.

Sea $(A_1 \,|\, B_1) = (a^1_{i,j} \,|\, b^1_i) = (A \,|\, B)$, es decir,

$$(A_1 \,|\, B_1) = \left(\begin{array}{cccc|c} a^1_{1,1} & a^1_{1,2} & \cdots & a^1_{1,n} & b^1_1 \\ a^1_{2,1} & a^1_{2,2} & \cdots & a^1_{2,n} & b^1_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a^1_{n,1} & a^1_{n,2} & \cdots & a^1_{n,n} & b^1_n \end{array}\right) = \left(\begin{array}{cccc|c} a_{1,1} & a_{1,2} & \cdots & a_{1,n} & b_1 \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} & b_n \end{array}\right).$$

Como A_1 es regular, al menos uno de los elementos $a^1_{i,1}$, $i = 1, 2, \ldots, n$, debe ser no nulo; sea $a^1_{i_1,1}$ uno de tales elementos (estrategia para la elección del mismo) que se llama primer pívot del proceso de eliminación.

Si $i_1 \neq 1$ se permutan entre sí las filas 1 e i_1, lo cual equivale en escritura matricial a multiplicar la matriz $(A_1 \,|\, B_1)$ por la izquierda por la matriz $P_{\sigma_{i_1,1}} \in \mathcal{M}_n(\mathbb{K})$ de permutación elemental.

En cualquier caso notamos

$$P_1 = \begin{cases} I & \text{si } a^1_{1,1} \text{ es el primer pívot,} \quad \text{en este caso } \det(P_1) = 1, \\ P_{\sigma_{i_1,1}} & \text{si } a^1_{i_1,1},\ i_1 \neq 1, \text{ es el primer pívot,} \quad \text{en este caso } \det(P_1) = -1. \end{cases}$$

Sea $P_1\,(A_1\,|\,B_1) = (\alpha^1_{i,j}\,|\,\beta^1_i)$, donde $\alpha^1_{1,1} \neq 0$, es decir,

$$P_1\,(A_1\,|\,B_1) = \left(\begin{array}{c|ccc|c} \boxed{\alpha^1_{1,1}} & \alpha^1_{1,2} & \cdots & \alpha^1_{1,n} & \beta^1_1 \\ \alpha^1_{2,1} & \alpha^1_{2,2} & \cdots & \alpha^1_{2,n} & \beta^1_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha^1_{n,1} & \alpha^1_{n,2} & \cdots & \alpha^1_{n,n} & \beta^1_n \end{array}\right).$$

Mediante combinaciones lineales de la primera fila, se anulan todos los elementos de la primera columna situados por debajo de la diagonal principal, dejando la primera fila inalterada; esto es equivalente a multiplicar por la izquierda la matriz $P_1\,(A_1\,|\,B_1)$ por la matriz

$$E_1 = \left(\begin{array}{c|ccc} 1 & & & \\ \hline -\dfrac{\alpha^1_{2,1}}{\alpha^1_{1,1}} & 1 & & \\ \vdots & & \ddots & \\ -\dfrac{\alpha^1_{n,1}}{\alpha^1_{1,1}} & & & 1 \end{array}\right) \in \mathcal{M}_n(\mathbb{K})\,.$$

Sea $(A_2\,|\,B_2) = (a^2_{i,j}\,|\,b^2_i) = E_1\,P_1\,(A_1\,|\,B_1) = E_1\,P_1\,(A\,|\,B)$, es decir,

$$(A_2\,|\,B_2) = \left(\begin{array}{c|ccc|c} a^2_{1,1} & a^2_{1,2} & \cdots & a^2_{1,n} & b^2_1 \\ \hline & a^2_{2,2} & \cdots & a^2_{2,n} & b^2_2 \\ & \vdots & \ddots & \vdots & \vdots \\ & a^2_{n,2} & \cdots & a^2_{n,n} & b^2_n \end{array}\right) = \left(\begin{array}{c|ccc|c} \alpha^1_{1,1} & \alpha^1_{1,2} & \cdots & \alpha^1_{1,n} & \beta^1_1 \\ \hline & a^2_{2,2} & \cdots & a^2_{2,n} & b^2_2 \\ & \vdots & \ddots & \vdots & \vdots \\ & a^2_{n,2} & \cdots & a^2_{n,n} & b^2_n \end{array}\right).$$

Además, como $\det(A_2) = \det(E_1)\,\det(P_1)\,\det(A_1) = \pm\det(A_1) = \pm\det(A)$, la matriz A_2 sigue siendo regular.

Descripción de la k-ésima iteración del proceso de eliminación.

Sea $(A_k\,|\,B_k) = (a^k_{i,j}\,|\,b^k_i)$ el resultado de la $(k-1)$-ésima iteración del proceso de eliminación, por lo que $(A_k\,|\,B_k) = E_{k-1}\,P_{k-1}\,(A_{k-1}\,|\,B_{k-1}) = \cdots = E_{k-1}\,P_{k-1}\cdots E_1\,P_1\,(A\,|\,B)$, es decir,

$$(A_k\,|\,B_k) \;=\; \left(\begin{array}{ccc|ccc|c} a^k_{1,1} & \cdots & a^k_{1,k-1} & a^k_{1,k} & a^k_{1,k+1} & \cdots & a^k_{1,n} & b^k_1 \\ & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & a^k_{k-1,k-1} & a^k_{k-1,k} & a^k_{k-1,k+1} & \cdots & a^k_{k-1,n} & b^k_{k-1} \\ \hline & & & a^k_{k,k} & a^k_{k,k+1} & \cdots & a^k_{k,n} & b^k_k \\ & & & a^k_{k+1,k} & a^k_{k+1,k+1} & \cdots & a^k_{k+1,n} & b^k_{k+1} \\ & & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & & a^k_{n,k} & a^k_{n,k+1} & \cdots & a^k_{n,n} & b^k_n \end{array}\right)$$

$$= \left(\begin{array}{ccc|ccccc} \alpha^1_{1,1} & \cdots & \alpha^1_{1,k-1} & \alpha^1_{1,k} & \alpha^1_{1,k+1} & \cdots & \alpha^1_{1,n} & \beta^1_1 \\ & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & \alpha^{k-1}_{k-1,k-1} & \alpha^{k-1}_{k-1,k} & \alpha^{k-1}_{k-1,k+1} & \cdots & \alpha^{k-1}_{k-1,n} & \beta^{k-1}_{k-1} \\ \hline & & & a^k_{k,k} & a^k_{k,k+1} & \cdots & a^k_{k,n} & b^k_k \\ & & & a^k_{k+1,k} & a^k_{k+1,k+1} & \cdots & a^k_{k+1n} & b^k_{k+1} \\ & & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & & a^k_{n,k} & a^k_{n,k+1} & \cdots & a^k_{n,n} & b^k_n \end{array}\right).$$

Como A_k es regular, al menos uno de los elementos $a^k_{i,k}$, $i = k, k+1, \ldots, n$, debe ser no nulo; sea $a^k_{i_k,k}$ uno de tales elementos (estrategia para la elección del mismo) que se llama k-ésimo pívot del proceso de eliminación.

Si $i_k \neq k$ se permutan entre sí las filas k e i_k, lo cual equivale en escritura matricial a multiplicar la matriz $(A_k \,|\, B_k)$ por la izquierda por la matriz $P_{\sigma_{i_k,k}} \in \mathcal{M}_n(\mathbb{K})$ de permutación elemental.

En cualquier caso notamos

$$P_k = \left\{ \begin{array}{ll} I & \text{si } a^k_{k,k} \text{ es el } k\text{-ésimo pívot,} \quad \text{en este caso } \det(P_k) = 1, \\ P_{\sigma_{i_k,k}} & \text{si } a^k_{i_k,k},\ i_k \neq k, \text{ es el } k\text{-ésimo pívot,} \quad \text{en este caso } \det(P_k) = -1. \end{array} \right.$$

Sea $P_k\,(A_k \,|\, B_k) = (\alpha^k_{i,j} \,|\, \beta^k_i)$, donde $\alpha^k_{k,k} \neq 0$, es decir,

$$P_k\,(A_k \,|\, B_k) = \left(\begin{array}{ccc|ccccc} \alpha^k_{1,1} & \cdots & \alpha^k_{1,k-1} & \alpha^k_{1,k} & \alpha^k_{1,k+1} & \cdots & \alpha^k_{1,n} & \beta^k_1 \\ & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & \alpha^k_{k-1,k-1} & \alpha^k_{k-1,k} & \alpha^k_{k-1,k+1} & \cdots & \alpha^k_{k-1,n} & \beta^k_{k-1} \\ \hline & & & \boxed{\alpha^k_{k,k}} & \alpha^k_{k,k+1} & \cdots & \alpha^k_{k,n} & \beta^k_k \\ & & & \alpha^k_{k+1,k} & \alpha^k_{k+1,k+1} & \cdots & \alpha^k_{k+1n} & \beta^k_{k+1} \\ & & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & & \alpha^k_{n,k} & \alpha^k_{n,k+1} & \cdots & \alpha^k_{n,n} & \beta^k_n \end{array}\right)$$

$$= \left(\begin{array}{ccc|ccccc} \alpha^1_{1,1} & \cdots & \alpha^1_{1,k-1} & \alpha^1_{1,k} & \alpha^1_{1,k+1} & \cdots & \alpha^1_{1,n} & \beta^1_1 \\ & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & \alpha^{k-1}_{k-1,k-1} & \alpha^{k-1}_{k-1,k} & \alpha^{k-1}_{k-1,k+1} & \cdots & \alpha^{k-1}_{k-1,n} & \beta^{k-1}_{k-1} \\ \hline & & & \boxed{\alpha^k_{k,k}} & \alpha^k_{k,k+1} & \cdots & \alpha^k_{k,n} & \beta^k_k \\ & & & \alpha^k_{k+1,k} & \alpha^k_{k+1,k+1} & \cdots & \alpha^k_{k+1n} & \beta^k_{k+1} \\ & & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & & \alpha^k_{n,k} & \alpha^k_{n,k+1} & \cdots & \alpha^k_{n,n} & \beta^k_n \end{array}\right).$$

Mediante combinaciones lineales de la k-ésima fila, se anulan todos los elementos de la k-ésima columna situados por debajo de la diagonal principal, dejando las k primeras filas inalteradas; esto es equivalente a multiplicar por la izquierda la matriz $P_k\,(A_k \,|\, B_k)$ por la matriz

$$E_k = \left(\begin{array}{ccc|c|c} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ \hline & & & 1 & \\ \hline & & & -\dfrac{\alpha^k_{k+1,k}}{\alpha^k_{k,k}} & 1 \\ & & & \vdots & & \ddots \\ & & & -\dfrac{\alpha^k_{n,k}}{\alpha^k_{k,k}} & & & 1 \end{array}\right) \in \mathcal{M}_n(\mathbb{K})\,.$$

Sea $(A_{k+1} \,|\, B_{k+1}) = (a^{k+1}_{i,j} \,|\, b^{k+1}_i) = E_k\,P_k\,(A_k \,|\, B_k) = \cdots = E_k\,P_k\,\cdots\,E_1\,P_1\,(A \,|\, B)$, es decir,

$$(A_{k+1} \,|\, B_{k+1}) \;=\; \left(\begin{array}{ccc|c|ccc|c} a^{k+1}_{1,1} & \cdots & a^{k+1}_{1,k-1} & a^{k+1}_{1,k} & a^{k+1}_{1,k+1} & \cdots & a^{k+1}_{1,n} & b^{k+1}_1 \\ & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & a^{k+1}_{k-1,k-1} & a^{k+1}_{k-1,k} & a^{k+1}_{k-1,k+1} & \cdots & a^{k+1}_{k-1,n} & b^{k+1}_{k-1} \\ \hline & & & a^{k+1}_{k,k} & a^{k+1}_{k,k+1} & \cdots & a^{k+1}_{k,n} & b^{k+1}_k \\ \hline & & & & a^{k+1}_{k+1,k+1} & \cdots & a^{k+1}_{k+1,n} & b^{k+1}_{k+1} \\ & & & & \vdots & \ddots & \vdots & \vdots \\ & & & & a^{k+1}_{n,k+1} & \cdots & a^{k+1}_{n,n} & b^{k+1}_n \end{array}\right)$$

$$= \left(\begin{array}{ccc|c|ccc|c} \alpha^1_{1,1} & \cdots & \alpha^1_{1,k-1} & \alpha^1_{1,k} & \alpha^1_{1,k+1} & \cdots & \alpha^1_{1,n} & \beta^1_1 \\ & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & \alpha^{k-1}_{k-1,k-1} & \alpha^{k-1}_{k-1,k} & \alpha^{k-1}_{k-1,k+1} & \cdots & \alpha^{k-1}_{k-1,n} & \beta^{k-1}_{k-1} \\ \hline & & & \alpha^k_{k,k} & \alpha^k_{k,k+1} & \cdots & \alpha^k_{k,n} & \beta^k_k \\ \hline & & & & a^{k+1}_{k+1,k+1} & \cdots & a^{k+1}_{k+1,n} & b^{k+1}_{k+1} \\ & & & & \vdots & \ddots & \vdots & \vdots \\ & & & & a^{k+1}_{n,k+1} & \cdots & a^{k+1}_{n,n} & b^{k+1}_n \end{array}\right).$$

Además, como $\det(A_{k+1}) = \det(E_k)\,\det(P_k)\,\det(A_k) = \pm\det(A_k) = \pm\det(A)$ la matriz A_{k+1} sigue siendo regular.

Finalización del proceso de eliminación.

Tras $(n-1)$ iteraciones del proceso de eliminación, la matriz

$$A_n = E_{n-1}\, P_{n-1}\, A_{n-1} = \cdots = E_{n-1}\, P_{n-1} \cdots E_1\, P_1\, A$$

es evidentemente triangular superior; tomando entonces

$$M = E_{n-1}\, P_{n-1} \cdots E_1\, P_1\,,$$

con lo que

$$M\, A = A_n\,,$$

y sabiendo que

$$\det(M) = \begin{cases} +1 & \text{si se han realizado un número par de cambios de filas,} \\ -1 & \text{si se han realizado un número impar de cambios de filas,} \end{cases}$$

se llega, como era nuestro objetivo, a que hemos encontrado una matriz M inversible tal que $M\, A$ es triangular superior.

Se verifica que

$$\det(A) = \pm\, \alpha_{1,1}^1\, \alpha_{2,2}^2 \cdots \alpha_{n-1,n-1}^{n-1}\, a_{n,n}^n\,,$$

es decir, el determinante de la matriz A es, módulo el signo, el producto de los pivotes de las diferentes iteraciones (aceptando que $a_{n,n}^n$ es el pívot de la n-ésima iteración que no se llega a realizar).

Todas las permutaciones de filas P_k, $k = 1, 2, \ldots, n-1$, y todas las combinaciones lineales de filas E_k, $k = 1, 2, \ldots, n-1$, que se han realizado con los elementos de la matriz A para obtener $M\, A$, se han realizado simultáneamente con los elementos del segundo miembro B para obtener también $M\, B = B_n$.

Teorema 8 *Dada $A \in \mathcal{M}_n(\mathbb{K})$, existe $M \in \mathcal{M}_n(\mathbb{K})$ inversible tal que $M\, A$ es triangular superior.*

Elección del pívot.

Nos situamos en la k-ésima iteración, $k = 1, 2, \ldots, n-1$, del proceso de eliminación; para elegir el pívot correspondiente, distintas estrategias son posibles:

- si $a^k_{k,k} \neq 0$ podemos elegirlo como pívot, con lo cual nos evitamos la permutación de filas;
- no obstante si $|a^k_{k,k}| \approx 0$, aún siendo distinto de 0, se pueden producir grandes errores de redondeo al dividir por un número muy pequeño;
- se llama estrategia de primer pívot no nulo a aquella que consiste en tomar como pívot el primer elemento no nulo (o mayor en valor absoluto que un cierto umbral de precisión) de la columna $a^k_{i,k}$, $i = k, k+1, \ldots, n$;
- se llama estrategia de pívot parcial a aquella que consiste en tomar como pívot el elemento $a^k_{i_k,k}$ tal que
$$|a^k_{i_k,k}| = \max_{i=k,k+1,\ldots,n} |a^k_{i,k}| \, ;$$
- se llama estrategia de pívot total a aquella que consiste en tomar como pívot el elemento $a^k_{i_k,j_k}$ tal que
$$|a^k_{i_k,j_k}| = \max_{i=k,k+1,\ldots,n \, ; \, j=k,k+1,\ldots,n} |a^k_{i,j}| \, ;$$
no obstante, en este caso puede haber también una permutación de columnas (si $j_k \neq k$) que se traduce en multiplicar la matriz A por la derecha por una matriz de permutación elemental, lo cual equivale a una permutación en el orden de las incógnitas (permutación que debe ser deshecha tras resolver el sistema lineal resultante).

En cualquier caso, para una matriz A arbitraria, si queremos evitar las divisiones por cero, así como los errores de redondeo, es indispensable la utilización de alguna estrategia de pívot. Hay que señalar, sin embargo, que existen matrices (las simétricas y definidas positivas, por ejemplo) para las que la utilización de tales técnicas no es indispensable.

Número de operaciones elementales necesarias.

En el proceso de eliminación, para pasar de A_k a A_{k+1}, $k = 1, 2, \ldots, n-1$, hay que realizar:

$$\begin{array}{rl} (n-k)^2 & \text{sumas,} \\ (n-k)^2 & \text{multiplicaciones,} \\ n-k & \text{divisiones;} \end{array}$$

por tanto, en **total** hay:

$$\begin{array}{rcll} (n-1)^2 + (n-2)^2 + \cdots + 1^2 & = & \frac{1}{6}\,(2\,n^3 - 3\,n^2 + n) & \text{sumas,} \\ (n-1)^2 + (n-2)^2 + \cdots + 1^2 & = & \frac{1}{6}\,(2\,n^3 - 3\,n^2 + n) & \text{multiplicaciones,} \\ (n-1) + (n-2) + \cdots + 1 & = & \frac{1}{2}\,(n^2 - n) & \text{divisiones.} \end{array}$$

En el segundo miembro, para pasar de B_k a B_{k+1}, $k = 1, 2, \ldots, n-1$, hay que realizar:

$$\begin{array}{rl} n-k & \text{sumas,} \\ n-k & \text{multiplicaciones;} \end{array}$$

por tanto, en **total** hay:

$$\begin{array}{rcll} (n-1) + (n-2) + \cdots + 1 & = & \frac{1}{2}\,(n^2 - n) & \text{sumas,} \\ (n-1) + (n-2) + \cdots + 1 & = & \frac{1}{2}\,(n^2 - n) & \text{multiplicaciones.} \end{array}$$

Si a las cantidades anteriores les sumamos las necesarias para el proceso de remonte, se tiene que el coste total del método de Gauss es:

$$\begin{array}{rcll} \frac{1}{6}\,(2\,n^3 + 3\,n^2 - 5\,n) & \approx & \frac{1}{3}\,n^3 & \text{sumas,} \\ \frac{1}{6}\,(2\,n^3 + 3\,n^2 - 5\,n) & \approx & \frac{1}{3}\,n^3 & \text{multiplicaciones,} \\ \frac{1}{2}\,(n^2 + n) & \approx & \frac{1}{2}\,n^2 & \text{divisiones.} \end{array}$$

En la evaluación del coste total del método de Gauss no se han tenido en cuenta las posibles permutaciones de filas (y columnas) necesarias en la estrategia de pívot utilizada y, en particular, el tiempo utilizado en la búsqueda del pívot óptimo.

4. Método de Gauss-Jordan

El **método de Gauss-Jordan** es un método directo de resolución del Problema 1 en el que se pueden diferenciar tres etapas:

1. determinación de una matriz $M \in \mathcal{M}_n(\mathbb{K})$ inversible tal que $M\,A$ sea diagonal;
2. cálculo simultáneo del vector $M\,B$;
3. resolución (inmediata) del sistema lineal

$$M\,A\,X = M\,B\,.$$

Al igual que con el método de Gauss, la matriz M no se llega a calcular nunca explícitamente para después multiplicarla por A y B; en la práctica se utiliza un proceso iterativo finito de eliminación análogo al ya visto para el método de Gauss con las únicas variantes que ahora se indican (el objetivo de cada iteración es, a partir del pívot elegido, anular todos los restantes elementos de columna en cuestión y no sólo los situados por debajo de la diagonal principal).

Descripción de la primera iteración del proceso de eliminación.

Esta iteración es completamente análoga a la primera iteración del método de Gauss (en este caso no hay elementos por encima de la diagonal principal en la primera columna).

Descripción de la k-ésima iteración del proceso de eliminación.

Sea $(A_k \,|\, B_k) = (a_{i,j}^k \,|\, b_i^k)$ el resultado de la $(k-1)$-ésima iteración del proceso de eliminación, por lo que $(A_k \,|\, B_k) = E_{k-1}\,P_{k-1}\,(A_{k-1} \,|\, B_{k-1}) = \cdots = E_{k-1}\,P_{k-1}\cdots E_1\,P_1\,(A \,|\, B)$, es decir,

$$(A_k \,|\, B_k) = \left(\begin{array}{ccc|ccc|c} a_{1,1}^k & & & a_{1,k}^k & a_{1,k+1}^k & \cdots & a_{1,n}^k & b_1^k \\ & \ddots & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & a_{k-1,k-1}^k & a_{k-1,k}^k & a_{k-1,k+1}^k & \cdots & a_{k-1,n}^k & b_{k-1}^k \\ \hline & & & a_{k,k}^k & a_{k,k+1}^k & \cdots & a_{k,n}^k & b_k^k \\ & & & a_{k+1,k}^k & a_{k+1,k+1}^k & \cdots & a_{k+1,n}^k & b_{k+1}^k \\ & & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & & a_{n,k}^k & a_{n,k+1}^k & \cdots & a_{n,n}^k & b_n^k \end{array}\right)$$

$$= \left(\begin{array}{ccc|ccc|c} \alpha_{1,1}^1 & & & a_{1,k}^k & a_{1,k+1}^k & \cdots & a_{1,n}^k & b_1^k \\ & \ddots & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & \alpha_{k-1,k-1}^{k-1} & a_{k-1,k}^k & a_{k-1,k+1}^k & \cdots & a_{k-1,n}^k & b_{k-1}^k \\ \hline & & & a_{k,k}^k & a_{k,k+1}^k & \cdots & a_{k,n}^k & b_k^k \\ & & & a_{k+1,k}^k & a_{k+1,k+1}^k & \cdots & a_{k+1,n}^k & b_{k+1}^k \\ & & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & & a_{n,k}^k & a_{n,k+1}^k & \cdots & a_{n,n}^k & b_n^k \end{array}\right).$$

Como A_k es regular, al menos uno de los elementos $a^k_{i,k}$, $i = k, k+1, \ldots, n$, debe ser no nulo; sea $a^k_{i_k,k}$ uno de tales elementos (estrategia para la elección) que se llama k-ésimo **pívot** de la eliminación.

Si $i_k \neq k$ se permutan entre sí las filas k e i_k, lo cual equivale en escritura matricial a multiplicar la matriz A_k por la izquierda por la matriz $P_{\sigma_{i_k,k}} \in \mathcal{M}_n(\mathbb{K})$ de permutación elemental.

En cualquier caso notamos

$$P_k = \begin{cases} I & \text{si } a^k_{k,k} \text{ es el } k\text{-ésimo pívot,} \quad \text{en este caso } \det(P_k) = 1, \\ P_{\sigma_{i_k,k}} & \text{si } a^k_{i_k,k},\ i_k \neq k, \text{ es el } k\text{-ésimo pívot,} \quad \text{en este caso } \det(P_k) = -1. \end{cases}$$

Sea $P_k\,(A_k \,|\, B_k) = (\alpha^k_{i,j} \,|\, \beta^k_i)$, donde $\alpha^k_{k,k} \neq 0$, es decir,

$$P_k\,(A_k \,|\, B_k) = \left(\begin{array}{ccc|ccccc} \alpha^k_{1,1} & & & \alpha^k_{1,k} & \alpha^k_{1,k+1} & \cdots & \alpha^k_{1,n} & \beta^k_1 \\ & \ddots & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & \alpha^k_{k-1,k-1} & \alpha^k_{k-1,k} & \alpha^k_{k-1,k+1} & \cdots & \alpha^k_{k-1,n} & \beta^k_{k-1} \\ \hline & & & \boxed{\alpha^k_{k,k}} & \alpha^k_{k,k+1} & \cdots & \alpha^k_{k,n} & \beta^k_k \\ & & & \alpha^k_{k+1,k} & \alpha^k_{k+1,k+1} & \cdots & \alpha^k_{k+1,n} & \beta^k_{k+1} \\ & & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & & \alpha^k_{n,k} & \alpha^k_{n,k+1} & \cdots & \alpha^k_{n,n} & \beta^k_n \end{array}\right)$$

$$= \left(\begin{array}{ccc|ccccc} \alpha^1_{1,1} & & & a^k_{1,k} & a^k_{1,k+1} & \cdots & a^k_{1,n} & b^k_1 \\ & \ddots & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & \alpha^{k-1}_{k-1,k-1} & a^k_{k-1,k} & a^k_{k-1,k+1} & \cdots & a^k_{k-1,n} & b^k_{k-1} \\ \hline & & & \boxed{\alpha^k_{k,k}} & \alpha^k_{k,k+1} & \cdots & \alpha^k_{k,n} & \beta^k_k \\ & & & \alpha^k_{k+1,k} & \alpha^k_{k+1,k+1} & \cdots & \alpha^k_{k+1,n} & \beta^k_{k+1} \\ & & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & & \alpha^k_{n,k} & \alpha^k_{n,k+1} & \cdots & \alpha^k_{n,n} & \beta^k_n \end{array}\right).$$

Mediante combinaciones lineales de la k-ésima fila, se anulan todos los elementos de la k-ésima columna excepto el elemento que pertenece a la diagonal principal; esto es equivalente a multiplicar por la izquierda la matriz $P_k\,(A_k \,|\, B_k)$ por la matriz

$$E_k = \left(\begin{array}{ccc|c|c} 1 & & & -\frac{\alpha_{1,k}^k}{\alpha_{k,k}^k} & \\ & \ddots & & \vdots & \\ & & 1 & -\frac{\alpha_{k-1,k}^k}{\alpha_{k,k}^k} & \\ \hline & & & 1 & \\ \hline & & & -\frac{\alpha_{k+1,k}^k}{\alpha_{k,k}^k} & 1 \\ & & & \vdots & & \ddots \\ & & & -\frac{\alpha_{n,k}^k}{\alpha_{k,k}^k} & & & 1 \end{array}\right) \in \mathcal{M}_n(\mathbb{K})\,.$$

Sea $(A_{k+1} \,|\, B_{k+1}) = (a_{i,j}^{k+1} \,|\, b_i^{k+1}) = E_k\,P_k\,(A_k \,|\, B_k) = \cdots = E_k\,P_k \cdots E_1\,P_1\,(A \,|\, B)$, es decir,

$$(A_{k+1} \,|\, B_{k+1}) = \left(\begin{array}{ccc|c|ccc|c} a_{1,1}^{k+1} & & & & a_{1,k+1}^{k+1} & \cdots & a_{1,n}^{k+1} & b_1^{k+1} \\ & \ddots & & & \vdots & \ddots & \vdots & \vdots \\ & & a_{k-1,k-1}^{k+1} & & a_{k-1,k+1}^{k+1} & \cdots & a_{k-1,n}^{k+1} & b_{k-1}^{k+1} \\ \hline & & & a_{k,k}^{k+1} & a_{k,k+1}^{k+1} & \cdots & a_{k,n}^{k+1} & b_k^{k+1} \\ \hline & & & & a_{k+1,k+1}^{k+1} & \cdots & a_{k+1,n}^{k+1} & b_{k+1}^{k+1} \\ & & & & \vdots & \ddots & \vdots & \vdots \\ & & & & a_{n,k+1}^{k+1} & \cdots & a_{n,n}^{k+1} & b_n^{k+1} \end{array}\right)$$

$$= \left(\begin{array}{ccc|c|ccc|c} \alpha_{1,1}^1 & & & & a_{1,k+1}^{k+1} & \cdots & a_{1,n}^{k+1} & b_1^{k+1} \\ & \ddots & & & \vdots & \ddots & \vdots & \vdots \\ & & \alpha_{k-1,k-1}^{k-1} & & a_{k-1,k+1}^{k+1} & \cdots & a_{k-1,n}^{k+1} & b_{k-1}^{k+1} \\ \hline & & & \alpha_{k,k}^k & \alpha_{k,k+1}^k & \cdots & \alpha_{k,n}^k & \beta_k^k \\ \hline & & & & a_{k+1,k+1}^{k+1} & \cdots & a_{k+1,n}^{k+1} & b_{k+1}^{k+1} \\ & & & & \vdots & \ddots & \vdots & \vdots \\ & & & & a_{n,k+1}^{k+1} & \cdots & a_{n,n}^{k+1} & b_n^{k+1} \end{array}\right).$$

Además, como $\det(A_{k+1}) = \det(E_k)\,\det(P_k)\,\det(A_k) = \pm\det(A_k) = \pm\det(A)$, la matriz A_{k+1} sigue siendo regular.

Finalización del proceso de eliminación.

Observamos que, a diferencia con en el método de Gauss donde $(n-1)$ iteraciones son suficientes, en el método de Gauss-Jordan es necesario realizar una n-ésima iteración, pues es preciso anular también los elementos de la última columna de A_n situados por encima de la diagonal principal. Resulta claro que en esta n-ésima iteración la matriz P_n no puede sino ser la matriz identidad I.

Por tanto, tras n iteraciones del proceso de eliminación, la matriz

$$A_{n+1} = E_n\, P_n\, A_n = \cdots = E_n\, P_n\, \cdots\, E_1\, P_1\, A$$

es evidentemente diagonal; tomando entonces

$$M = E_n\, P_n\, \cdots\, E_1\, P_1$$

con lo que

$$M\, A = A_{n+1}\,,$$

y sabiendo que

$$\det(M) = \begin{cases} +1 & \text{si se han realizado un número par de cambios de filas,} \\ -1 & \text{si se han realizado un número impar de cambios de filas,} \end{cases}$$

se llega, como era nuestro objetivo, a que hemos encontrado una matriz M inversible tal que $M\, A$ es diagonal.

Se verifica que

$$\det(A) = \pm\, \alpha_{1,1}^1\, \alpha_{2,2}^2\, \cdots\, \alpha_{n,n}^n\,,$$

es decir, el determinante de la matriz A es, módulo el signo, el producto de los pivotes de las diferentes iteraciones.

Todas las permutaciones de filas P_k, $k = 1, 2, \ldots, n$, y todas las combinaciones lineales de filas E_k, $k = 1, 2, \ldots, n$, que se han realizado con los elementos de la matriz A para obtener $M\, A$, se han realizado simultáneamente con los elementos del segundo miembro B para obtener también $M\, B = B_{n+1}$.

Teorema 9 *Dada $A \in \mathcal{M}_n(\mathbb{K})$ inversible, existe $M \in \mathcal{M}_n(\mathbb{K})$ inversible tal que $M\, A$ es diagonal.*

Observación 10 *También en el método de Gauss-Jordan es necesaria la utilización de una estrategia de pívot; todos los comentarios realizados para el método de Gauss permanecen válidos.*

Número de operaciones elementales necesarias.

En el proceso de eliminación, para pasar de A_k a A_{k+1}, $k = 1, 2, \ldots, n$, hay que realizar:

$$\begin{array}{cl} (n-k)\,(n-1) & \text{sumas,} \\ (n-k)\,(n-1) & \text{multiplicaciones,} \\ n-1 & \text{divisiones;} \end{array}$$

por tanto, en **total** hay:

$$\begin{array}{ccc} [(n-1)+(n-2)+\cdots+0]\,(n-1) & = & \frac{1}{2}\,(n^3-2\,n^2+n) \quad \text{sumas,} \\ [(n-1)+(n-2)+\cdots+0]\,(n-1) & = & \frac{1}{2}\,(n^3-2\,n^2+n) \quad \text{multiplicaciones,} \\ n\,(n-1) & = & n^2-n \quad \text{divisiones.} \end{array}$$

En el segundo miembro, para pasar de B_k a B_{k+1}, $k = 1, 2, \ldots, n$, hay que realizar:

$$\begin{array}{cl} n-1 & \text{sumas,} \\ n-1 & \text{multiplicaciones;} \end{array}$$

por tanto, en **total** hay:

$$\begin{array}{ccl} n\,(n-1) & = & n^2-n \quad \text{sumas,} \\ n\,(n-1) & = & n^2-n \quad \text{multiplicaciones.} \end{array}$$

Si a las cantidades anteriores les sumamos las necesarias para la resolución de un sistema diagonal (sólo n divisiones), se tiene que el coste total del método de Gauss-Jordan es:

$$\begin{array}{ccl} \frac{1}{2}\,(n^3-n) & \approx & \frac{1}{2}\,n^3 \quad \text{sumas,} \\ \frac{1}{2}\,(n^3-n) & \approx & \frac{1}{2}\,n^3 \quad \text{multiplicaciones,} \\ n^2 & = & n^2 \quad \text{divisiones.} \end{array}$$

Tampoco aquí se ha tenido en cuenta el tiempo utilizado en realizar las posibles permutaciones de filas (y columnas) necesarias en la estrategia de pívot utilizada, ni el tiempo preciso para localizar el pívot óptimo.

5. Factorización y método $L\,U$

Definición 11 *Dada $A \in \mathcal{M}_n(\mathbb{K})$ inversible, se llama factorización $L\,U$ a la descomposición, si es posible,*

$$A = L\,U\,,$$

siendo $L \in \mathcal{M}_n(\mathbb{K})$ triangular inferior, inversible y con unos en la diagonal principal y $U \in \mathcal{M}_n(\mathbb{K})$ triangular superior e inversible.

Construcción de la factorización $L\,U$ de una matriz.

Supongamos que en el proceso de eliminación del método de Gauss no se hace ninguna permutación de filas, es decir, $P_k = I$, $k = 1, 2, \ldots, n-1$; esto implica que ningún elemento de la forma $a^k_{k,k}$, $k = 1, 2, \ldots, n-1$, debe ser nulo y, además, no nos interesamos por los posibles errores de redondeo en el caso de que alguno de dichos elementos (que son los pivotes de las iteraciones correspondientes) sea pequeño en valor absoluto.

Notamos que este caso $\alpha^k_{i,j} = a^k_{i,j}$, $k = 1, 2, \ldots, n-1$, $i, j = 1, 2, \ldots, n$, y además

$$A_n = M\,A = \begin{pmatrix} a^1_{1,1} & a^1_{1,2} & \cdots & a^1_{1,n} \\ & a^2_{2,2} & \cdots & a^2_{2,n} \\ & & \ddots & \vdots \\ & & & a^n_{n,n} \end{pmatrix}.$$

Tomamos entonces

$$U = A_n$$

que es triangular superior e inversible. Además, como en este caso

$$M = E_{n-1} \cdots E_1\,,$$

y como todas las matrices E_k, $k = 1, 2, \ldots, n-1$, son triangulares inferior, inversibles y con unos en la diagonal principal, su producto también lo es, así como su inversa que notamos mediante

$$L = M^{-1} = {E_1}^{-1} \cdots {E_{n-1}}^{-1}.$$

En consecuencia se tiene que

$$A_n = M\,A \iff L\,U = A\,.$$

Veamos una forma cómoda de construir la matriz L; basta observar que

$$E_k{}^{-1} = \left(\begin{array}{ccc|c|c} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ \hline & & & 1 & \\ \hline & & & \frac{a^k_{k+1,k}}{a^k_{k,k}} & 1 \\ & & & \vdots & & \ddots \\ & & & \frac{a^k_{n,k}}{a^k_{k,k}} & & & 1 \end{array}\right), \quad k = 1, 2, \ldots, n-1,$$

con lo que

$$L = E_1{}^{-1} \cdots E_{n-1}{}^{-1} = \begin{pmatrix} 1 & & & \\ \frac{a^1_{2,1}}{a^1_{1,1}} & 1 & & \\ \frac{a^1_{3,1}}{a^1_{1,1}} & \frac{a^2_{3,2}}{a^2_{2,2}} & 1 & \\ \vdots & \vdots & \vdots & \ddots \\ \frac{a^1_{n,1}}{a^1_{1,1}} & \frac{a^2_{n,2}}{a^2_{2,2}} & \frac{a^3_{n,3}}{a^3_{3,3}} & \cdots & 1 \end{pmatrix}.$$

Teorema 12 *Sea $A \in \mathcal{M}_n(\mathbb{K})$ inversible y tal que todas las submatrices principales son también inversibles. Entonces la factorización LU existe y es única.*

La factorización LU de una matriz está asociada de forma natural a un método directo de resolución del Problema 1, llamado método LU, en el que se pueden diferenciar dos etapas:

1. determinación de la factorización LU de la matriz A;
2. resolución mediante un proceso de descenso seguido de uno de remonte del sistema lineal, ya que
$$AX = B \iff LY = B \quad \text{y} \quad UX = Y\,.$$

El coste del método LU es el mismo que el de Gauss, si no se tienen en cuenta las posibles permutaciones de filas (y columnas) de este último ni el tiempo invertido en la búsqueda del pívot; indicamos que en el proceso de descenso no es preciso hacer las n divisiones, ya que los elementos diagonales de L son unos.

6. Factorización y método de Cholesky

Definición 13 *Dada $A \in \mathcal{M}_n(\mathbb{R})$ simétrica e inversible, se llama factorización de Cholesky a la descomposición, si es posible,*

$$A = C\,C^t\,,$$

siendo $C \in \mathcal{M}_n(\mathbb{R})$ triangular inferior e inversible.

Teorema 14 *Si $A \in \mathcal{M}_n(\mathbb{R})$ es simétrica y definida positiva, entonces la factorización de Cholesky existe. Además, si se impone que todos los elementos diagonales de la matriz C sean positivos, entonces la factorización es única.*

Corolario 15 *Dada $A \in \mathcal{M}_n(\mathbb{R})$ simétrica, existe la factorización de Cholesky si y sólo si es definida positiva.*

Sea $A \in \mathcal{M}_n(\mathbb{R})$ tal que admite factorización de Cholesky, siendo $C = (c_{i,j}) \in \mathcal{M}_n(\mathbb{R})$; se verifica entonces que

$$\det(A) = \prod_{i=1}^{n} {c_{i,i}}^2\,.$$

Observación 16 *Si bien siempre que se habla de la factorización de Cholesky se hace referencia a matrices reales, todos los resultados (incluida la unicidad en el caso de que los elementos diagonales de C sean positivos) siguen siendo válidos en el caso complejo, cambiando simétrica por hermítica y transpuesta por conjugada.*

La factorización de Cholesky de una matriz está asociada de forma natural a un método directo de resolución del Problema 1, llamado método de Cholesky, en el que se pueden diferenciar dos etapas:

1. determinación de la factorización de Cholesky de la matriz A;
2. resolución mediante un proceso de descenso seguido de uno de remonte del sistema lineal, ya que

$$A\,X = B \iff C\,Y = B \quad \text{y} \quad C^t\,X = Y\,.$$

Construcción de la factorización de Cholesky de una matriz.

Sea $A = (a_{i,j}) \in \mathcal{M}_n(\mathbb{R})$ simétrica y definida positiva, de forma que la factorización de Cholesky existe y es única (supuestos positivos los elementos diagonales de la matriz C). Sea entonces

$$C = \begin{pmatrix} c_{1,1} & & & \\ c_{2,1} & c_{2,2} & & \\ \vdots & \vdots & \ddots & \\ c_{n,1} & c_{n,2} & \cdots & c_{n,n} \end{pmatrix} .$$

Igualando A con el producto $C\,C^t$ y recordando que C es triangular inferior se obtiene

$$a_{i,j} = \sum_{k=1}^{n} c_{i,k}\, c_{j,k} = \sum_{k=1}^{\min\{i,j\}} c_{i,k}\, c_{j,k}\,, \quad i,j = 1,2,\ldots,n\,.$$

Puesto que A es simétrica, basta con que las relaciones anteriores sean verificadas para $i \leq j$, por ejemplo, con lo que

$$a_{i,j} = \sum_{k=1}^{i} c_{i,k}\, c_{j,k}\,, \quad i,j = 1,2,\ldots,n\,, \quad i \leq j\,.$$

Así, haciendo $i = 1$ se obtiene:

j	$a_{1,j}$	$c_{j,1}$
1	$a_{1,1} = {c_{1,1}}^2$	$c_{1,1} = +\sqrt{a_{1,1}}$
2	$a_{1,2} = c_{1,1}\, c_{2,1}$	$c_{2,1} = \dfrac{a_{1,2}}{c_{1,1}}$
$\vdots$	$\vdots$	$\vdots$
j	$a_{1,j} = c_{1,1}\, c_{j,1}$	$c_{j,1} = \dfrac{a_{1,j}}{c_{1,1}}$
$\vdots$	$\vdots$	$\vdots$
n	$a_{1,n} = c_{1,1}\, c_{n,1}$	$c_{n,1} = \dfrac{a_{1,n}}{c_{1,1}}$

Supuesto conocidas las $(i-1)$, $i = 2, 3, \ldots, n$, primeras columnas de C, para calcular la i-ésima columnas se tiene:

j	$a_{i,j}$	$c_{j,i}$
i	$a_{i,i} = \sum_{k=1}^{i} {c_{i,k}}^2$	$c_{i,i} = +\sqrt{a_{i,i} - \sum_{k=1}^{i-1} {c_{i,k}}^2}$
$i+1$	$a_{i,i+1} = \sum_{k=1}^{i} c_{i,k}\, c_{i+1,k}$	$c_{i+1,i} = \dfrac{a_{i,i+1} - \sum_{k=1}^{i-1} c_{i,k}\, c_{i+1,k}}{c_{i,i}}$
$\vdots$	$\vdots$	$\vdots$
j	$a_{i,j} = \sum_{k=1}^{i} c_{i,k}\, c_{j,k}$	$c_{j,i} = \dfrac{a_{i,j} - \sum_{k=1}^{i-1} c_{i,k}\, c_{j,k}}{c_{i,i}}$
$\vdots$	$\vdots$	$\vdots$
n	$a_{i,n} = \sum_{k=1}^{i} c_{i,k}\, c_{n,k}$	$c_{n,i} = \dfrac{a_{i,n} - \sum_{k=1}^{i-1} c_{i,k}\, c_{n,k}}{c_{i,i}}$

Como puede verse, la importancia de las fórmulas anteriores reside en que haciendo $i = 1, 2, \ldots, n$, es posible ir calculando las respectivas columnas de C, las cuales dependen en cada momento de la correspondiente columna de A (o fila, puesto que es simétrica) y de las columnas anteriores de C (ya calculadas).

Número de operaciones elementales necesarias.

Para calcular la i-ésima, $i = 1, 2, \ldots, n$, columna de la matriz C, hay que realizar:

$$\begin{array}{cl} (i-1)\,(n-i+1) & \text{sumas,} \\ (i-1)\,(n-i+1) & \text{multiplicaciones,} \\ n-i & \text{divisiones,} \\ 1 & \text{raíz cuadrada;} \end{array}$$

por lo que el coste total de la factorización de Cholesky es:

$$\begin{array}{rcll} 0\,n+1\,(n-1)+\cdots+(n-1)\,1 & = & \frac{1}{6}\,(n^3-n) & \text{sumas,} \\ 0\,n+1\,(n-1)+\cdots+(n-1)\,1 & = & \frac{1}{6}\,(n^3-n) & \text{multiplicaciones,} \\ (n-1)+(n-2)+\cdots+0 & = & \frac{1}{2}\,(n^2-n) & \text{divisiones,} \\ n\,1 & = & n & \text{raíces cuadradas.} \end{array}$$

A las cantidades anteriores hay que sumar el coste de un proceso de descenso y otro de remontada, por lo que el coste total del método de Cholesky es:

$$\begin{array}{rcll} \frac{1}{6}\,(n^3+6\,n^2-7\,n) & \approx & \frac{1}{6}\,n^3 & \text{sumas,} \\ \frac{1}{6}\,(n^3+6\,n^2-7\,n) & \approx & \frac{1}{6}\,n^3 & \text{multiplicaciones,} \\ \frac{1}{2}\,(n^2+3\,n) & \approx & \frac{1}{2}\,n^2 & \text{divisiones,} \\ n & = & n & \text{raíces cuadradas.} \end{array}$$

Notamos que el número de operaciones necesarias para resolver un sistema lineal por el método de Cholesky es inferior al necesario para el método de Gauss; esto hace que el método de Cholesky sea el utilizado preferentemente para resolver sistemas de matriz simétrica y definida positiva.

7. Factorización $Q\,R$ y método de Householder

Definición 17 *Se llama matriz de Householder a toda matriz de la forma*

$$\begin{aligned} &\text{si} \quad \mathbb{K}=\mathbb{R}: \quad H(Z) \;=\; I-\frac{2}{Z^t\,Z}\,Z\,Z^t\in\mathcal{M}_n(\mathbb{R})\,, \quad Z\in\mathbb{R}^n-\{0\}\,,\\ &\text{si} \quad \mathbb{K}=\mathbb{C}: \quad H(Z) \;=\; I-\frac{2}{Z^*\,Z}\,Z\,Z^*\in\mathcal{M}_n(\mathbb{C})\,, \quad Z\in\mathbb{C}^n-\{0\}\,; \end{aligned}$$

convenimos que la matriz I es también una matriz de Householder.

Proposición 18 *Toda matriz de Householder es simétrica-hermítica, ortogonal-unitaria y de determinante -1 (excepto la matriz identidad cuyo determinante es 1). Además, si $\alpha\in\mathbb{K}-\{0\}$, entonces $H(\alpha\,Z)=H(Z)$.*

Teorema 19 *Sea $X=(x_i)\in\mathbb{K}^n$ tal que $\displaystyle\sum_{i=2}^{n}|x_i|>0$; existen entonces dos matrices de Householder $H(Z)$, $H(Z')$, $Z,Z'\in\mathbb{K}^n-\{0\}$, tales que las $(n-1)$ últimas componentes de los vectores $H(Z)\,X$ y $H(Z')\,X$, son nulas. Más concretamente, si $x_1=|x_1|\,e^{\varphi_{x_1}\,i}$, entonces*

$$H(X\pm\|X\|_2\,e^{\varphi_{x_1}\,i}\,E_1)\,X=\mp\|X\|_2\,e^{\varphi_{x_1}\,i}\,E_1\,,$$

donde E_1 denota el primer vector de la base canónica de $\mathbb{K}^n$.

Corolario 20 *Si $\displaystyle\sum_{i=2}^{n}|x_i|=0$, con $x_1\neq 0$, el resultado anterior sigue siendo cierto, pero en este caso las dos matrices de Householder son $H(X)$ e I.*

A nivel práctico, se prefiere trabajar con la expresión $H(X+\|X\|_2\,e^{\varphi_{x_1}\,i}\,E_1)$ en vez de con $H(X-\|X\|_2\,e^{\varphi_{x_1}\,i}\,E_1)$, ya que esta última puede presentar en el denominador un problema de cancelación, así como una división por un nímero muy pequeño.

Si $\mathbb{K}=\mathbb{R}$, entonces $\varphi_{x_1}=0,\pi$, por lo que $e^{\varphi_{x_1}\,i}=\pm 1$ y la expresión anterior se convierte en

$$H(X\pm\|X\|_2\,E_1)\,X=\mp\|X\|_2\,E_1\,;$$

observamos que en cualquier caso siempre hay una opción en la cual el vector resultante tiene primera componente positiva.

Definición 21 *Dada $A \in \mathcal{M}_n(\mathbb{K})$ inversible, se llama factorización $Q\,R$ a la descomposición, si es posible,*

$$A = Q\,R\,,$$

donde $Q \in \mathcal{M}_n(\mathbb{K})$ es ortogonal-unitaria y $R \in \mathcal{M}_n(\mathbb{K})$ es triangular superior e inversible.

Sea A una matriz que admite una factorización $Q\,R$. Más importante que la matriz Q es su inversa $H = Q^{-1}$ (que es su transpuesta-conjugada), ya que

$$A = Q\,R \iff H\,A = R\,,$$

de donde

$$A\,X = B \iff H\,A\,X = H\,B \iff R\,X = H\,B\,.$$

La matriz H, que evidentemente también es ortogonal-unitaria, se obtiene como un producto de matrices de Householder; en la práctica se calcula directamente la matriz $H\,A$ y el vector $H\,B$ mediante el proceso iterativo finito que se describe a continuación.

La factorización $Q\,R$ de una matriz está asociada de forma natural a un método directo de resolución del Problema 1, llamado método de Householder, en el que se pueden diferenciar tres etapas:

1. determinación de una matriz $H \in \mathcal{M}_n(\mathbb{K})$ ortogonal-unitaria tal que $H\,A$ es triangular superior;
2. cálculo simultáneo del vector $H\,B$;
3. resolución mediante un proceso de remonte del sistema lineal

$$H\,A\,X = H\,B\,.$$

Observación 22 *Según el Teorema visto al hablar del condicionamiento, en el método de Householder se verifica que el sistema lineal final tiene el mismo condicionamiento que el sistema inicial, lo cual es una ventaja desde el punto de vista de la estabilidad numérica del problema.*

Descripción de la primera iteración.

Sea $A_1 = (a^1_{i,j}) = A$, es decir,

$$A_1 = \begin{pmatrix} a^1_{1,1} & a^1_{1,2} & \cdots & a^1_{1,n} \\ a^1_{2,1} & a^1_{2,2} & \cdots & a^1_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a^1_{n,1} & a^1_{n,2} & \cdots & a^1_{n,n} \end{pmatrix} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix}.$$

Sea $C_1 \in \mathbb{K}^n - \{0\}$ la primera columna de A_1; sabemos que existe $H_1 \in \mathcal{M}_n(\mathbb{K})$ matriz de Householder tal que el vector $H_1\, C_1$ tiene solamente la primera componente no nula.

Sea $A_2 = (a^2_{i,j}) = H_1\, A_1 = H_1\, A$; en consecuencia

$$A_2 = \left(\begin{array}{c|ccc} a^2_{1,1} & a^2_{1,2} & \cdots & a^2_{1,n} \\ \hline & a^2_{2,2} & \cdots & a^2_{2,n} \\ & \vdots & \ddots & \vdots \\ & a^2_{n,2} & \cdots & a^2_{n,n} \end{array}\right)$$

tiene nulos todos los elementos de la primera columna situados por debajo de la diagonal principal.

Además, como $\det(A_2) = \det(H_1)\, \det(A_1) = \pm \det(A_1) = \pm \det(A)$, la matriz A_2 es también inversible.

Descripción de la k-ésima iteración.

Sea $A_k = (a_{i,j}^k)$ el resultado de la $(k-1)$-ésima iteración, por lo que $A_k = H_{k-1}A_{k-1} = \cdots = H_{k-1} \cdots H_1 A$, es decir,

$$A_k = \left(\begin{array}{ccc|c|ccc} a_{1,1}^k & \cdots & a_{1,k-1}^k & a_{1,k}^k & a_{1,k+1}^k & \cdots & a_{1,n}^k \\ & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ & & a_{k-1,k-1}^k & a_{k-1,k}^k & a_{k-1,k+1}^k & \cdots & a_{k-1,n}^k \\ \hline & & & a_{k,k}^k & a_{k,k+1}^k & \cdots & a_{k,n}^k \\ \hline & & & a_{k+1,k}^k & a_{k+1,k+1}^k & \cdots & a_{k+1,n}^k \\ & & & \vdots & \vdots & \ddots & \vdots \\ & & & a_{n,k}^k & a_{n,k+1}^k & \cdots & a_{n,n}^k \end{array}\right)$$

$$= \left(\begin{array}{ccc|c|ccc} a_{1,1}^2 & \cdots & a_{1,k-1}^2 & a_{1,k}^2 & a_{1,k+1}^2 & \cdots & a_{1,n}^2 \\ & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ & & a_{k-1,k-1}^k & a_{k-1,k}^k & a_{k-1,k+1}^k & \cdots & a_{k-1,n}^k \\ \hline & & & a_{k,k}^k & a_{k,k+1}^k & \cdots & a_{k,n}^k \\ \hline & & & a_{k+1,k}^k & a_{k+1,k+1}^k & \cdots & a_{k+1,n}^k \\ & & & \vdots & \vdots & \ddots & \vdots \\ & & & a_{n,k}^k & a_{n,k+1}^k & \cdots & a_{n,n}^k \end{array}\right)$$

Sea $C_k \in \mathbb{K}^n - \{0\}$ la k-ésima columna de A_k y sea $\tilde{C}_k \in \mathbb{K}^{n-k+1} - \{0\}$ las $(n-k+1)$ últimas componentes de C_k; sabemos que existe $\tilde{H}_k \in \mathcal{M}_{n-k+1}(\mathbb{K})$ matriz de Householder tal que el vector $\tilde{H}_k \tilde{C}_k$ tiene solamente la primera componente no nula. Tomamos entonces

$$H_k = \left(\begin{array}{c|c} I & 0 \\ \hline 0 & \tilde{H}_k \end{array}\right) \in \mathcal{M}_n(\mathbb{K})\,.$$

Notamos que se verifica que si

$$\tilde{H}_k = H(\tilde{Z}_k)\,, \quad \text{con} \quad \tilde{Z}_k \in \mathbb{K}^{n-k+1} - \{0\}\,,$$

entonces

$$H_k = H(Z_k)\,, \quad \text{con} \quad Z_k = \left(\begin{array}{c} 0 \\ \hline \tilde{Z}_k \end{array}\right) \in \mathbb{K}^n - \{0\}\,.$$

Sea $A_{k+1} = (a_{ij}^{k+1}) = H_k\, A_k = \cdots = H_k \cdots H_1\, A$; en consecuencia

$$A_{k+1} = \left(\begin{array}{ccc|c|ccc} a_{1,1}^{k+1} & \cdots & a_{1,k-1}^{k+1} & a_{1,k}^{k+1} & a_{1,k+1}^{k+1} & \cdots & a_{1,n}^{k+1} \\ & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ & & a_{k-1,k-1}^{k+1} & a_{k-1,k}^{k+1} & a_{k-1,k+1}^{k+1} & \cdots & a_{k-1,n}^{k+1} \\ \hline & & & a_{k,k}^{k+1} & a_{k,k+1}^{k+1} & \cdots & a_{k,n}^{k+1} \\ \hline & & & & a_{k+1,k+1}^{k+1} & \cdots & a_{k+1,n}^{k+1} \\ & & & & \vdots & \ddots & \vdots \\ & & & & a_{n,k+1}^{k+1} & \cdots & a_{n,n}^{k+1} \end{array}\right)$$

$$= \left(\begin{array}{ccc|c|ccc} a_{1,1}^{2} & \cdots & a_{1,k-1}^{2} & a_{1,k}^{2} & a_{1,k+1}^{2} & \cdots & a_{1,n}^{2} \\ & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ & & a_{k-1,k-1}^{k} & a_{k-1,k}^{k} & a_{k-1,k+1}^{k} & \cdots & a_{k-1,n}^{k} \\ \hline & & & a_{k,k}^{k+1} & a_{k,k+1}^{k+1} & \cdots & a_{k,n}^{k+1} \\ \hline & & & & a_{k+1,k+1}^{k+1} & \cdots & a_{k+1,n}^{k+1} \\ & & & & \vdots & \ddots & \vdots \\ & & & & a_{n,k+1}^{k+1} & \cdots & a_{n,n}^{k+1} \end{array}\right)$$

tiene nulos todos los elementos de las k primeras columnas situados por debajo de la diagonal principal.

Además, como $\det(A_{k+1}) = \det(H_k)\, \det(A_k) = \pm \det(A_k) = \pm \det(A)$, la matriz A_{k+1} es también inversible.

Notamos que a diferencia con la factorización $L\,U$ (método de Gauss sin estrategia de pívot) donde la matriz A_{k+1} conserva las k primeras filas de la matriz A_k, en la factorización $Q\,R$ se conservan solamente las $k-1$ primeras filas.

Finalización del proceso.

Tras $(n-1)$ iteraciones, la matriz

$$A_n = H_{n-1}\, A_{n-1} = H_{n-1} \cdots H_1\, A$$

es evidentemente triangular superior; tomando entonces

$$H = H_{n-1} \cdots H_1\,,$$

con lo que

$$H\, A = A_n\,,$$

se llega, como era nuestro objetivo, a que hemos encontrado una matriz H ortogonal-unitaria (por ser producto de matrices ortogonales-unitarias) tal que $H\, A$ es triangular superior.

Se verifica que

$$\det(A) = \pm\, a_{1,1}^2\, a_{2,2}^3 \cdots a_{n-1,n-1}^n\, a_{n,n}^n\,.$$

Todas las transformaciones de Householder H_k, $k = 1, 2, \ldots, n-1$ que se han realizado con los elementos de la matriz A para obtener $H\, A$, deben de realizarse simultáneamente con los elementos del segundo miembro B para obtener también $H\, B$.

Teorema 23 *Dada $A \in \mathcal{M}_n(\mathbb{K})$, existe $Q \in \mathcal{M}_n(\mathbb{K})$ ortogonal-unitaria y $R \in \mathcal{M}_n(\mathbb{K})$ triangular superior tales que*

$$A = Q\, R\,;$$

además se puede elegir R de manera que todos sus elementos diagonales sean no negativos. Si la matriz A es inversible, la matriz R también lo es y por tanto se puede concluir que existe su factorización $Q\, R$; esta factorización es única si se impone la positividad de los elementos diagonales de R.

8. Generalidades sobre los métodos iterativos

Los métodos iterativos de resolución del Problema 1 tratan de construir una sucesión de "soluciones aproximadas" (aproximaciones de la solución) $\{X_k = (x_i^k)_{i=1}^n\}_{k\in\mathbb{N}}$, convergente hacia la solución (exacta) $X = (x_i)_{i=1}^n$ del mismo.

Normalmente estos métodos se presentan bajo la forma

$$X_0 \quad \text{dado}, \quad X_{k+1} = C\,X_k + V\,, \quad k = 0, 1, \ldots,$$

donde $C \in \mathcal{M}_n(\mathbb{K})$ y $V \in \mathbb{K}^n$ dependen de los datos del problema tratado.

Las dos cuestiones que se asocian a cualquier método iterativo como el anterior son:

- convergencia de $\{X_k\}_{k\in\mathbb{N}}$ hacia X, sin lo cual el método deja de ser útil;
- rapidez de convergencia y estimaciones del error $\|X_k - X\|$; esto es necesario pues en la práctica sólo se pueden calcular los "primeros elementos" de la sucesión y es preciso que el úttimo elemento que se calcule esté "suficientemente próximo" a la solución exacta.

Un ejemplo.

Consideremos el siguiente ejemplo de resolución de un sistema lineal mediante un método iterativo:

$$\left\{\begin{array}{rcrcrcr} 3\,x_1 & + & x_2 & + & x_3 & = & -1 \\ -x_1 & + & 4\,x_2 & + & x_3 & = & -8 \\ 2\,x_1 & + & x_2 & + & 5\,x_3 & = & -14 \end{array}\right\} \quad \text{Solución:} \quad \left\{\begin{array}{rcr} x_1 & = & 1 \\ x_2 & = & -1 \\ x_3 & = & -3 \end{array}\right\}$$

Despejando de cada ecuación la incógnita correspondiente nos queda:

$$\left\{\begin{array}{rcrcrcr} x_1 & = & \frac{1}{3}(& -1 & - & x_2 & - \; x_3) \\ x_2 & = & \frac{1}{4}(& -8 & + & x_1 & - \; x_3) \\ x_3 & = & \frac{1}{5}(& -14 & - & 2\,x_1 & - \; x_2) \end{array}\right\}$$

Lo anterior sugiere el siguiente método iterativo para calcular la solución del sistema:

$$\left\{ \begin{array}{c} x_1^0 \\ x_2^0 \\ x_3^0 \end{array} \right\} \text{ dado, } \left\{ \begin{array}{rcrcrcr} x_1^{k+1} & = & \frac{1}{3}(& -1 & - & x_2^k & - & x_3^k) \\ x_2^{k+1} & = & \frac{1}{4}(& -8 & + & x_1^k & - & x_3^k) \\ x_3^{k+1} & = & \frac{1}{5}(& -14 & - & 2\,x_1^k & - & x_2^k) \end{array} \right\}, \; k = 0, 1, \ldots$$

Tomando como solución inicial $x_1^0 = x_2^0 = x_3^0 = 0$, haciendo los cálculos en doble precisión y redondeando a la décima cifra decimal, obtenemos:

k	x_1^k	x_2^k	x_3^k
1	-0.3333333333	-2.0000000000	-2.8000000000
2	+1.2666666667	-1.3833333333	-2.2666666667
3	+0.8833333333	-1.1166666667	-3.0300000000
4	+1.0488888889	-1.0216666667	-2.9300000000
5	+0.9838888889	-1.0052777778	-3.0152222222
6	+1.0068333333	-1.0002222222	-2.9925000000
7	+0.9975740741	-1.0001666667	-3.0026888889
8	+1.0009518519	-0.9999342593	-2.9989962963
9	+0.9996435185	-1.0000129630	-3.0003938889
10	+1.0001356173	-0.9999906481	-2.9998548148
11	+0.9999484877	-1.0000023920	-3.0000561173
12	+1.0000195031	-0.9999988488	-2.9999789167
13	+0.9999925885	-1.0000003951	-3.0000080315
14	+1.0000028088	-0.9999998450	-2.9999969564
15	+0.9999989338	-1.0000000587	-3.0000011545
16	+1.0000004044	-0.9999999779	-2.9999995618
17	+0.9999998466	-1.0000000085	-3.0000001662
18	+1.0000000582	-0.9999999968	-2.9999999369
19	+0.9999999779	-1.0000000012	-3.0000000239
20	+1.0000000084	-0.9999999995	-2.9999999909
21	+0.9999999968	-1.0000000002	-3.0000000034
22	+1.0000000012	-0.9999999999	-2.9999999987
23	+0.9999999995	-1.0000000000	-3.0000000005
24	+1.0000000002	-1.0000000000	-2.9999999998
25	+0.9999999999	-1.0000000000	-3.0000000001
26	+1.0000000000	-1.0000000000	-3.0000000000
27	+1.0000000000	-1.0000000000	-3.0000000000

El anterior método iterativo se debe mejorar si en cada iteración se utilizan las componentes ya calculadas; esto conduce a:

$$\left\{ \begin{array}{c} x_1^0 \\ x_2^0 \\ x_3^0 \end{array} \right\} \text{ dado, } \left\{ \begin{array}{rcrcrcr} x_1^{k+1} & = & \frac{1}{3}(& -1 & - & x_2^k & - & x_3^k) \\ x_2^{k+1} & = & \frac{1}{4}(& -8 & + & x_1^{k+1} & - & x_3^k) \\ x_3^{k+1} & = & \frac{1}{5}(& -14 & - & 2\,x_1^{k+1} & - & x_2^{k+1}) \end{array} \right\}, \; k = 0, 1, \ldots$$

Tomando de nuevo como solución inicial $x_1^0 = x_2^0 = x_3^0 = 0$, haciendo los cálculos en doble precisión y redondeando a la décima cifra decimal, obtenemos:

k	x_1^k	x_2^k	x_3^k
1	-0.3333333333	-2.0833333333	-2.2500000000
2	+1.1111111111	-1.1597222222	-3.0125000000
3	+1.0574074074	-0.9825231481	-3.0264583333
4	+1.0029938272	-0.9926369599	-3.0026701389
5	+0.9984356996	-0.9997235404	-2.9994295718
6	+0.9997177040	-1.0002131810	-2.9998444454
7	+1.0000192088	-1.0000340864	-3.0000008662
8	+1.0000116509	-0.9999968707	-3.0000052862
9	+1.0000007190	-0.9999984987	-3.0000005879
10	+0.9999996955	-0.9999999292	-2.9999998924
11	+0.9999999405	-1.0000000418	-2.9999999678
12	+1.0000000032	-1.0000000072	-2.9999999998
13	+1.0000000024	-0.9999999995	-3.0000000011
14	+1.0000000002	-0.9999999997	-3.0000000001
15	+0.9999999999	-1.0000000000	-3.0000000000
16	+1.0000000000	-1.0000000000	-3.0000000000
17	+1.0000000000	-1.0000000000	-3.0000000000

El criterio seguido para detener el cálculo de las iteraciones es que dos consecutivas, una vez redondeadas a diez cifras decimales, sean iguales. Como puede observarse se ha producido con el segundo método una mejora en la rapidez de convergencia hacia la solución exacta del problema.

9. Convergencia de un método iterativo

Dado el Problema 1, supongamos que $C \in \mathcal{M}_n(\mathbb{K})$ y $V \in \mathbb{K}^n$ son tales que $I - C$ es inversible y que la única solución de

$$X = C\,X + V$$

coincide con la del Problema 1; esto se consigue, por ejemplo, si existe $F \in \mathcal{M}_n(\mathbb{K})$ inversible tal que $F\,A = I - C$ y $F\,B = V$.

La expresión anterior sugiere la construcción del método iterativo

$$X_0 \quad \text{dado}, \quad X_{k+1} = C\,X_k + V\,, \quad k = 0, 1, \ldots\,;$$

la matriz C recibe el nombre de **matriz del método iterativo**. Es claro a partir de esta construcción, que el método iterativo, de converger, lo hace hacia la solución del Problema 1.

Definición 24 *Se dice que el método iterativo anterior es convergente si $\{X_k\}_{k\in\mathbb{N}}$ converge hacia X, solución del Problema 1, cualquiera que sea X_0.*

Teorema 25 *Las siguientes proposiciones son equivalentes:*

1. *el método iterativo anterior es convergente;*
2. $$\rho(C) < 1\,;$$
3. *para al menos una norma matricial subordinada*
$$\|C\| < 1\,.$$

Observación 26 *El método iterativo anterior no es sino un método de aproximaciones sucesivas (también llamado método de Picard) utilizado para encontrar un punto fijo de la aplicación*

$$\left.\begin{array}{rcl} G\colon \mathbb{K}^n & \to & \mathbb{K}^n\,, \\ X & \mapsto & G(X) = C\,X + V\,. \end{array}\right\}$$

Observamos que la condición de convergencia sobre la norma asegura que la aplicación anterior es contractiva, pues

$$\|G(X) - G(Y)\| = \|C\,X + V - C\,Y - V\| = \|C\,(X - Y)\| \leq \|C\|\,\|X - Y\|\,.$$

10. Velocidad de convergencia

Consideramos un método iterativo como el definido anteiormente; supuesta la matriz C normal entonces llamando

$$E_k = X_k - X\,, \quad k = 0, 1, \dots\,,$$

se tiene que

$$\|X_k - X\|_2 = \|E_k\|_2 = \|C^k\, E_0\|_2 \leq \|C^k\|_2\, \|E_0\|_2 = (\rho(C))^k\, \|E_0\|_2\,,$$

de donde se deduce, en este caso, que el método iterativo tiene una convergencia más rápida en cuanto $\rho(C)$ (que debe ser menor que 1 para que haya convergencia) sea más peque no.

Teorema 27 *Sea $\|\cdot\|$ una norma vectorial y consideremos un método iterativo. Se verifica que:*

1. $$\lim_{k\to\infty} \sup_{\|X_0 - X\|=1} \|X_k - X\|^{1/k} = \rho(C)\,;$$

2. *si se considera un segundo método iterativo, cuya matriz es $\tilde{C}$ (todo lo relacionado con este segundo método iterativo es notado con una tilde) y se tiene que*

   $$\rho(C) < \rho(\tilde{C})\,, \quad y \quad X_0 = \tilde{X}_0\,,$$

   entonces para todo $\epsilon > 0$ existe $l = l(\epsilon) \in \mathbb{N}$ tal que para todo $k \geq l$ se tiene que

   $$\sup_{\|X_0 - X\|=1} \left(\frac{\|\tilde{X}_k - X\|}{\|X_k - X\|} \right)^{1/k} \geq \frac{\rho(\tilde{C})}{\rho(C) + \epsilon}\,.$$

11. Descripción de los métodos iterativos clásicos

Se trata de ver distintas maneras de construir C y V a partir de A y B.

En los métodos clásicos se parte de una descomposición

$$A = M - N\,,$$

con M inversible y "fácil de invertir." Entonces

$$A\,X = B \iff M\,X = N\,X + B \iff X = M^{-1}\,N\,X + M^{-1}\,B\,,$$

de donde se construye el método iterativo

$$X_0 \quad \text{dado}, \quad X_{k+1} = M^{-1}\,N\,X_k + M^{-1}\,B\,, \quad k = 0, 1, \ldots\,;$$

se observa que en este caso

$$\begin{aligned} C &= M^{-1}\,N = I - M^{-1}\,A\,, \\ V &= M^{-1}\,B\,, \end{aligned}$$

de donde se concluye, en particular, que $I - C = M^{-1}\,A$ es inversible.

En la práctica, en cada iteración se resuelve el sistema lineal

$$X_0 \quad \text{dado}, \quad M\,X_{k+1} = N\,X_k + B\,, \quad k = 0, 1, \ldots\,,$$

de donde la necesidad de que M sea "fácil de invertir". Una primera idea es que en la descomposición de A como diferencia de M y N, dichas matrices "no se solapen," por ejemplo tomando M como la parte diagonal (supuesta inversible) y $-N$ el resto, o bien M la parte triangular inferior y $-N$ la parte estrictamente triangular superior, etc. Parece intuitivo entonces que "mientras más grande" sea M "más rápidez de convergencia" tendrá el método, siendo el caso límite aquel en que $M = A$ y $N = 0$ en el que el método converge en una sola iteración.

Para describir los métodos iterativos clásicos de Jacobi, Gauss-Seidel y relajación, utilizamos la descomposición

$$A = D - E - F\,,$$

donde D es parte diagonal de la matriz, $-E$ la parte estrictamente triangular inferior y $-F$ la parte estrictamente triangular superior; suponemos que $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$, con lo que D es inversible.

El método de Jacobi es aquel en el que se toma

$$\begin{aligned} M &= D\,, \\ N &= E+F\,, \end{aligned}$$

de donde

$$\begin{aligned} J = C &= D^{-1}\,(E+F)\,, \\ V &= D^{-1}\,B\,. \end{aligned}$$

Para describir la forma de las iteraciones observamos que:

$$\begin{aligned} A\,X = B &\iff (D-(E+F))\,X = B \\ &\iff D\,X = (E+F)\,X + B \\ &\iff X = D^{-1}\,(E+F)\,X + D^{-1}\,B\,, \end{aligned}$$

lo que induce el método iterativo

$$X_0 \quad \text{dado,} \quad X_{k+1} = D^{-1}\,(E+F)\,X_k + D^{-1}\,B\,, \quad k = 0, 1, \ldots\,;$$

o bien

$$X_0 \quad \text{dado,} \quad D\,X_{k+1} = (E+F)\,X_k + B\,, \quad k = 0, 1, \ldots\,.$$

Si desarrollamos las ecuaciones que determinan la forma de las iteraciones obtenemos:

$$\left\{\begin{array}{rcccccccccc} a_{1,1}\,x_1^{k+1} &=& & & -\ a_{1,2}\,x_2^k & - & \cdots & - & a_{1,n}\,x_n^k & + & b_1 \\ a_{2,2}\,x_2^{k+1} &=& -\ a_{2,1}\,x_1^k & & & - & \cdots & - & a_{2,n}\,x_n^k & + & b_2 \\ \quad\cdots & & & & & & & & & & \\ a_{n,n}\,x_n^{k+1} &=& -\ a_{n,1}\,x_1^k & - & a_{n,2}\,x_2^k & - & \cdots & & & + & b_n \end{array}\right\}$$

El método de Gauss-Seidel es aquel en el que se toma

$$\begin{aligned} M &= D - E\,, \\ N &= F\,, \end{aligned}$$

de donde

$$\begin{aligned} \mathcal{L}_1 = C &= (D-E)^{-1}\,F\,, \\ V &= (D-E)^{-1}\,B\,. \end{aligned}$$

Para describir la forma de las iteraciones observamos que:

$$\begin{aligned} A\,X = B &\iff ((D-E)-F)\,X = B \\ &\iff (D-E)\,X = F\,X + B \\ &\iff X = (D-E)^{-1}\,F\,X + (D-E)^{-1}\,B\,, \end{aligned}$$

lo que induce el método iterativo

$$X_0 \quad \text{dado,} \quad X_{k+1} = (D-E)^{-1}\,F\,X_k + (D-E)^{-1}\,B\,, \quad k = 0, 1, \ldots\,;$$

o bien

$$X_0 \quad \text{dado,} \quad (D-E)\,X_{k+1} = F\,X_k + B\,, \quad k = 0, 1, \ldots\,,$$

o incluso

$$X_0 \quad \text{dado,} \quad D\,X_{k+1} = E\,X_{k+1} + F\,X_k + B\,, \quad k = 0, 1, \ldots\,.$$

Si desarrollamos las ecuaciones que determinan la forma de las iteraciones obtenemos:

$$\left\{ \begin{array}{lcccccccccc} a_{1,1}\,x_1^{k+1} & = & & & - & a_{1,2}\,x_2^k & - & \cdots & - & a_{1,n}\,x_n^k & + & b_1 \\ a_{2,2}\,x_2^{k+1} & = & - & a_{2,1}\,x_1^{k+1} & & & - & \cdots & - & a_{2,n}\,x_n^k & + & b_2 \\ \quad \cdots & & & & & & & & & & & \\ a_{n,n}\,x_n^{k+1} & = & - & a_{n,1}\,x_1^{k+1} & - & a_{n,2}\,x_2^{k+1} & - & \cdots & & & + & b_n \end{array} \right\}$$

Por último, el método de relajación es aquel en el que se toma

$$\begin{aligned} M &= \tfrac{1}{\omega}\, D - E\,, \\ N &= \tfrac{1-\omega}{\omega}\, D + F\,, \end{aligned}$$

donde $\omega \in \mathbb{R} - \{0\}$ (llamado parámetro de relajación), de donde

$$\begin{aligned} \mathcal{L}_\omega = C &= \left(\tfrac{1}{\omega}\, D - E\right)^{-1} \left(\tfrac{1-\omega}{\omega}\, D + F\right)\,, \\ V &= \left(\tfrac{1}{\omega}\, D - E\right)^{-1} B\,. \end{aligned}$$

El método de Gauss-Seidel es un caso particular del método de relajación, correspondiente a $\omega = 1$. Cuando $\omega < 1$ se habla de sub-relajación y cuando $\omega > 1$ se habla de sobre-relajación.

Para describir la forma de las iteraciones observamos que:

$$\begin{aligned} A\,X = B &\iff ((\tfrac{1}{\omega}\, D - E) - (\tfrac{1-\omega}{\omega}\, D + F)\, X = B \\ &\iff (\tfrac{1}{\omega}\, D - E)\, X = (\tfrac{1-\omega}{\omega}\, D + F)\, X + B \\ &\iff X = (\tfrac{1}{\omega}\, D - E)^{-1}\, (\tfrac{1-\omega}{\omega}\, D + F)\, X + (\tfrac{1}{\omega}\, D - E)^{-1}\, B\,, \end{aligned}$$

lo que induce el método iterativo

$$X_0 \quad \text{dado,} \quad X_{k+1} = (\tfrac{1}{\omega}\, D - E)^{-1}\, (\tfrac{1-\omega}{\omega}\, D + F)\, X_k + (\tfrac{1}{\omega}\, D - E)^{-1}\, B\,, \quad k = 0, 1, \dots\,;$$

o bien

$$X_0 \quad \text{dado,} \quad (\tfrac{1}{\omega}\, D - E)\, X_{k+1} = (\tfrac{1-\omega}{\omega}\, D + F)\, X_k + B\,, \quad k = 0, 1, \dots\,,$$

o incluso

$$X_0 \quad \text{dado,} \quad \tfrac{1}{\omega}\, D\, X_{k+1} = E\, X_{k+1} + \tfrac{1-\omega}{\omega}\, D\, X_k + F\, X_k + B\,, \quad k = 0, 1, \dots\,.$$

Si desarrollamos las ecuaciones que determinan la forma de las iteraciones obtenemos:

$$\left\{ \begin{array}{rcccccccccc} \frac{1}{\omega}\, a_{1,1}\, x_1^{k+1} & = & + & \frac{1-\omega}{\omega}\, a_{1,1}\, x_1^k & - & a_{1,2}\, x_2^k & - \cdots - & a_{1,n}\, x_n^k & + & b_1 \\ \frac{1}{\omega}\, a_{2,2}\, x_2^{k+1} & = & - & a_{2,1}\, x_1^{k+1} & + & \frac{1-\omega}{\omega}\, a_{2,2}\, x_2^k & - \cdots - & a_{2,n}\, x_n^k & + & b_2 \\ \quad \cdots & & & & & & & & & \\ \frac{1}{\omega}\, a_{n,n}\, x_n^{k+1} & = & - & a_{n,1}\, x_1^{k+1} & - & a_{n,2}\, x_2^{k+1} & - \cdots + & \frac{1-\omega}{\omega}\, a_{n,n}\, x_n^k & + & b_n \end{array} \right\}$$

12. Convergencia de los métodos clásicos

Teorema 28 *Sea $A \in \mathcal{M}_n(\mathbb{K})$ definida positiva, descompuesta en la forma*

$$A = M - N\,, \quad M, N \in \mathcal{M}_n(\mathbb{K})\,, \quad M \textit{ inversible}\,.$$

Si la matriz $M^t + N$ (resp. $M^* + N$), que es simétrica-hermítica, es definida positiva, entonces

$$\rho(M^{-1}\,N) < 1\,,$$

con lo que el método iterativo correspondiente es convergente.

Teorema 29 (Teorema de Ostrowski-Reich) *Si la matriz $A \in \mathcal{M}_n(\mathbb{K})$ es definida positiva y si $\omega \in (0, 2)$, entonces el método de relajación converge (condición suficiente de convergencia del método de relajación).*

Teorema 30 *El radio espectral de la matriz $\mathcal{L}_\omega$ del método de relajación verifica que*

$$\rho(\mathcal{L}_\omega) \geq |1 - \omega|\,;$$

por tanto una condición necesaria para que el método de relación converja es que $\omega \in (0, 2)$.

Corolario 31 *Si la matriz $A \in \mathcal{M}_n(\mathbb{K})$ es definida positiva, el método de relajación converge si y sólo si $\omega \in (0, 2)$.*

Caso de matrices tridiagonales.

Teorema 32 *Sea $A \in \mathcal{M}_n(\mathbb{K})$ tridiagonal. Entonces los radios espectrales de las matrices de Jacobi y de Gauss-Seidel correspondientes verifican que*

$$\rho(\mathcal{L}_1) = (\rho(J))^2 .$$

En consecuencia, ambos métodos convergen o divergen simultáneamente y en caso de converger el método de Gauss-Seidel lo hace más rápidamente.

Teorema 33 *Sea $A \in \mathcal{M}_n(\mathbb{K})$ tridiagonal y tal que $\mathrm{sp}(J) \subset \mathbb{R}$. Entonces los métodos de Jacobi y de relajación con $\omega \in (0,2)$ convergen o diverge simultáneamente; en caso de convergencia la función $(0,2) \to \mathbb{R}/\omega \mapsto \rho(\mathcal{L}_\omega)$ tiene el siguiente aspecto:*

Teorema 34 *Sea $A \in \mathcal{M}_n(\mathbb{K})$ definida positiva y tridiagonal. Entonces los métodos de Jacobi, Gauss-Seidel y relajación con $\omega \in (0,2)$ convergen; además, la función $(0,2) \to \mathbb{R}/\omega \mapsto \rho(\mathcal{L}_\omega)$ tiene el aspecto dado anteriormente, por lo que existe un único parámetro de relajación optimal*

$$\omega_0 = \frac{2}{1 + \sqrt{1 - (\rho(J))^2}}$$

tal que

$$\rho(\mathcal{L}_{\omega_0}) = \min_{\omega \in (0,2)} \rho(\mathcal{L}_\omega) = \omega_0 - 1 < \rho(\mathcal{L}_1) = (\rho(J))^2 < \rho(J)$$

si $\rho(J) > 0$; si $\rho(J) = 0$ entonces $\omega_0 = 1$ y $\rho(\mathcal{L}_1) = \rho(J) = 0$.