Introducción al Álgebra Lineal Numérica

M. Dolores Martínez
e-mail: dolores.martinez@uv.es

Javier Pastor
e-mail: pastorv@uv.es

Departamento de Matemática Aplicada
Universidad de Valencia

Septiembre de 2014

Resumen

Estos apuntes corresponden a la materia impartida por los autores en la asignatura Métodos Numéricos para el Álgebra Lineal (Licenciatura de Matemáticas, plan 2000, Facultad de Matemáticas de la Universidad de Valencia) desde el curso 2003-2004 hasta el 2011-2012, a partir del que entró en vigor el nuevo Grado en Matemáticas.

Índice general

Capítulo 1

Introducción

En este primer tema se presentan diversas aplicaciones que conducen a la resolución de sistemas de ecuaciones lineales. En las dos primeras situaciones se llega a un sistema con matriz de coeficientes con estructura especial: matriz tridiagonal, simétrica y definida positiva. Este hecho nos llevará a prestar especial atención a dichas clases de matrices durante todo el curso.

Como tercer problema aplicado estudiaremos la cuestión de ajustar una nube de puntos 'de la mejor manera posible' mediante una recta. Este problema se traducirá en el de encontrar soluciones de un sistema lineal sobredeterminado; es decir, con más ecuaciones que incógnitas, que evidentemente puede no tener soluciones en el sentido clásico o habitual, y se introducirá y analizará para esta situación el concepto de solución de mínimos cuadrados.

Por último, comentaremos como influyen en el carácter de las soluciones de ecuaciones diferenciales lineales los valores propios de una matriz. En el último capítulo de este manual trataremos la problemática de aproximar valores y vectores propios de una matriz.

1.1. Interpolación mediante una clase de splines cúbicos

Los splines se utilizan para aproximar formas complicadas por su simplicidad de representación y facilidad de cálculo. En particular, en el terreno de gráficos por ordenador.

Recordemos que los valores de un polinomio de grado menor o igual que 3, $q(x)$, y de su derivada primera en los extremos de un intervalo $[a, b]$ determinan de forma unívoca éste (interpolación de Hermite). En efecto, si $q(x) = q_0 + q_1 x + q_2 x^2 + q_3 x^3$, entonces debemos resolver el sistema de ecuaciones lineales, compatible determinado,

$$\begin{bmatrix} 1 & a & a^2 & a^3 \\ 1 & b & b^2 & b^3 \\ 0 & 1 & 2a & 3a^2 \\ 0 & 1 & 2b & 3b^2 \end{bmatrix} \begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} q(a) \\ q(b) \\ q'(a) \\ q'(b) \end{bmatrix},$$

para determinar sus coeficientes en la base canónica $\{1, x, x^2, x^3\}$. Una situación similar se presenta con la interpolación mediante polinomios de grado cualquiera. No obstante, existe una expresión más cómoda de $q(x)$ en términos de una adecuada base de polinomios y de las llamadas diferencias divididas asociadas a q:

$$\begin{aligned} q(x) &= q[a] + q[a,a](x-a) + q[a,a,b](x-a)^2 \\ &\quad + q[a,a,b,b](x-a)^2(x-b), \end{aligned}$$

donde

$$\begin{aligned} q[a] &: = q(a),\ q[a,a] := q'(a), \\ q[a,a,b] &: = \frac{q(b) - q(a) - q'(a)(b-a)}{(b-a)^2} \\ q[a,a,b,b] &: = \frac{(q'(b) + q'(a))(b-a) + 2(q(a) - q(b))}{(b-a)^3}. \end{aligned}$$

Sea $u : [a,b] \to \mathbb{R}$. Consideramos una partición uniforme del intervalo $[a,b]$ de n subintervalos: $x_i = a + ih$, $i = 0, 1, ..., n$, $h := (b-a)/n$.

Vamos a plantearnos el problema de interpolación siguiente. Buscamos una función $u_h : [a,b] \to \mathbb{R}$ verificando:

(i) u_h es de clase C^2,

(ii) $u_h \mid_{[x_i, x_{i+1}]}$, $0 \le i \le n-1$, es un polinomio de grado menos o igual que 3,

(iii) $u_h(x_i) = u(x_i)$, $0 \le i \le n$ y $u_h'(a) = u'(a)$, $u_h'(b) = u'(b)$.

A la función u_h así determinada se le llama *spline cúbico* asociado a u (aunque hay otros splines asociados, según cuáles sean las condiciones adicionales que impongamos en los extremos, como por ejemplo la de conocer los valores de las derivadas segundas en a y b).

Sabemos que si conocieramos los valores de u y de u'en los nodos existiría una única función cumpliendo las condiciones previas excepto (i). En este caso sólo podríamos asegurar que la función obtenida uniendo los polinomios de tercer grado obtenidos en cada subintervalo por interpolación de Hermite es de clase C^1.

Vamos a comprobar que las derivadas primeras de u_h en los nodos interiores están determinadas por la condiciones impuestas a u_h. Por lo tanto, habremos probado que existe algún splin cúbico asociado a u y además que es único.

Supongamos que existe u_h. Para simplificar usaremos la notación $u_i := u_h(x_i) = u(x_i)$, $u_i' := u_h'(x_i)$, $0 \le i \le n$. Consideremos un subintervalo genérico de la partición $[x_i, x_{i+1}]$. Por el comentario inicial sabemos que

$$\begin{aligned} u_h \mid_{[x_i,x_{i+1}]}(x) &= u_i + u_i'(x - x_i) + \frac{u_{i+1} - u_i - u_i'h}{h^2}(x - x_i)^2 \\ &\quad + \frac{(u_{i+1}' + u_i')h + 2(u_i - u_{i+1})}{h^3}(x - x_i)^2(x - x_{i+1}), \end{aligned} \tag{1.1}$$

y por tanto la derivada segunda de u_h en el intervalo es

$$u_h'' \mid_{[x_i,x_{i+1}]}(x) = 2\frac{u_{i+1}-u_i-u_i'h}{h^2} + 2\frac{(u_{i+1}'+u_i')h+2(u_i-u_{i+1})}{h^3}((x-x_{i+1})+2(x-x_i)).$$

En consecuencia

$$u_h''(x_i^+) = 6\frac{u_{i+1}-u_i}{h^2} - 2\frac{(u_{i+1}'+2u_i')}{h}$$

y

$$u_h''(x_{i+1}^-) = 6\frac{u_i-u_{i+1}}{h^2} + 2\frac{(2u_{i+1}'+u_i')}{h},$$

entendidas como derivadas laterales. Si u_h es de clase C^2 las derivadas laterales segundas en cada nodo interior de la partición deben coincidir. Por tanto, tenemos

$$6\frac{u_{i+1}-u_i}{h^2} - 2\frac{(u_{i+1}'+2u_i')}{h} = 6\frac{u_{i-1}-u_i}{h^2} + 2\frac{(2u_i'+u_{i-1}')}{h} \quad (1\le i\le n-1),$$

es decir

$$u_{i-1}'+4u_i'+u_{i+1}' = 3\frac{u_{i+1}-u_{i-1}}{h} \quad (1\le i\le n-1),$$

lo que podemos reescribir en forma matricial:

$$\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix} \begin{bmatrix} u_1' \\ u_2' \\ \vdots \\ u_{n-2}' \\ u_{n-1}' \end{bmatrix} = \begin{bmatrix} 3\frac{u_2-u_0}{h} - u_0' \\ 3\frac{u_3-u_1}{h} \\ \vdots \\ 3\frac{u_{n-1}-u_{n-3}}{h} \\ 3\frac{u_n-u_{n-2}}{h} - u_n' \end{bmatrix}. \tag{1.2}$$

Observar que la matriz de coeficientes, A, es invertible:

$$\begin{aligned} v^TAv &= 3(v_1^2+v_{n-1}^2) + 2\sum_{i=2}^{n-2} v_i^2 + \sum_{i=1}^{n-2}(v_i+v_{i+1})^2 \\ &\ge 2\sum_{i=1}^{n-1} v_i^2 > 0 \quad (v\in\mathbb{R}^{n-1} \text{ no nulo}), \end{aligned}$$

de lo que se sigue la regularidad de A. En efecto, $Av=0$, entonces $v^TAv=0$, y en consecuencia, $v=0$. Resaltar que A es simétrica, definida positiva y tridiagonal.

En consecuencia, las derivadas primeras están unívocamente determinadas por (1.2), lo que junto con (iii) permite precisar quién es u_h: es la función que se obtiene por interpolación segmentaria de Hermite con los datos hallados. Tenemos así probada la existencia y unicidad del splin cúbico.

Nota 1.1. *Observar que a diferencia de la interpolación a trozos en la que trabajamos en cada subintervalo, el splin cúbico asociado depende de los valores de la función en todos los nodos, lo que se traduce en la necesidad de resolver sistemas de ecuaciones lineales. A costa de este esfuerzo conseguimos una aproximación más regular con menos datos sobre la función a interpolar.*

Ejemplo 1.1. *Vamos a obtener el splin cúbico asociado a* $u(x) = x^2 - 3x + 2$ *en* $[0, 1]$*, dividido éste en* 5 *subintervalos. Resolviendo el correspondiente sistema*

$$\begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} u'_1 \\ u'_2 \\ u'_3 \\ u'_4 \end{bmatrix} = \begin{bmatrix} -12.6 \\ -13.2 \\ -10.8 \\ -7.4 \end{bmatrix},$$

obtenemos que las derivadas en los nodos de la partición son por orden

$$[-3; -2.6; -2.2, ; -1.8; -1.4; -1]\,,$$

valores que coinciden exactamente con las derivadas de u (lo que era predecible puesto que se trata de un polinomio de grado menor o igual que 3*). Por tanto, el polinomio obtenido con aritmética exacta es u.*

Se cumple el siguiente resultado sobre la bondad de la aproximación mediante la clase de splines cúbicos analizada (véase [7, Th. 2.4.3.3]).

Teorema 1.1. *Sea* $u \in C^4([a, b])$ *cumpliendo*

$$\left|u^{(iv)}(x)\right| \leq L \quad (x \in [a, b]).$$

Sea $n \in \mathbb{N}$*. Consideramos la partición uniforme del intervalo* $[a, b]$ *de paso* $h = (b-a)/n$*, y* u_h *el splin cúbico que interpola a u en los nodos de la partición y cuya derivada toma los valores* $u'(a)$ *y* $u'(b)$ *en a y b respectivamente. Entonces*

$$\left|u^{(k)}(x) - u_h^{(k)}(x)\right| \leq 2Lh^{4-k} \quad (x \in [a, b], \quad k = 0, 1, 2, 3).$$

Nota 1.2. *Existen otro tipo de splines cúbicos como son los llamados* ***splines cúbicos naturales****. En las condiciones (i)-(iii) que definen el splin cúbico se sustituye en (iii) el conocimiento de la derivada en los extremos por la condición de anularse la derivada segunda en los extremos. En el extremo a, por ejemplo, obtendríamos de dicha condición que*

$$\frac{3(u_1 - u_0)}{2h} - \frac{u'_1}{2} = u'_0,$$

con lo que la primera ecuación del sistema (1.2) quedaría

$$\frac{7u'_1}{2} + u'_2 = \frac{3}{2h}(2u_2 - u_0 - u_1).$$

Un cambio análogo sufriría la última ecuación. Es evidente que la nueva matriz de coeficientes goza de las mismas características que en el caso expuesto con detalle.

1.2. Diferencias finitas para un problema de contorno

Para α, $\beta \in \mathbb{R}$ y f, $g : [a,b] \to \mathbb{R}$ consideramos el siguiente problema de contorno asociado a una ecuación diferencial lineal de segundo orden:

$$\begin{cases} -y'' + f(x)y = g(x), \ \ x \in [a,b], \\ y(a) = \alpha,\ y(b) = \beta \ \ \text{(condiciones de contorno separadas)}. \end{cases} \tag{PC}$$

Una solución de (PC) es una función $u : [a,b] \to \mathbb{R}$, dos veces derivable en el intervalo $[a,b]$, de forma que

$$-u''(x) + f(x)u(x) = g(x),\ \forall x \in [a,b];\ u(a) = \alpha,\ u(b) = \beta.$$

Ejemplo 1.2. *La función $u(x) = x^{-2}$, $x \in [1,2]$, es solución del problema de contorno*

$$\begin{cases} -y'' + 6x^{-2}y = 0, \\ y(1) = 1,\ y(2) = 1/4. \end{cases}$$

El siguiente ejemplo pone de manifiesto la problemática sobre la existencia y unicidad de solución de un problema de contorno como el propuesto.

Ejemplo 1.3. *Suponiendo conocido que todas las soluciones de $y'' + y = 0$ son de la forma $u(x) = A\sin(x) + B\cos(x)$, donde $A, B \in \mathbb{R}$ son constantes arbitrarias, es fácil convencerse de que el problema*

$$\begin{cases} y'' + y = 0, \\ y(0) = 0,\ y(\pi) = 1, \end{cases}$$

no tiene solución, mientras que

$$\begin{cases} y'' + y = 0, \\ y(0) = 0,\ y(\pi) = 0, \end{cases}$$

tiene como soluciones todas las funciones de la forma $u(x) = A\sin(x)$. Sin embargo, el problema

$$\begin{cases} y'' + y = 0, \\ y(0) = 0,\ y(\pi/2) = 1, \end{cases}$$

tiene como única solución $u(x) = \sin(x)$.

En el ámbito de los problemas de valores iniciales asociados a ecuaciones lineales no hay dificultades en cuanto a la existencia y unicidad de solución, de tal manera que si sustituimos las condiciones de contorno por las condiciones iniciales $y(c) = \alpha$, $y'(c) = \beta$, donde $c \in [a,b]$ y $\alpha, \beta \in \mathbb{R}$ son arbitrarios, el correspondiente problema tiene solución y es única, bajo la condición $f, g \in C([a,b])$. Bajo esta misma condición y la condición $f \geq 0$ en $[a,b]$, se puede probar que la solución (PC) existe y es única. En adelante asumimos dichas condiciones y denotaremos a la solución mediante u. Observar que $u \in C^2([a,b])$.

Consideremos una partición uniforme del intervalo $[a,b]$ de n subintervalos: $x_i = a + ih$, $i = 0, 1, ..., n$, $h := (b-a)/n$. Nuestro objetivo consiste en aproximar el valor de u en los nodos internos de la partición, es decir, aproximar $u(x_i)$, $1 \leq i \leq n-1$.

Por ser u solución de (PC) sabemos en particular que

$$-u''(x_i) + f(x_i)u(x_i) = g(x_i) \quad (1 \leq i \leq n-1). \tag{1.3}$$

Vía el desarrollo de Taylor vamos a expresar la cantidad desconocida $u''(x_i)$ con las incógnitas $u(x_i)$. Para ello supondremos adicionalmente que $f, g \in C^2([a,b])$, lo que implicará que $u \in C^4([a,b])$.

Trabajamos con un nodo interno x_i, $1 \leq i \leq n-1$. Por Taylor, existen puntos $\rho_i \in (x_i, x_{i+1})$ y $\sigma_i \in (x_{i-1}, x_i)$ tales que

$$u(x_{i+1}) = u(x_i) + u'(x_i)h + u''(x_i)\frac{h^2}{2} + u'''(x_i)\frac{h^3}{6} + u^{(iv)}(\rho_i)\frac{h^4}{24}$$

y

$$u(x_{i-1}) = u(x_i) - u'(x_i)h + u''(x_i)\frac{h^2}{2} - u'''(x_i)\frac{h^3}{6} + u^{(iv)}(\sigma_i)\frac{h^4}{24},$$

de donde sumando ambas igualdades

$$u''(x_i) = \frac{-u(x_{i+1}) + 2u(x_i) - u(x_{i-1})}{h^2} + r_i(h^2),$$

donde $r_i(h^2)$ representa el término en h^2 restante. Reescribiendo la relación (1.3) se tiene

$$\frac{-u(x_{i+1}) + 2u(x_i) - u(x_{i-1})}{h^2} + r_i(h^2) + f(x_i)u(x_i) = g(x_i) \quad (1 \leq i \leq n-1).$$

Usamos la notación $f_i = f(x_i)$ y $g_i = g(x_i)$. La idea consiste en aproximar $u(x_i)$ por las cantidades u_i que verifiquen el sistema de ecuaciones lineales

$$-u_{i+1} + (2 + h^2 f_i)u_i - u_{i-1} = h^2 g_i \quad (1 \leq i \leq n-1),$$

es decir, despreciar los terminos desconocidos $r_i(h^2)$. Como $u(x_0) = \alpha$ y $u(x_n) = \beta$, el sistema se puede escribir matricialmente como sigue:

$$\begin{bmatrix} 2+f_1h^2 & -1 & & & \\ -1 & 2+f_2h^2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2+f_{n-2}h^2 & -1 \\ & & & -1 & 2+f_{n-1}h^2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-2} \\ u_{n-1} \end{bmatrix}$$

$$= \begin{bmatrix} g_1h^2 + \alpha \\ g_2h^2 \\ \vdots \\ g_{n-2}h^2 \\ g_{n-1}h^2 + \beta \end{bmatrix}. \tag{1.4}$$

Sea A_h la matriz de coeficientes del sistema. Nótese que A_h es tridiagonal, estrictamente diagonalmente dominante, si $f > 0$, y simétrica. Además, es definida positiva. En efecto,

$$\begin{aligned} v^T A_h v &= v_1^2 + v_{n-1}^2 + \sum_{i=1}^{n-1} f_i h^2 v_i^2 + \sum_{i=1}^{n-2} (v_i - v_{i+1})^2 \\ &\geq v_1^2 + v_{n-1}^2 + \sum_{i=1}^{n-2} (v_i - v_{i+1})^2 > 0, \end{aligned}$$

para todo $v \in \mathbb{R}^{n-1}$ no nulo. En particular, A_h es regular.

En el siguiente teorema se pone de manifiesto que bajo las hipótesis realizadas

$$\max_{1 \leq i \leq n-1} |u_i - u(x_i)| = O(h^2).$$

Teorema 1.2 ([7, Th. 7.4.10])**.** *Supongamos que el problema (PC) tiene solución única $u \in C^4([a,b])$ y que $f(x) \geq 0$, $x \in [a,b]$. Sea*

$$\left| u^{(iv)}(x) \right| \leq M \quad (x \in [a,b]).$$

Sea $[u_1; u_2; \ldots; u_{n-1}]$ la solución del sistema (1.4). Entonces

$$|u_i - u(x_i)| \leq \frac{M}{24} h^2 (x_i - a)(b - x_i) \quad (i = 1, 2, \ldots, n-1).$$

Ejemplo 1.4. *Considérese el problema de contorno*

$$\begin{cases} y'' = 0, \; x \in [0,1], \\ y(0) = 0, \; y(1) = 1, \end{cases}$$

cuya única solución es $u(x) = x$, $x \in [0,1]$. Consideremos una partición de $[0,1]$ en cuatro subintervaloses ($n = 4$). Hemos de resolver el sistema

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

cuya solución es

$$\begin{pmatrix} 0.25 \\ 0.5 \\ 0.75 \end{pmatrix}.$$

En este caso especial obtenemos los valores exactos, ya que las derivadas cuartas de la solución son cero y, por lo tanto, no hay error de discretización.

Ejemplo 1.5. *La única solución del problema de contorno*

$$\begin{cases} y'' + y = 0, \; x \in [0,1], \\ y(0) = 0, \; y(1) = 1, \end{cases}$$

es $u(x) = \sin(x)/\sin(1)$, $x \in [0,1]$. *Para* $n = 4$ *tenemos el sistema*

$$\begin{pmatrix} 1.9375 & -1 & 0 \\ -1 & 1.9375 & -1 \\ 0 & -1 & 1.9375 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

cuya solución es

$$\begin{pmatrix} 0.2943 \\ 0.5702 \\ 0.8104 \end{pmatrix}.$$

Así,

$$|u_1 - u(0.25)| \leq 2.7 \times 10^{-4}.$$

1.3. Un sistema lineal sobredeterminado: la recta de regresión lineal

Consideremos un conjunto de puntos del plano $\{(x_i, y_i)\}_{i=1}^m$. Nos planteamos el problema de encontrar una recta $y = \alpha x + \beta$ que contenga dichos puntos. Es evidente que este problema no tiene solución a menos que los puntos estén alineados. Los parámetros α, $\beta \in \mathbb{R}$ que determinan la recta han de ser solución del sistema de ecuaciones lineales

$$-\alpha x_i + y_i = \beta \quad (1 \leq i \leq m), \tag{1.5}$$

que tiene más ecuaciones que incógnitas para $m \geq 3$, de ahí que se le califique como sobredeterminado. Por la posible inexistencia de solución de (1.5) nos planteamos la búsqueda de solución en el sentido de que se verifique lo mejor posible el sistema (1.5). Nos centraremos en el problema de encontrar la llamada solución de *mínimos cuadrados*, que consiste en resolver el problema de minimización

$$\min_{\alpha,\beta \in \mathbb{R}} \sum_{i=1}^{m} (-\alpha x_i + y_i - \beta)^2.$$

Se trata pues de minimizar globalmente la función $f(\alpha, \beta) = \sum_{i=1}^m (-\alpha x_i + y_i - \beta)^2$, α, $\beta \in \mathbb{R}$. Es conocido que el mínimo, de existir, debe ser un punto crítico de la función de ahí que deba verificar el sistema lineal cuadrado

$$\begin{pmatrix} \sum_{i=1}^m x_i^2 & \sum_{i=1}^m x_i \\ \sum_{i=1}^m x_i & m \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m x_i y_i \\ \sum_{i=1}^m y_i \end{pmatrix}. \tag{1.6}$$

Este sistema es el llamado *sistema de ecuaciones normales* asociado a (1.5), que es un sistema cuadrado con matriz simétrica semidefinida positiva.

Para lo que sigue es conveniente introducir la siguiente notación: la *media* de x

$$\overline{x} = \frac{1}{m} \sum_{i=1}^{m} x_i,$$

y la *covarianza* de x e y

$$S_{xy} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{m}\sum_{i=1}^{m} x_i y_i - \overline{xy},$$

también llamada varianza si $x = y$, y los correspondientes análogos $\overline{y}$ y S_{yy}.

El determinante de la matriz de coeficientes de (1.6) vale

$$m\sum_{i=1}^{m} x_i^2 - (\sum_{i=1}^{m} x_i)^2 = m\sum_{i=1}^{m}(x_i - \overline{x})^2.$$

Por tanto, la matriz es singular si, y sólo si,

$$m\, x_i = \sum_{j=1}^{m} x_j \quad (1 \le i \le m),$$

o, equivalentemente, si todas las abscisas de todos los puntos son iguales (lo que implicaría que los puntos estarían en una recta vertical). Por tanto, si hay al menos dos abscisas distintas, lo que supondremos en lo que sigue, entonces existe una única solución de (1.6). Obsérvese que esta condición es equivalente a solicitar que la matriz de coeficientes de (1.5) tenga rango máximo, en este caso 2.

Así, la solución del sistema (1.6) viene dada por

$$\alpha_0 = \frac{m\sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m} y_i \sum_{i=1}^{m} x_i}{d} = \frac{S_{xy}}{S_{xx}},$$

y

$$\beta_0 = \frac{\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i - \sum_{i=1}^{m} x_i y_i \sum_{i=1}^{m} x_i}{d} = \overline{y} - \overline{x}\frac{S_{xy}}{S_{xx}},$$

de tal suerte que la *recta de regresión lineal* es

$$y - \overline{y} = \frac{S_{xy}}{S_{xx}}(x - \overline{x}).$$

La matriz hessiana asociada a f es

$$H := \frac{1}{2}\begin{bmatrix} \frac{\partial^2 f}{\partial^2 \alpha} & \frac{\partial^2 f}{\partial\alpha\partial\beta} \\ \frac{\partial^2 f}{\partial\alpha\partial\beta} & \frac{\partial^2 f}{\partial^2 \beta} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} x_i^2 & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & m \end{bmatrix},$$

que en las condiciones impuestas es definida positiva. En efecto,

$$\begin{aligned} [v_1, v_2]H[v_1, v_2]' &= v_1^2\sum_{i=1}^{m} x_i^2 + 2v_1v_2\sum_{i=1}^{m} x_i + mv_2^2 \\ &= m(S_{xx}v_1^2 + \overline{x}^2 v_1^2 + 2\overline{x}v_1v_2 + v_2^2) \\ &= m(S_{xx}v_1^2 + (\overline{x}v_1 + v_2)^2) > 0, \end{aligned}$$

si $v_1^2 + v_2^2 > 0$. Por consiguiente, el punto (α_0, β_0) es un punto de mínimo global (recordemos el razonamiento: por Taylor

$$\begin{aligned} f(\alpha,\beta) &= f(\alpha_0,\beta_0) + f_\alpha(\alpha_0,\beta_0)(\alpha-\alpha_0) + f_\beta(\alpha_0,\beta_0)(\beta-\beta_0) \\ &\quad + \frac{1}{2}(f_{\alpha\alpha}(\rho,\tau)(\alpha-\alpha_0)^2 + 2f_{\alpha\beta}(\rho,\tau)(\alpha-\alpha_0)(\beta-\beta_0) + f_{\beta\beta}(\rho,\tau)(\beta-\beta_0)^2) \\ &= f(\alpha_0,\beta_0) + [\alpha-\alpha_0, \beta-\beta_0]\, H\, [\alpha-\alpha_0, \beta-\beta_0]' \geq f(\alpha_0,\beta_0), \end{aligned}$$

donde la última desigualdad es consecuencia de ser H definida positiva).

Finalmente, utilizando que $\beta_0 = \overline{y} - \overline{x}\alpha_0$, el valor del mínimo es

$$\begin{aligned} f(\alpha_0,\beta_0) &= \sum_{i=1}^{m} (\alpha_0(x_i - \overline{x}) + \overline{y} - y_i)^2 \\ &= m\left(\alpha_0^2 S_{xx} + S_{yy} - 2\alpha_0 S_{xy}\right) = m\left(S_{yy} - S_{xy}^2/S_{xx}\right) \\ &= mS_{yy}\left(1 - \frac{S_{xy}^2}{S_{yy}S_{xx}}\right) \equiv mS_{yy}\left(1 - r^2\right), \end{aligned}$$

donde

$$r := \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

el llamado *coeficiente de correlación*. Nótese que $|r| \leq 1$. Por tanto, cuando $r = \pm 1$ los puntos están alineados y viceversa.

Nótese que $S_{yy} = 0$ es equivalente a tener todos los puntos alineados horizontalmente, en concreto sobre la recta $y = \overline{y}$.

1.4. Ecuaciones diferenciales lineales y valores propios

Un modelo sencillo para el crecimiento de una población $p(t)$ (también por ejemplo para describir desintegraciones radiactivas de primer orden) es el descrito por la ecuación diferencial lineal de primer orden

$$p'(t) = a\, p(t), \quad t \geq 0,$$

que establece que la variación instantánea de la población es proporcional a el número de individuos de la población en dicho instante. Esto es equivalente a afirmar que $p(t) = p(0)e^{at}$, $t \geq 0$, y por tanto, el comportamiento para valores grandes del tiempo depende de a, de tal manera que la población tiende a extinguirse si $a < 0$ y crece indefinidamente si $a > 0$. Obsérvese que se puede interpretar a como el único valor propio de la matriz $A = [a]$ de tal forma que el caracter de las soluciones de nuestra ecuación diferencial depende de los valores propios de la matriz A que la define.

La modelización de circuitos eléctricos o de sistemas formados por masas puntuales unidas por muelles, conduce a la resolución de ecuaciones diferenciales del tipo

$$y' = Ay,$$

donde A es una matriz cuadrada de tamaño n de números reales. Si por analogía con el caso unidimensional buscamos una solución del sistema vectorial de la forma $y(x) = ve^{\lambda x}$, donde v es un vector no nulo de tamaño n arbitrario y λ es un número complejo arbitrario, entonces se debería cumplir

$$Av = \lambda v,$$

es decir, λ debería ser un valor propio asociado A y v un vector propio asociado a λ.

Consideremos un caso ilustrativo sencillo, en el que tenemos un sistema formado por una masa puntual m suspendida de un techo por medio de un muelle de longitud l. Por la ley de Hooke, el muelle realiza una fuerza restauradora proporcional a el estiramiento que se ha producido en el muelle. Por tanto, si en equilibrio el muelle mide $l + \triangle l$ como consecuencia del peso mg de la masa puntual, entonces

$$k \triangle l = mg$$

donde $k > 0$ es la constante del muelle que mide su rigidez. Fijemos un sistema de referencia formado por una recta vertical y sobre la que se producirá el movimiento de nuestra masa, con origen en la mencionada posición de equilibrio. El sentido creciente de los valores de y es el de alejamiento del soporte o techo. Sea $y(t)$ la función que nos proporciona la posición de la masa respecto del sistema de referencia en el instante t. Suponemos además que como consecuencia de producirse el movimiento en un medio viscoso (aire, agua, etc.), se produce una fuerza de rozamiento de magnitud proporcional a la velocidad, es decir, $F_r = -\mu y'(t)$, con $\mu > 0$. Por la segunda ley de Newton, la fuerza total, $my''(t)$, coincide con la suma de todas las fuerzas que actúan sobre nuestra masa, de tal forma que se cumple

$$my''(t) = -\mu y'(t) - k(y(t) + \triangle l) + mg,$$

y teniendo en cuenta la condición de equilibrio llegamos a

$$y''(t) = -\frac{\mu}{m} y'(t) - \frac{k}{m} y(t).$$

Realizemos la reducción estandar del orden de la ecuación diferencial, que consiste en introducir dos nuevas variables: $y_1 = y$, $y_2 = y'$. Entonces la ecuación se escribe

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}' = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{\mu}{m} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

que tiene la estructura anunciada. Recuérdese que el conjunto de soluciones del sistema diferencial lineal anterior es un espacio vectorial de dimensión 2. Los valores propios de la matriz de coeficientes son

$$\frac{-\mu \pm \sqrt{\mu^2 - 4km}}{2m}.$$

Si $\mu^2 - 4km > 0$, entonces tenemos dos valores propios negativos λ_1 y λ_2, de tal forma que la solución general es

$$y(t) = \alpha e^{\lambda_1 t} + \beta e^{\lambda_2 t} \ \ (\alpha, \beta \in \mathbb{R}),$$

es decir, las soluciones tienden monótonamente a cero cuando $t \to +\infty$.

Si $\mu^2 - 4km < 0$, entonces tenemos valores propios complejos conjugados λ_1 y λ_2, con $\mathrm{Re}(\lambda_1) = -\mu/(2m) < 0$, de tal forma que

$$y(t) = e^{\mathrm{Re}\,\lambda_1\, t}(\alpha \cos(\mathrm{Im}\,\lambda_1\, t) + \beta \sin(\mathrm{Im}\,\lambda_1\, t)) \ \ (\alpha, \beta \in \mathbb{R}),$$

esto es, las soluciones oscilan mientras tienden a cero cuando $t \to +\infty$.

Por último, si $\mu^2 - 4km = 0$, tenemos un único valor propio negativo $\lambda = -\mu/(2m)$ y, por tanto,

$$y(t) = (\alpha + \beta t)e^{\lambda\, t} \ \ (\alpha, \beta \in \mathbb{R}),$$

que son funciones que también tienden a cero cuando $t \to +\infty$.

Capítulo 2

Complementos sobre matrices

2.1. Normas vectoriales y normas matriciales

En adelante K representará el cuerpo de los números reales $\mathbb{R}$ o el de los números complejos $\mathbb{C}$. Usaremos la notación de MATLAB, de tal foma que los vectores de K^n se representarán como $x = [x(1); x(2); \ldots; x(n)]$, donde $x(i) \in K$, $1 \leq i \leq n$, es decir, como columnas

$$\begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ x(n) \end{bmatrix}.$$

También podemos hablar de vector fila $x' = [x(1), x(2), \ldots, x(n)]$, transpuesto del vector columna x.

Dados $x, y \in K^n$ y $\alpha \in K$ podemos realizar las siguientes operaciones habituales de suma de vectores

$$x + y = [x(1) + y(1); x(2) + y(2); \ldots; x(n) + y(n)],$$

y de producto por un escalar

$$\alpha x = [\alpha x(1); \alpha x(2); \ldots; \alpha x(n)],$$

que confieren a K^n estructura de K espacio vectorial. Denotaremos la base canónica como $\{e_i\}_{i=1}^n$, $e_i(j) = \delta_{ij}$, $1 \leq i, j \leq n$, donde $\delta_{i,j}$ representa la función delta de Kronecker que vale uno si $i = j$ y cero en cualquier otro caso.

Disponemos también de un producto entre un vectores fila y un vector columna

$$x'y = \sum_{i=1}^{n} x(i)y(i),$$

llamado usualmente producto escalar.

Las matrices A de m filas por n columnas sobre K las referenciaremos escribiendo $A \in K^{m\times n}$, y las identificaremos implícitamente con los vectores de K^{mn}, (esto se puede conseguir por ejemplo entendiendo la matriz como el vector columna obtenido al colocar una columna de la matriz tras otra), si bien las representaremos indistintamente en la forma

$$A = \begin{bmatrix} A(1,1) & A(1,2) & \dots & A(1,n) \\ A(2,1) & A(2,2) & \dots & A(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ A(m,1) & A(m,2) & \dots & A(m,n) \end{bmatrix},$$

donde $A(i,j) \in K$, $1 \le i \le m$, $1 \le j \le n$, o indicando columnas

$$A = [A(:,1), A(:,2), \dots, A(:,n)],$$

donde $A(:,j) = [A(1,j); A(2,j); \dots; A(m,j)]$, $1 \le j \le n$, o bien destacando las filas

$$A = [A(1,:); A(2,:); \dots; A(m,:)],$$

donde $A(i,:) = [A(i,1), A(i,2), \dots, A(i,n)]$, $1 \le i \le m$.

Por supuesto, los vectores de K^n se pueden interpretar como matrices de tamaño $n \times 1$ y los vectores fila de n componentes como matrices de tamaño $1 \times n$. En lo que sigue, haremos implícitamente uso de las identificaciones mencionadas entre vectores y matrices.

Como se recordó anteriormente, $K^{m\times n}$ con las operaciones de suma de matrices y de producto por un escalar habituales es un K espacio vectorial de dimensión mn.

Se define el producto de una matriz $A \in K^{m\times n}$ por un vector $x \in K^n$ como un nuevo vector $Ax \in K^m$ determinado por

$$(Ax)(i) = A(i,:)x \quad (1 \le i \le m).$$

Nótese que

$$Ax = \sum_{j=1}^{n} x(j)A(:,j),$$

es decir, Ax se puede interpretar como un vector combinación lineal de las columnas de A.

De esta forma disponemos de una aplicación lineal $\mathcal{L}_A : K^n \to K^m$; $\mathcal{L}_A x = Ax$. Es conocido que toda aplicación lineal $\mathcal{L} : K^n \to K^m$ viene determinada por una matriz en la forma anterior, como multiplicación por una adecuada matriz. Obsérvese que A está determinada por cómo actúa sobre una base de K^n, como por ejemplo la canónica; de hecho, $Ae_j = A(:,j)$, $1 \le j \le n$.

Disponemos además del producto de matrices, que extiende el mencionado producto de vectores, y que en términos de las aplicaciones lineales que

representan, corresponde a la operación composición de aplicaciones: sean $A \in K^{m\times n}$ y $B \in K^{n\times s}$, se define $AB \in K^{m\times s}$ como

$$(AB)(i,j) = A(i,:)B(:,j), \quad 1 \le i \le m,\ 1 \le j \le s,$$

es decir, el producto de la fila i-ésima de A por la columna j-ésima de B nos da la entrada (i,j) de la matriz producto. Recuérdese que esta operación no es en general conmutativa, ni tan siquiera para matrices cuadradas.

Obsérvese que

$$AB = [AB(:,1), AB(:,2), \ldots, AB(:,s)]\,,$$

y que también

$$AB = \sum_{k=1}^{n} A(:,k)B(k,:),$$

es decir, se puede escribir como suma de matrices en las que todas las columnas son multiplos de un mismo vector.

Definición 2.1. *Una norma vectorial sobre K^n es una aplicación $\|\| : K^n \to \mathbb{R}$ cumpliendo:*

(i) $\|x\| \ge 0$, $\forall x \in K^n$, *y* $\|x\| = 0$ *si, y sólo si,* $x = 0$,

(ii) $\|x + y\| \le \|x\| + \|y\|$, $\forall x, y \in K^n$,

(iii) $\|\alpha x\| = |\alpha|\,\|x\|$, $\forall x \in K^n$, $\alpha \in K$.

Nótese que las normas sobre $\mathbb{C}^n$ pueden ser consideradas normas sobre $\mathbb{R}^n$ por simple restricción.

Las clases más importantes de normas vectoriales son las llamadas p-normas

$$\|x\|_p = (\sum_{i=1}^{n} |x(i)|^p)^{1/p} \quad (p \ge 1),$$

en especial los casos $p = 1, 2$, y la llamada norma infinito ($p = \infty$)

$$\|x\|_\infty = \max_{1\le i\le n} (|x(i)|).$$

Obsérvese que la p-norma sobre K^ny la correspondiente p-norma sobre K^s, $s \ne n$, son diferentes.

Ejercicio 2.1. *Compruébese que para toda norma $\|\|$ sobre K^n se cumple*

$$\|x\| \le cte.\, \|x\|_\infty \quad (x \in K^n).$$

Ejercicio 2.2. *Para toda norma $\|\|$ sobre K^n se cumple*

$$|\|x\| - \|y\|| \le \|x - y\| \quad (x, y \in K^n),$$

lo que en particular establece la continuidad de la norma.

Recuérdese que todas las normas sobre K^n son equivalentes, queriéndose indicar con ello que fijadas dos normas $\|\|\|_*$ y $\|\|\|_\circ$ sobre K^n, existen constantes $c_1, c_2 > 0$ tales que

$$c_1 \|x\|_* \leq \|x\|_\circ \leq c_2 \|x\|_* \quad (x \in K^n).$$

Esta propiedad fundamental, no válida en espacios normados de dimensión infinita, permite hablar de una única convergencia de sucesiones, ya que para cualquier norma la convergencia es la convergencia componente a componente, es decir, si tenemos la sucesión de vectores $\{x_k\}_{k=1}^{\infty}$ de K^n y $x \in K^n$, entonces

$$\lim x_k = x \longleftrightarrow \lim x_k(i) = x(i) \ \ (1 \leq i \leq n).$$

Ejercicio 2.3. *Pruébense la siguientes relaciones que muestran la equivalencia de las $p-$normas para $p = 1, 2, \infty$:*

(i) $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2, \quad \forall x \in K^n.$

(ii) $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty, \quad \forall x \in K^n.$

Nótese que la equivalencia de normas es transitiva, y que por tanto, de las anteriores relaciones se deduce la equivalencia entre $\|\|\|_1$ y $\|\|\|_\infty$.

El análisis de los problemas y algoritmos que vamos a ir introduciendo en los próximos capítulos, requiere con frecuencia de disponer de alguna forma de medir o cuantificar la proximidad entre matrices. Este papel lo jugarán las normas matriciales.

Definición 2.2. *Una norma matricial sobre $K^{m\times n}$ es cualquier norma vectorial sobre K^{mn}.*

Evidentemente todas las normas matriciales son equivalentes.

La norma $\|\|\|_2$ sobre K^{mn} es una norma matricial que recibe habitualmente el nombre de norma de Frobenius:

$$\begin{aligned} \|A\|_F &:= \|[A(:,1); A(:,2); \ldots; A(:,n)]\|_2 \\ &= \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |A(i,j)|^2 \right)^{1/2} \end{aligned}$$

mientras que en el caso de $\|\|\|_\infty$ se usa la notación

$$\|A\|_{\text{máx}} := \|[A(:,1); A(:,2); \ldots; A(:,n)]\|_\infty = \max_{1\leq i\leq m, 1\leq j\leq n} |A(i,j)|\,,$$

con el objetivo de facilitar su distinción frente a otras normas matriciales que introduciremos a continuación asociadas, en un sentido alternativo, a las normas vectoriales.

Como iremos viendo, cuando tratamos con matrices hay normas que, por sus propiedades adicionales, son más útiles que otras. Éstas están definidas en base a considerar la matriz como una aplicación que actúa entre espacios normados.

Teorema 2.1. *Sean* $\|\|\|_{\widehat{n}}$ *y* $\|\|\|_{\widehat{m}}$ *normas vectoriales sobre* K^n *y* K^m *respectivamente. Para toda* $A \in K^{m\times n}$ *el conjunto*

$$\{\|Ax\|_{\widehat{m}} : \|x\|_{\widehat{n}} = 1\},$$

está acotado superiormente. La aplicación $\|\|\|_{\widehat{m}\widehat{n}} : K^{m\times n} \to \mathbb{R}$ *determinada por*

$$\|A\|_{\widehat{m}\widehat{n}} = \sup\{\|Ax\|_{\widehat{m}} : \|x\|_{\widehat{n}} = 1\},$$

define un norma matricial sobre $K^{m\times n}$*, que llamaremos norma matricial subordinada a las normas prefijadas. Además,*

$$\begin{aligned} \|A\|_{\widehat{m}\widehat{n}} &= \text{máx}\{\|Ax\|_{\widehat{m}} : \|x\|_{\widehat{n}} = 1\} \\ &= \text{mín}\{\alpha \geq 0 : \|Ax\|_{\widehat{m}} \leq \alpha \|x\|_{\widehat{n}}\,,\ x \in K^n\}. \end{aligned}$$

Demostración. Definimos la aplicación $h : \{x \in K^n : \|x\|_{\widehat{n}} = 1\} \to R$; $h(x) = \|Ax\|_{\widehat{m}}$. Es una aplicación continua ya que es composición de la norma con una aplicación lineal. Al estar definida sobre un conjunto cerrado y acotado, y por lo tanto compacto en K^n, alcanza sus valores extremos, es decir, existe el supremo del conjunto y de hecho es un máximo como se anuncia al final del enunciado. Por lo tanto, la aplicación $\|\|\|_{\widehat{m}\widehat{n}}$ está bien definida. Veamos que cumple la propiedad (i) de toda norma. Si $\|A\|_{\widehat{m}\widehat{n}} = 0$, entonces en particular

$$0 = \|e_j\|_{\widehat{n}} \left\|Ae_j/\|e_j\|_{\widehat{n}}\right\|_{\widehat{m}} = \|Ae_j\|_{\widehat{m}} = \|A(:,j)\|_{\widehat{m}}\,, \quad 1 \leq j \leq n,$$

de donde se deduce que todas las columnas de la matriz son el vector nulo. Las propiedades (ii)-(iii) de la norma son consecuencia inmediata de las propiedades del supremo de un conjunto de números reales.

Finalmente, si $\alpha \geq 0$ cumple $\|Ax\|_{\widehat{m}} \leq \alpha \|x\|_{\widehat{n}}$, $\forall x \in K^n$, entonces

$$\|Ax\|_{\widehat{m}} \leq \alpha,\ \forall x \in K^n,\ \|x\|_{\widehat{n}} = 1,$$

lo que implica que $\|A\|_{\widehat{m}\widehat{n}} \leq \alpha$, probándose que es el menor de lo números con dicha propiedad, ya que si $y \in K^n$ no es el vector nulo, entonces

$$\|A(y/\|y\|_{\widehat{n}})\|_{\widehat{m}} \leq \|A\|_{\widehat{m}\widehat{n}}\,,$$

es decir,

$$\|Ay\|_{\widehat{m}} \leq \|A\|_{\widehat{m}\widehat{n}} \|y\|_{\widehat{n}}\,, \quad \forall y \in K^n.$$

□

Ejercicio 2.4. *Pruébese la primera afirmación del teorema anterior mediante la mencionada equivalencia de normas. (*$\|Ax\|_\infty \leq \text{máx}_{1\leq i\leq m}(\sum_{i=1}^n |A(i,j)|\,|x(j)|) \leq n\,\|A\|_{\text{máx}}\,\|x\|_\infty$*, por ejemplo)*

Proposición 2.1. *Sean* $\|\|\|_{\widehat{n}}$, $\|\|\|_{\widehat{m}}$ *y* $\|\|\|_{\widehat{s}}$ *normas vectoriales sobre* K^n, K^m *y* K^s *respectivamente. Sean* $\|\|\|_{\widehat{m}\widehat{n}}$, $\|\|\|_{\widehat{n}\widehat{s}}$ *y* $\|\|\|_{\widehat{m}\widehat{s}}$ *las correspondientes normas matriciales subordinadas sobre* $K^{m\times n}$, $K^{n\times s}$ *y* $K^{m\times s}$ *respectivamente. Sean* $A \in K^{m\times n}$ *y* $B \in K^{n\times s}$. *Entonces*

$$\|AB\|_{\widehat{m}\widehat{s}} \leq \|A\|_{\widehat{m}\widehat{n}} \|B\|_{\widehat{n}\widehat{s}} \quad \textit{(Multiplicatividad).}$$

Además, si I_n *es la matriz identidad, determinada por* $I_n(i,j) = \delta_{ij}$, $1 \leq i,j \leq n$, *entonces*

$$\|I_n\|_{\widehat{n}\widehat{n}} = 1.$$

Demostración. Por el teorema anterior tenemos

$$\|ABx\|_{\widehat{m}} \leq \|A\|_{\widehat{m}\widehat{n}} \|Bx\|_{\widehat{n}} \leq \|A\|_{\widehat{m}\widehat{n}} \|B\|_{\widehat{n}\widehat{s}} \|x\|_{\widehat{s}},$$

que conduce sin dificultad a la desigualdad anunciada al tomar supremos.

La última afirmación es consecuencia inmediata de la definición de norma subordinada. □

La norma de Frobenius no es una norma subordinada ya que $\|I_n\|_F = \sqrt{n}$, y la norma máximo tampoco, ya que no es multiplicativa:

$$2 = \left\| \begin{bmatrix} 2 & 0 \\ 1 & 0 \end{bmatrix} \right\|_{\text{máx}} > \left\| \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \right\|_{\text{máx}} \left\| \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \right\|_{\text{máx}} = 1.$$

No obstante, la norma de Frobenius sí es multiplicativa: sean $A \in K^{m\times n}$ y $B \in K^{n\times s}$, entonces

$$\begin{aligned} \|AB\|_F^2 &= \sum_{i=1}^{m}\sum_{j=1}^{s} |A(i,:)B(:,j)|^2 \leq \sum_{i=1}^{m}\sum_{j=1}^{s} (\|A(i,:)\|_2^2 \|B(:,j)\|_2^2) \\ &= \sum_{i=1}^{m} \|A(i,:)\|_2^2 \sum_{j=1}^{s} \|B(:,j)\|_2^2 = \|A\|_F^2 \|B\|_F^2, \end{aligned}$$

habiendo utilizado en la primera desigualdad la llamada desigualdad de Cauchy-Schwarz

$$\left|\overline{x}' y\right| \leq \|x\|_2 \|y\|_2 \quad (x, y \in K^n).$$

Cuando tenemos una familia de normas sobre K^s, $s \in \mathbb{N}$, determinadas por el mismo patrón, que denotamos sin distinción como $\|\|$, entonces hablamos de la norma subordinada correspondiente utilizando la misma notación, entendiéndose

$$\|A\| = \max_{\|x\|=1} \|Ax\| \quad (A \in K^{m\times n}).$$

Entendidas en este sentido, las normas matriciales que utilizaremos en adelante son las normas subordinadas a las p-normas, para $p = 1, 2, \infty$. Así, para $A \in K^{m\times n}$,

$$\|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p,$$

y se cumplen

$$\|Ax\|_p \leq \|A\|_p \|x\|_p \quad (x \in K^n),$$

y para $B \in K^{n\times s}$,

$$\|AB\|_p \leq \|A\|_p \|B\|_p .$$

Nota 2.1. *Notar que por la multiplicatividad de la norma de Frobenius se cumple*

$$\|Ax\|_2 \leq \|A\|_F \|x\|_2 ,$$

pero es en realidad una cota demasiado burda comparada con la de la norma $\|\|_2$*, es decir,* $\|A\|_2 \leq \|A\|_F$*.*

Ejercicio 2.5. *Sea* $D \in K^{m\times n}$ *diagonal (*$D(i,j) = 0$ *si* $i \neq j$*). Entonces*

$$\|D\|_p = \|diag(D)\|_\infty \quad (1 \leq p \leq \infty).$$

($\|Dx\|_p^p = \sum_{i=1}^m \left|\sum_{j=1}^n D(i,j)x(j)\right|^p = \sum_{i=1}^s |D(i,i)x(i)|^p \leq \|diag(D)\|_\infty^p \|x\|_p^p$*,* $s = \min(m,n)$*. Para la otra desigualdad considerar* e_j*,* $1 \leq j \leq s$*.)*

Ejercicio 2.6. *Sea* $x \in \mathbb{C}^n$*. Compruébese* $\|x\|_p$ *calculada como normal vectorial, coincide con el valor de la norma matricial subordinada* $\|\|_p$ *de la matriz* $x \in \mathbb{C}^{n\times 1}$*. (Para vectores filas la propiedad no es cierta ya que se cumple* $\|A\|_p = \|A^*\|_q$*,* $1/p + 1/q = 1$*)*

Nótese que las normas matriciales $\|\|_p$ tienen sentido tanto sobre $\mathbb{R}$ como sobre $\mathbb{C}$ y que en principio podían tomar valores diferentes sobre una matriz real; es decir, si $A \in \mathbb{R}^{m\times n}$, entonces

$$\max_{x\in\mathbb{R}^n, \|x\|_p=1} \|Ax\|_p \leq \max_{x\in\mathbb{C}^n, \|x\|_p=1} \|Ax\|_p ,$$

aunque como se verá a continuación para $p = 1, \infty$, no hay tal distinción.

Proposición 2.2. *Sea* $A \in K^{m\times n}$*. Entonces*

(i) $\|A\|_1 = \max_{1\leq j\leq n} \|A(:,j)\|_1$,

(ii) $\|A\|_\infty = \max_{1\leq i\leq m} \|A(i,:)\|_1$.

Demostración. (i) Como $\|A(:,j)\|_1 = \|Ae_j\|_1 \leq \|A\|_1 \|e_j\|_1 = \|A\|_1$, tenemos probada la desigualdad $\geq$. Por otra parte,

$$\begin{aligned} \|Ax\|_1 &= \left\|\sum_{j=1}^n x(j)A(:,j)\right\|_1 \leq \sum_{j=1}^n |x(j)| \, \|A(:,j)\|_1 \\ &\leq \max_{1\leq j\leq n} \|A(:,j)\|_1 \|x\|_1 . \end{aligned}$$

(ii) Sea $\gamma = \max_{1\le i\le m} \|A(i,:)\|_1$. Puesto que $\|Ax\|_\infty = \max_{1\le i\le m} \left|\sum_{j=1}^n A(i,j)x(j)\right| \le \|x\|_\infty \gamma$ entonces $\|A\|_\infty \le \gamma$. Por otra parte, si $\gamma = \|A(i_0,:)\|_1$, consideramos el vector $x \in K^n$ con, para $1 \le j \le n$,

$$x(j) = \begin{cases} \frac{\overline{A(i_0,j)}}{|A(i_0,j)|}, \text{ si } A(i_0,j) \neq 0, \\ 1, \text{ si } A(i_0,j) = 0. \end{cases}$$

Entonces $\|x\|_\infty = 1$, y

$$\|A\|_\infty \ge \|Ax\|_\infty \ge |A(i_0,:)x| = \|A(i_0,:)\|_1 = \gamma.$$

□

Nota 2.2. *El anterior resultado pone de manifiesto como conseguir para ambas normas un vector unitario, es decir, con norma* 1*, para el que se alcanza la norma de la matriz. Veamos un ejemplo de ello. Sea*

$$A = \begin{bmatrix} i/2 & 1+i \\ 0 & -i \end{bmatrix} \in \mathbb{C}^{2\times 2}.$$

Entonces, $\|A\|_1 = \max(1/2, 1+\sqrt{2}) = 1+\sqrt{2} = \|Ae_2\|_1$*, luego podríamos tomar como vector unitario donde se alcanza la norma* e_2*. Además,* $\|A\|_\infty = \max(1/2+\sqrt{2}, 1) = 1/2+\sqrt{2}$*. Si tomamos* $x_* = [-i; (1-i)/\sqrt{2}]$ *se cumple* $\|x_*\|_\infty = 1$ *y*

$$\begin{aligned} \|Ax_*\|_\infty &= \left\|[1/2+\sqrt{2}; -i(1-i)/\sqrt{2}]\right\|_\infty \\ &= \max(1/2+\sqrt{2}, 1) = 1/2+\sqrt{2} = \|A\|_\infty . \end{aligned}$$

2.2. Algunas clases de matrices

Definición 2.3. *Una matriz* $A \in \mathbb{R}^{n\times n}$ *se dice que es regular o invertible si existe* $B \in \mathbb{R}^{n\times n}$ *tal que*

$$B\,A = I_n. \tag{2.1}$$

Proposición 2.3. *Sea* $A \in \mathbb{R}^{n\times n}$*. Equivalen:*

(i) *A es regular,*

(ii) *Si* $Ax = 0$*, entonces* $x = 0$*,*

(iii) *Para todo* $b \in \mathbb{R}^n$ *existe un* $x \in \mathbb{R}^n$ *tal que* $Ax = b$*.*

Demostración. (i)⇒(ii) Sea $B \in \mathbb{R}^{n\times n}$ tal que $B\,A = I_n$. Si $Ax = 0$, entonces

$$x = I_n x = BAx = B0 = 0.$$

(ii)⇔(iii) La primera propiedad expresa que el conjunto de n elementos formado por las columnas de A es linealmente independiente y la segunda que es sistema generador, propiedades que sabemos son equivalentes.

(iii)⇒(i) Para cada $j = 1, 2, ..., n$, sea v_j verificando $Av_j = e_j$. Consideramos la matriz

$$C = [v_1\, v_2\, ...\, v_n] \in \mathbb{R}^{n\times n}.$$

Como $(AC)(:, j) = AC(:, j)$, es ya sencillo convencerse de que

$$AC = I_n,$$

es decir, C es regular. En consecuencia, C verifica (iii), por lo que existe $D \in \mathbb{R}^{n\times n}$ tal que

$$CD = I_n.$$

Así,

$$A = AI_n = ACD = I_nD = D,$$

de donde,

$$CA = I_n,$$

lo que establece la regularidad de A. □

Nota 2.3. *El resultado anterior expresa que los endomorfismos sobre espacios vectoriales de dimensión finita son inyectivos si y sólo si son suprayectivos. Esta propiedad no es en general cierta para espacios de dimensión infinita. La alternativa de Fredholm establece dicha propiedad para los operadores de la forma $I - A$ con A un operador compacto.*

Observar que de los argumentos anteriores se deduce que A es regular si y sólo si existe $B \in \mathbb{R}^{n\times n}$ tal que

$$B\,A = AB = I_n.$$

Recordar que en general el producto de matrices no es conmutativo.

Si $C \in \mathbb{R}^{n\times n}$ cumple

$$C\,A = AC = I_n,$$

entonces

$$C = CI_n = CAB = I_nB = B$$

por lo que en definitiva podemos establecer que una matriz B verificando (2.1), si existe, es única y se le llama matriz inversa de A. Usaremos la notación habitual $B = A^{-1}$.

Ejercicio 2.7 (Normas con peso)**.** *Sea $A \in K^{n\times n}$ regular y sea $\|\|\|$ una norma sobre K^n. Pruébese que la aplicación*

$$\|x\|_A = \|Ax\|\,,\ x \in K^n,$$

es una norma sobre K^n.

Definición 2.4. *La matriz transpuesta de $A \in \mathbb{R}^{m\times n}$ es un elemento de $\mathbb{R}^{n\times m}$, que denotaremos mediante A', y que está determinada por*

$$A'(i, j) = A(j, i) \quad 1 \le i \le n,\ 1 \le j \le m.$$

Proposición 2.4. *Sean $A, B \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$ y $\alpha, \beta \in \mathbb{R}$. Se cumplen:*

(i) $x'x = \|x\|_2^2$,

(ii) $A'(:,i) = A(i,:)', \quad 1 \le i \le n$,

(iii) $(A')' = A$,

(iv) $(\alpha A + \beta B)' = \alpha A' + \beta B'$,

(v) $(AB)' = B'A'$,

(vi) *A es regular si y sólo si A' es regular,*

(vii) $\|A\|_\infty = \|A'\|_1$.

Definición 2.5. *Una matriz es simétrica si coincide con su transpuesta.*

Definición 2.6. *Una matriz $A \in \mathbb{R}^{n\times n}$ diremos que es (estrictamente) diagonalmente dominante por filas si*

$$|A(i,i)| \ge (>) \sum_{j=1, j\neq i}^{n} |A(i,j)|, \quad 1 \le i \le n,$$

y que es (estrictamente) diagonalmente dominante por columnas si A' cumple la correspondiente propiedad por filas.

Proposición 2.5. *Las matrices estrictamente diagonalmente dominantes son regulares.*

Demostración. Bastará que probemos el resultado para $A \in \mathbb{R}^{n\times n}$ estrictamente diagonalmente dominante por filas, ya que el caso de columnas es consecuencia inmediata de éste por transposición. Si $Ax = 0$, entonces

$$|A(i,i)|\,|x(i)| \le \sum_{j=1, j\neq i}^{n} |A(i,j)|\,|x(j)|, \quad 1 \le i \le n.$$

Si $|x(i_0)| = \|x\|_\infty$, entonces de la anterior desigualdad para i_0 y tras simplificar $\|x\|_\infty > 0$ se obtiene

$$|A(i_0,i_0)| \le \sum_{j=1, j\neq i_0}^{n} |A(i_0,j)|,$$

lo que contradice la hipótesis sobre A. □

Nota 2.4. *La anterior propiedad no es cierta si la dominancia diagonal no es estricta en todas las filas. Basta tomar una matriz con una fila entera de ceros.*

Definición 2.7. *Sean $A \in \mathbb{R}^{n\times n}$ y $k \in \mathbb{Z} \cap [-n+1, n-1]$. La diagonal k-ésima de A es el vector columna*

$$diag(A,k) = \begin{cases} [A(1,k+1),\, A(2,k+2), \cdots,\, A(n-k,n)]', & si\ k \geq 0 \\ [A(1-k,1),\, A(2-k,2), \cdots,\, A(n,n+k)]', & si\ k < 0 \end{cases}$$

Las diagonales correspondientes a $k > 0$ las llamaremos diagonales positivas, y diagonales negativas las correspondientes a $k < 0$. La diagonal $k = 0$ se le llama diagonal principal. Una diagonal es nula si todas sus componentes son nulas.

Una matriz es triangular superior si todas sus diagonales negativas son nulas, diagonal inferior si son nulas todas las diagonales positivas, y diagonal si es triangular inferior y superior.

Una matriz banda $A \in \mathbb{R}^{n\times n}$ con ancho de banda $[p,q]$ $(0 \leq p, q \leq n-1)$ es aquella que cumple que las diagonales $diag(A,r)$, $-n+1 \leq r \leq -q-1$, $p+1 \leq r \leq n-1$, son todas nulas. Así, las matrices triangulares superiores son matrices banda con ancho de banda $[n-1,0]$, las triangulares inferiores matrices banda con ancho de banda $[0,n-1]$, y las matrices diagonales matrices banda con ancho de banda $[0,0]$. El parámetro p representa el ancho de la banda positiva y q el ancho de la banda negativa.

Las matrices de la clase Hessenberg superior (inferior) son las matrices banda con ancho de banda $[n-1,1]$ $([1,n-1])$.

Una matriz es tridiagonal si es una matriz banda con ancho de banda $[1,1]$.

Nota 2.5. *Toda matriz se puede transformar mediante transformaciones ortogonales de semejanza (por ejemplo utilizando matrices de Householder; al ser semejantes se conservan los valores propios, no así los vectores propios, pero están relacionados) en una matriz Hessenberg superior. Si la matriz de partida es simétrica, la matriz de Hessenberg superior obtenida es también simétrica y por tanto, es tridiagonal y simétrica. En el tema de aproximación de valores propios probaremos el resultado mencionado e introduciremos técnicas especificas para aproximar los valores propios de una matriz de Hessenberg superior.*

Ejercicio 2.8. *Probar que las matrices triangulares son regulares si y sólo si $A(i,i) \neq 0$, $1 \leq i \leq n$.*

Nota 2.6. *Recordemos que las permutaciones de $\{1,...,n\}$ son las aplicaciones biyectivas $\sigma : \{1,...,n\} \to \{1,...,n\}$. Lo indicaremos escribiendo $\sigma \in \mathcal{S}_n$ y $\sigma \equiv [\sigma(1)\,\sigma(2)\cdots\sigma(n)]$, es decir, identificamos la permutación mediante la imagen escrita como un vector fila.*

Definición 2.8. *Sea $\sigma \in \mathcal{S}_n$. Llamaremos matriz permutación, asociada a σ, a la matriz*

$$P_\sigma = [e_{\sigma(1)}\, e_{\sigma(2)}\, \ldots\, e_{\sigma(n)}] \in \mathbb{R}^{n\times n},$$

donde $\{e_j\}_{j=1}^n$ representa la base canónica de $\mathbb{R}^n$.

En el análisis matricial del método de eliminación de Gauss usaremos un caso particular de estas matrices, las matrices asociadas a transposiciones, es decir, a permutaciones que cambian dos elementos y dejan los demás invariantes.

Proposición 2.6. *Sean $A \in \mathbb{R}^{n\times n}$ y $\sigma, \gamma \in \mathcal{S}_n$. Entonces*

(i) $(AP_\sigma)(:,j) = A(:,\sigma(j)), \quad 1 \le j \le n,$

(ii) $P_\sigma P_\gamma = P_{\sigma\circ\gamma},$

(iii) $(P'_\sigma A)(i,:) = A(\sigma(i),:), \quad 1 \le i \le n,$

(iv) $P_{\sigma^{-1}} = P_\sigma^{-1} = P'_\sigma.$

Demostración. (i) $(AP_\sigma)(:,j) = AP_\sigma(:,j) = Ae_{\sigma(j)} = A(:,\sigma(j))$.
(ii) Consecuencia de la anterior.
(iii) $[(P'_\sigma A)(i,:)]' = (P'_\sigma A)'(:,i) = (A'P_\sigma)(:,i) = A'(:,\sigma(i)) = A(\sigma(i),:)$.
(iv) La primera igualdad es consecuencia de (ii). Para la segunda,

$$(P'_\sigma P_\sigma)(i,j) = P_\sigma(\sigma(i),j) = e_{\sigma(j)}(\sigma(i)) = \delta_{i,j}, \quad 1 \le i,j \le n,$$

es decir, $P'_\sigma P_\sigma = I_n$. □

Ejercicio 2.9. *Sea $D \in \mathbb{R}^{n\times n}$ diagonal. Estúdiese como reordenar en orden decreciente los elementos de la diagonal principal de D mediante el producto por matrices permutación. (Sea $\sigma \in S_n$ tal que $D(\sigma(1),\sigma(1)) \ge D(\sigma(2),\sigma(2)) \ge \ldots \ge D(\sigma(n),\sigma(n))$. Entonces $(P'_\sigma D P_\sigma)(i,j) = D(\sigma(i),\sigma(j))$).*

Definición 2.9. *Una matriz es definida (semidefinida) positiva si para todo $x \in \mathbb{R}^n$, no nulo, se cumple*

$$x'Ax > (\ge) 0.$$

En el tema de introducción ya se mostro que las matrices definidas positivas son regulares, lo que no se puede trasladar al caso de las matrices semidefinidas positivas (basta con considerar la matriz identicamente nula).

Ejercicio 2.10. *La matriz $B \in \mathbb{R}^{n\times n}$ es definida positiva si, y sólo si, B' es definida positiva.*

Proposición 2.7. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica y semidefinida positiva. Entonces*

(i) $A(i,i) \ge 0, \quad 1 \le i \le n,$

(ii) $|A(i,j)| \le (A(i,i) + A(j,j))/2, \quad 1 \le i,j \le n,$

(iii) $\|A\|_{\text{máx}} = \|diag(A)\|_\infty$.

(iv) *Las matrices $A_k := A(1:k, 1:k)$, $1 \le k \le n$, son simétricas semidefinidas positivas.*

Demostración. Comencemos destacando que

$$e_i' A e_j = A(i,j), \quad 1 \leq i,j \leq n.$$

Ahora (i) es consecuencia de ser A semidefinida positiva.

Por otra parte, al ser A también simétrica tenemos

$$0 \leq (e_j + e_i)' A (e_j + e_i) = A(i,i) + A(j,j) + 2A(i,j)$$

y

$$0 \leq (e_j - e_i)' A (e_j - e_i) = A(i,i) + A(j,j) - 2A(i,j),$$

de donde se obtiene (ii) al despejar $A(i,j)$ de ambas desigualdades.

Por (ii) tenemos $|A(i,j)| \leq \|diag(A)\|_\infty$, $1 \leq i,j \leq n$, de donde

$$\|A\|_{\text{máx}} \leq \|diag(A)\|_\infty .$$

Por tanto (iii) queda probada ya que la otra desigualdad es trivial.

Por otra parte, es evidente que las matrices A_k son simétricas. Si consideramos $x \in \mathbb{R}^k$ entonces

$$x' A_k x = y' A y \geq 0,$$

donde $y = [x; zeros(n-k,1)]$. □

Nota 2.7. *Las desigualdades de la anterior proposición son estrictas si la matriz es definida positiva ($i \neq j$). En principio estas propiedades permitirían descartar que una matriz es semidefinida positiva.*

Para toda $A \in \mathbb{R}^{n\times n}$, la nueva matriz $A'A$ es simétrica semidefinida positiva. En efecto, es evidente que es simétrica, y además

$$x' A' A x = (Ax)' Ax = \|Ax\|_2^2 \geq 0.$$

De esta relación se deduce sin dificultad el siguiente resultado:

Proposición 2.8. *Sea $A \in \mathbb{R}^{n\times n}$. Equivalen:*

(i) *A es regular,*

(ii) *$A'A$ es definida positiva.*

(iii) *AA' es definida positiva.*

Proposición 2.9. *Sea $A, P \in \mathbb{R}^{n\times n}$ con A definida positiva y P regular. Entonces la matriz $P'AP$ es definida positiva.*

Demostración. La clave es la relación

$$x' P' A P x = (Px)' A P x$$

junto con que $Px = 0$ si y sólo si $x = 0$. □

Ejercicio 2.11. *¿Si $P'AP$ es definida positiva, para una matriz P regular, entonces A es definida positiva? (Sí; para x no nulo tenemos $x'Ax = y'P'APy > 0$, siendo $x = Py$).*

Definición 2.10. *Un conjunto $\{x_i\}_{i=1}^k \subset \mathbb{R}^n$ se dice que es ortonormal si*

$$x_i' x_j = \delta_{ij}, \quad 1 \leq i, j \leq k.$$

Notar que si un conjunto de n vectores de $\mathbb{R}^n$ es ortonormal, entonces es una base, pues son independientes. En efecto, si tenemos una combinación lineal

$$\sum_{i=1}^n \lambda_i x_i = 0,$$

entonces

$$0 = x_j'(\sum_{i=1}^n \lambda_i x_i) = \lambda_j, \quad 1 \leq j \leq n.$$

En el anterior razonamiento se refleja la propiedad interesante de las bases ortonormales: las coordenadas de un vector respecto de dicha base se obtienen con facilidad, de hecho

$$x = \sum_{i=1}^n (x_i' x)\, x_i, \quad x \in \mathbb{R}^n.$$

Obsérvese que en este caso

$$\|x\|_2^2 = x'x = \sum_{i=1}^n (x_i' x)^2.$$

Nota 2.8 (Proceso de ortogonalización de Gram-Schmidt)**.** *El método que recordamos permite, a partir de un sistema independiente de vectores, obtener otro que genera el mismo subespacio vectorial y que es un sistema ortonormal.*

Sea $\{x_i\}_{i=1}^k \subset \mathbb{R}^n$ un sistema linealmente independiente y $S = \left\langle \{x_i\}_{i=1}^k \right\rangle$. Probaremos el resultado por inducción sobre k. Supongamos que existen $\{y_i\}_{i=1}^{k-1} \subset \mathbb{R}^n$ sistema ortonormal tal que $\left\langle \{y_i\}_{i=1}^{k-1} \right\rangle = \left\langle \{x_i\}_{i=1}^{k-1} \right\rangle$. Construimos

$$v_k = x_k - \sum_{i=1}^{k-1} \lambda_i y_i,$$

donde los escalares de la combinación lineal, λ_i, se determinan exigiendo que v_k sea ortogonal a y_i, $1 \leq j \leq k-1$, es decir,

$$\lambda_i = y_i'\, x_k, \quad 1 \leq i \leq k-1.$$

Finalmente elegiríamos $y_k = v_k / \|v_k\|_2$. Es evidente que se cumpliría $\left\langle \{y_i\}_{i=1}^k \right\rangle \supseteq \left\langle \{x_i\}_{i=1}^k \right\rangle$ y serían iguales por tener la misma dimensión k.

La siguiente consecuencia nos será útil para probar que toda matriz simétrica es diagonalizable.

Proposición 2.10. *Sea $x \in \mathbb{R}^n$ con $\|x\|_2 = 1$. Existen vectores $\{x_i\}_{i=1}^{n-1}$ de $\mathbb{R}^n$ tales que $\{x\} \cup \{x_i\}_{i=1}^{n-1}$ es una base ortonormal de $\mathbb{R}^n$.*

Demostración. Basta con asegurarse que podemos completar $\{x\}$ a una base de $\mathbb{R}^n$ (si $x(i_0) \neq 0$, entonces $\{x\} \cup \{e_i : 1 \leq i \leq n,\, i \neq i_0\}$ es una base) y luego aplicar Gram-Schmidt comenzando con x. □

Definición 2.11. *Una matriz $A \in \mathbb{R}^{n\times n}$ es ortogonal si*

$$A'A = I_n.$$

Notar que la propiedad de ser A ortogonal se puede expresar de forma equivalente diciendo que

$$A(:,i)'A(:,j) = \delta_{ij}, \quad 1 \leq i,j \leq n,$$

es decir, que las columnas de A son un conjunto ortonormal de vectores. Evidentemente, A es ortogonal si y sólo si A' es ortogonal. Por tanto, lo que se diga sobre las filas de las matrices ortogonales es cierto por columnas y viceversa. El siguiente resultado es ya evidente.

Proposición 2.11. *Sea $A \in \mathbb{R}^{n\times n}$. Equivalen*

(i) *A es ortogonal,*

(ii) *$\{A(:,j)\}_{j=1}^n$ es una base ortonormal de $\mathbb{R}^n$.*

(iii) *$\{A(i,:)\}_{i=1}^n$ es una base ortonormal de $\mathbb{R}^n$.*

Corolario 2.1. *Sea $x \in \mathbb{R}^n$ con $\|x\|_2 = 1$. Existe $P \in \mathbb{R}^{n\times n}$, ortogonal, tal que $P(:,1) = x$.*

En la siguiente proposición se establece una de las propiedades más interesantes de las matrices ortogonales frente a la norma matricial subordinada a la norma euclídea.

Proposición 2.12. *Sea $A \in \mathbb{R}^{m\times n}$ y sean $P \in \mathbb{R}^{m\times m}$, $Q \in \mathbb{R}^{n\times n}$, ortogonales. Entonces*

(i) $\|P\|_2 = 1$,

(ii) $\|PAQ\|_2 = \|A\|_2$.

Demostración. Para (i) basta tener en cuenta que para todo $x \in \mathbb{R}^n$ se tiene

$$\|Px\|_2^2 = (Px)'Px = x'P'Px = x'x = \|x\|_2^2 .$$

Para (ii) por la multiplicatividad tenemos

$$\|PAQ\|_2 \leq \|P\|_2 \|A\|_2 \|Q\|_2 = \|A\|_2 .$$

Por otra parte, por la desigualdad probada,

$$\|A\|_2 = \left\|P'PAQQ'\right\|_2 \leq \|PAQ\|_2 ,$$

lo que acaba de probar (ii). □

Nota 2.9. *Si $P \in \mathbb{R}^{m\times n}$ verifica $P'P = I_n$, entonces $\|P\|_2 = 1$.*

2.3. Valores y vectores propios de una matriz

Comencemos recordando el concepto de determinante de una matriz $A \in \mathbb{R}^{n\times n}$, o de un conjunto de n vectores $\mathbb{R}^n$, y algunas de sus propiedades básicas.

Definición 2.12. *Una aplicación*

$$\begin{array}{rcl} \det & : & \mathbb{R}^{n\times n} \equiv \mathbb{R}^n \times \mathbb{R}^n \times \ldots \times \mathbb{R}^n \to \mathbb{R} \\ A & = & [A(:,1)\, A(:,2) \ldots A(:,n)] \longmapsto \det(A) \end{array}$$

es un determinante sobre las matrices $\mathbb{R}^{n\times n}$ si cumple:

(i) det *es lineal respecto de las columnas de A,*

(ii) *si $A(:,i) = A(:\,j)$ para $i \neq j$, $1 \leq i,j \leq n$, entonces* $\det(A) = 0$, *y*

(iii) $\det(I_n) = 1$.

Nota 2.10. *La anterior noción de determinante se extiende sin dificultad a $K^{n\times n}$ con K un cuerpo.*

Se puede probar que las anteriores propiedades determinan de forma unívoca la aplicación y que viene dada por

$$\det(A) = \sum_{\sigma\in S_n} (-1)^{s(\sigma)} \prod_{i=1}^{n} A(i,\sigma(i)),$$

donde $s(\sigma)$ representa el número de inversiones en σ respecto de la permutación identidad (cambios respecto del orden natural entre los primeros n números naturales).

Proposición 2.13. *Sean $A, B \in \mathbb{R}^{n\times n}$ y $\alpha \in \mathbb{R}$. Se cumplen:*

(i) $\det(A) = \det(A')$.

(ii) $\det(A) = 0$ *si, y sólo si, las columnas de A son linealmente dependientes.*

(iii) *Si $C(:,i) = A(:,i) + \alpha A(:,j)$, y $C(:,k) = A(:,k)$, $k \neq i$, entonces* $\det(A) = \det(C)$.

(iv) $\det(AB) = \det(A)\det(B)$.

(v) *Si A es regular, entonces* $\det(A^{-1}) = \det(A)^{-1}$.

(vi) *Sea $1 \leq j \leq n$. Entonces* $\det(A) = \sum_{i=1}^{n} A(i,j)\, A_{ij}$, *donde* $A_{ij} = (-1)^{i+j} M_{ij}$ *(A_{ij} cofactor de A(i,j)) y M_{ij} es el determinante de la matriz obtenida al eliminar de A la fila i-ésima y la columna j-ésima.*

(vii) *Si A es regular, entonces* $A^{-1}(i,j) = A_{ji}/\det(A)$.

Definición 2.13. *La matriz conjugada de $A \in \mathbb{C}^{m\times n}$ es la matriz $\overline{A} \in \mathbb{C}^{m\times n}$ determinada por*

$$\overline{A}(i,j) = \overline{A(i,j)}, \quad 1 \leq i \leq m,\ 1 \leq j \leq n.$$

Ejercicio 2.12. *Probar las siguientes afirmaciones para matrices de dimensiones adecuadas:* $\overline{\overline{A}} = A$, $\overline{A+B} = \overline{A} + \overline{B}$, $\overline{AB} = \overline{A}\,\overline{B}$, $\overline{\alpha A} = \overline{\alpha}\overline{A}$, $\overline{A}' = \overline{A'}$, $\overline{A}^{-1} = \overline{A^{-1}}$.

Definición 2.14. *El número complejo λ es una valor propio de $A \in \mathbb{R}^{n\times n}$ si existe $x \in \mathbb{C}^n$, no nulo, tal que*

$$Ax = \lambda x.$$

En este caso, diremos que x es un vector propio asociado a λ.

Mediante $\sigma(A)$ denotaremos el conjunto de los valores propios asociados a A.

La condición anterior se puede interpretar como que las columnas de la matriz $A - \lambda$ son linealmente dependientes, o equivalentemente

$$\det(A - \lambda) = 0,$$

que es la llamada ecuación característica. Así $\sigma(A)$ es el conjunto formado por las raíces del llamado polinomio característico $\det(A-\lambda)$, que es un polinomio de grado n y que, por tanto, tendrá en $\mathbb{C}$ exactamente n raíces si las contamos tantas veces como indique su orden de multiplicidad.

Ejercicio 2.13. $\det(A) = \prod_{\lambda\in\sigma(A)} \lambda^{k(\lambda)}$. *(En efecto,* $P_A(\mu) = \det(A-\mu) = (-1)^n \prod_{\lambda\in\sigma(A)}(\mu-\lambda)^{k(\lambda)}$, *luego,* $\det(A) = P_A(0) = (-1)^{2n} \prod_{\lambda\in\sigma(A)} \lambda^{k(\lambda)}$).

Nota 2.11. *Dado un polinomio $P(\lambda) = \lambda^n + a_1\lambda^{n-1} + \ldots + a_{n-1}\lambda + a_n$, la matriz (matriz 'companion')*

$$A = \begin{bmatrix} -a_1 & -a_2 & \cdots & -a_n \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}$$

tiene por ecuación característica $P(\lambda) = 0$.

Sea $\lambda \in \sigma(A)$. El conjunto de vectores propios asociado

$$S_\lambda(A) = \{x \in \mathbb{C}^n : Ax = \lambda x\},$$

es un $\mathbb{C}$-espacio vectorial de dimensión (llamada dimensión geométrica) entre uno y el orden de multiplicidad de λ como raíz del polinomio característico (multiplicidad algebraica que denotaremos mediante $k(\lambda)$). Si $\lambda \in \sigma(A) \cap \mathbb{R}$ entonces $S_\lambda = \{x \in \mathbb{R}^n : Ax = \lambda x\}$ es un espacio vectorial real de dimensión entre 1 y $k(\lambda)$.

Recordemos porqué podemos considerar vectores propios reales en este caso. Si $x \in S_\lambda$, entonces la descomposición en parte real e imaginaria de la relación $Ax = \lambda x$ es

$$A\,\mathrm{Re}(x) + iA\,\mathrm{Im}(x) = \lambda\,\mathrm{Re}(x) + i\lambda\,\mathrm{Im}(x),$$

por lo que $\mathrm{Re}(x)$, $\mathrm{Im}(x) \in S_\lambda$.

Ejercicio 2.14. $\sigma(A) = \sigma(A')$.

Nota 2.12. *Para $A \in \mathbb{C}^{n\times n}$ se cumple que*

$$\mathbb{C}^n = \bigoplus_{\lambda\in\sigma(A)} \ker(A-\lambda)^{k(\lambda)}.$$

Proposición 2.14. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica. Entonces*

(i) $\sigma(A) \subset \mathbb{R}$.

(ii) *Si A es semidefinida positiva, entonces $\sigma(A) \subset [0, \infty)$.*

(iii) *Si A es definida positiva, entonces $\sigma(A) \subset (0, \infty)$.*

Demostración. Sean $\lambda \in \sigma(A)$ y $x \in$ tales que $Ax = \lambda x$. Entonces, teniendo en cuenta que A es simétrica y real,

$$\lambda = \frac{\overline{x}'Ax}{\overline{x}'x} = \left(\frac{\overline{x}'Ax}{\overline{x}'x}\right)' = \frac{x'A\overline{x}}{\overline{x}'x} = \overline{\lambda}, \tag{2.2}$$

lo que prueba (i). Los apartados (ii) y (iii) son consecuencia de que al ser los valores propios reales podemos considerar vectores propios asociados reales, y de la primera igualdad de (2.2) (el cociente se llama cociente de Rayleigh) □

Ejercicio 2.15. *¿Es cierto el resultado anterior sin ser la matriz simétrica? No, la matriz*

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

es definida positiva y $\sigma(A) = \{1 \pm i\}$.

Definición 2.15. *Una matriz $A \in \mathbb{R}^{n\times n}$ es diagonalizable si existe $P \in \mathbb{R}^{n\times n}$, regular, tal que*

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

es diagonal.

En realidad en la definición anterior admitimos sólo que sea diagonalizable en $\mathbb{R}$. La condición de ser diagonalizable se puede reinterpretar como sigue. Por una parte

$$\det(A-\lambda) = \det(P^{-1}AP-\lambda) = \det(D-\lambda) = \prod_{i=1}^{n}(D(i,i)-\lambda)$$

y por tanto, $\sigma(A) = \sigma(D) = diag(D)$. Además, como $PD = AP$, entonces

$$AP(:,j) = D(j,j)P(:,j), \quad 1 \le j \le n.$$

Al ser las columnas de P una base de $\mathbb{R}^n$, por ser regular, y puesto que acabamos de comprobar que las columnas de P son vectores propios asociados a los valores propios de A, resulta que existe una base de $\mathbb{R}^n$ formada por vectores propios asociados a valores propios de A. El recíproco es, evidentemente, cierto, de ahí que pueda usarse esta propiedad como definición alternativa de matriz diagonalizable.

Hay matrices que no son diagonalizables ni en $\mathbb{C}$, ya que existe $\lambda \in \sigma(A)$ tal que $dim(S_\lambda) < k(\lambda)$, donde $k(\lambda)$ es la multiplicidad del valor propio como raíz del polinomio característico. Téngase en cuenta que vectores propios asociados a valores propios distintos son independientes.

Ejercicio 2.16. *La matriz*

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

no es diagonalizable en $\mathbb{R}$ ya que $\sigma(A) = \{\pm i\}$, pero si lo es en $\mathbb{C}$ ya que

$$\begin{bmatrix} \frac{1}{2} & -\frac{1}{2}i \\ \frac{1}{2} & \frac{1}{2}i \end{bmatrix} A \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix} = \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix}.$$

La matriz

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

tiene 0 como único valor propio, pero no es diagonalizable, ni en $\mathbb{C}$, por la situación antes mencionada. Veamos de todas formas una prueba directa de la afirmación. Si existiera $P \in \mathbb{C}^{n\times n}$, regular, tal que $P^{-1}AP = zeros(2)$, entonces realizando los productos de matrices llegaríamos a que $P(2,:)$ es el vector nulo, lo que contradice la regularidad de P.

Por contra, toda matriz con valores propios reales simples es diagonalizable. También las matrices simétricas, como vemos a continuación.

Teorema 2.2. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica. Existe $P \in \mathbb{R}^{n\times n}$ ortogonal tal que $P'AP$ es diagonal.*

Demostración. Por inducción sobre n. Para $n = 1$ es evidente que podemos conseguir el resultado deseado. Supongamos el resultado cierto para $n-1$. Sea $\lambda \in \sigma(A) \subset \mathbb{R}$ y $x \in \mathbb{R}^n$ vector propio asociado con $\|x\|_2 = 1$. Por un resultado previo, existe $Q \in \mathbb{R}^{n\times n}$, ortogonal, con $Q(:,1) = x$. Entonces

$$(Q'AQ)(:,1) = \lambda Q'x = \lambda e_1.$$

Como $Q'AQ$ es simétrica por serlo A, entonces

$$Q'AQ = \begin{bmatrix} \lambda & 0 \\ 0 & B \end{bmatrix}$$

donde $B \in \mathbb{R}^{(n-1)\times(n-1)}$. Como B es simétrica ($B(i,j) = (Q'AQ)(i+1, j+1)$), entonces por la hipótesis de inducción existe $R \in \mathbb{R}^{(n-1)\times(n-1)}$, ortogonal, tal que $E = R^{'}BR$ es diagonal. Tomamos

$$P = Q \begin{bmatrix} 1 & 0 \\ 0 & R \end{bmatrix} \in \mathbb{R}^{n\times n}$$

que es ortogonal

$$P'P = \begin{bmatrix} 1 & 0 \\ 0 & R' \end{bmatrix} Q'Q \begin{bmatrix} 1 & 0 \\ 0 & R \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & R' \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & R \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & R'R \end{bmatrix} = I_n,$$

y además

$$P'AP = \begin{bmatrix} 1 & 0 \\ 0 & R' \end{bmatrix} \begin{bmatrix} \lambda & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & R \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & R'BR \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & E \end{bmatrix}.$$

□

Nota 2.13. *En dimensión infinita hay un resultado similar para operadores compactos y autoadjuntos T sobre un espacio de Hilbert X separable; se prueba que existe una base Hilbertiana de X formada por vectores propios asociados al operador T. Es frecuente utilizar bases especiales formadas por funciones propias asociadas a un operador diferencial (por ejemplo en el tratamiento de problemas mixtos asociados a EDPs podemos utilizar desarrollos en serie de Fourier respecto de bases obtenidas a partir de un operador diferencial de tipo Sturm-Liouville)*

Nota 2.14. *La base ortonormal de $\mathbb{R}^n$ formada por vectores propios de una matriz simétrica se puede obtener hallando una base ortonormal en cada subespacio propio y uniéndolas todas ellas.*

Nota 2.15 (Teorema de Schur)**.** *Con el mismo tipo de argumento que en el teorema previo, se puede probar que para toda $A \in \mathbb{R}^{n\times n}$ existe $Q \in \mathbb{C}^{n\times n}$ unitaria ($\overline{Q}'Q = I_n$) tal que*

$$\overline{Q}'AQ$$

es triangular superior. Observar que la diferencia estriba en que al no ser simétrica no podemos afirmar que los valores propios son reales ni que la transpuesta de la primera fila coincida con la primera columna.

Aparentemente, este resultado permite obtener los valores propios de una matriz, pero hay que destacar que la anterior factorización se basa en el conocimiento de valores y vectores propios, y sin conocerlos no es fáctible obtenerla en un número finito de pasos.

En general, no se puede afirmar que la matriz Q pueda elegirse real. En efecto, sea

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

Si existiera $Q \in \mathbb{R}^{2\times 2}$, regular, tal que $Q^{-1}AQ$ es triangular superior, entonces se comprueba fácilmente que $Q(:,1)$ debe ser el vector nulo.

Corolario 2.2. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica definida positiva. Entonces*

$$\det(A_k) > 0, \quad 1 \leq k \leq n.$$

Demostración. Por ser A simétrica, existe $P \in \mathbb{R}^{n\times n}$, ortogonal, tal que $D = P'AP$ es diagonal. Por ser A simétrica definida positiva, entonces $\sigma(A) \subset (0,\infty)$, por lo que $D(i,i) > 0$, $1 \leq i \leq n$. Así,

$$\det(A) = \det(D) = \prod_{i=1}^{n} D(i,i) > 0.$$

El resultado es consecuencia de que las matrices A_k gozan de la misma propiedad que A. □

Nota 2.16. *El resultado anterior es inmediato si recordamos que el producto de los valores propios nos proporciona el determinante y que éstos son todos positivos. En el tema de matrices especiales veremos que el recíproco también es cierto, y, por tanto, tendremos una caracterización para matrices simétricas de la propiedad de ser definida positiva, que se conoce como Criterio de Sylvester.*

Corolario 2.3. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica. Entonces A es definida positiva si, y sólo si, $\sigma(A) \subset (0,\infty)$.*

Demostración. Por ser A simétrica, existe $P \in \mathbb{R}^{n\times n}$, ortogonal, tal que $D = P'AP$ es diagonal. Por tener todas las entradas de la diagonal principal positivas, D es definida positiva. Como ya probó en Proposición 2.9, $A = PDP'$ es definida positiva. □

Definición 2.16. *El radio espectral de $A \in \mathbb{R}^{n\times n}$ es el número real no negativo*

$$r(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}.$$

Corolario 2.4. *Si $A \in \mathbb{R}^{n\times n}$ es simétrica, entonces*

$$\|A\|_2 = r(A).$$

Demostración. Por el teorema previo existe $P \in \mathbb{R}^{n\times n}$ ortogonal tal que $D = P'AP$ es diagonal. Por tanto

$$\|A\|_2 = \left\|PDP'\right\|_2 = \|D\|_2 = \|diag(D)\|_\infty = r(A).$$

□

La relación no es cierta en general; considerar por ejemplo una matriz cuyo único valor propio es el cero, pero no es identicamente nula. El siguiente teorema relaciona completamente las cantidades $\|A\|$ y $r(A)$, donde $\|\|$ representa una norma matricial subordinada.

Teorema 2.3. *Sea $A \in \mathbb{R}^{n\times n}$. Entonces*

(i) *Para toda norma matricial $\|\|$ subordinada a una norma vectorial compleja $\|\|$ se tiene*

$$r(A) \leq \|A\| .$$

(ii) *Para todo $\varepsilon > 0$, existe una norma matricial subordinada, $\|\|_\varepsilon$, tal que*

$$\|A\|_\varepsilon \leq r(A) + \varepsilon.$$

Demostración. Sea $x \in \mathbb{C}^n$ un vector propio no nulo asociado al valor propio $\lambda \in \mathbb{C}$. Entonces,

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

de donde se deduce (i).

Probemos (ii). Por el teorema de Schur existe $Q \in \mathbb{C}^{n\times n}$, regular, tal que $T = Q^{-1}AQ$ es triangular superior, y por tanto, $\sigma(A) = \sigma(T) = \{T(i,i)\}_{i=1}^n$. Para $\delta > 0$, aún por determinar, consideramos la matriz

$$D_\delta = diag([1 \, \delta \, \ldots \, \delta^{n-1}]),$$

cuya inversa es $D_{\delta^{-1}}$, y calculamos

$$D_{\delta^{-1}}TD_\delta = \begin{bmatrix} T(1,1) & \delta T(1,2) & \delta^2 T(1,3) & \ldots & \delta^{n-1}T(1,n) \\ 0 & T(2,2) & \delta T(2,3) & \ldots & \delta^{n-2}T(2,n) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & T(n-1,n-1) & \delta T(n-1,n) \\ 0 & 0 & \ldots & 0 & T(n,n) \end{bmatrix}.$$

Elegimos ahora δ cumpliendo

$$\sum_{k=i+1}^{n} \delta^{k-i} \, |T(i,k)| \leq \varepsilon, \quad 1 \leq i \leq n-1,$$

lo que es factible ya que las expresiones de la izquierda de las desigualdades anteriores son funciones de δ que tienden a cero cuando $\delta \to 0$. Con ello conseguimos que

$$\left\|D_{\delta^{-1}}Q^{-1}AQD_\delta\right\|_\infty \leq r(A) + \varepsilon, \tag{2.3}$$

por lo que sólo resta introducir una norma vectorial adecuada.

Introducimos la matriz regular $B = QD_\delta$ y definimos la norma

$$\|x\|_\varepsilon = \left\|B^{-1}x\right\|_\infty, \quad x \in \mathbb{C}^n.$$

Entonces, para toda $S \in \mathbb{R}^{n\times n}$ se tiene

$$\begin{aligned}\|S\|_\varepsilon &= \max_{\|x\|_\varepsilon=1,\, x\in\mathbb{C}^n} \|Sx\|_\varepsilon = \max_{\|B^{-1}x\|_\infty=1,\, x\in\mathbb{C}^n} \left\|B^{-1}Sx\right\|_\infty \\ &= \max_{\|y\|_\infty=1,\, y\in\mathbb{C}^n} \left\|B^{-1}SBy\right\|_\infty = \left\|B^{-1}SB\right\|_\infty,\end{aligned}$$

lo que proporciona (ii) por la desigualdad (2.3). □

Nota 2.17. *Evidentemente primer apartado del teorema es cierto para normas matriciales sobre $\mathbb{C}^{n\times n}$ que sean multiplicativas, como por ejemplo la norma de Frobenius.*

Notar que en (ii) la norma $\|\|_\varepsilon$ la podemos encontrar tanto compleja como real, ya que $\|\|_\varepsilon$ se puede restringir a $\mathbb{R}^n$ obteniéndose una norma vectorial real, y además

$$\max_{\|x\|_\varepsilon=1,\, x\in\mathbb{R}^n} \|Sx\|_\varepsilon \leq \max_{\|x\|_\varepsilon=1,\, x\in\mathbb{C}^n} \|Sx\|_\varepsilon.$$

Ejercicio 2.17. *Sean $A \in \mathbb{R}^{n\times n}$ y $p = 1, 2, \infty$. Entonces*

$$\max_{\|x\|_p=1,\, x\in\mathbb{R}^n} \|Ax\|_p = \max_{\|x\|_p=1,\, x\in\mathbb{C}^n} \|Ax\|_p.$$

(Según afirma P. G. Ciarlet, el resultado anterior no es cierto de forma general).

2.4. La factorización mediante valores singulares (SVD)

Definición 2.17 (Singular Value Decomposition)**.** *Sean $m, n \in \mathbb{N}$. Sea $A \in \mathbb{R}^{m\times n}$. Una factorización mediante valores singulares de A, o SVD, es aquella de la forma*

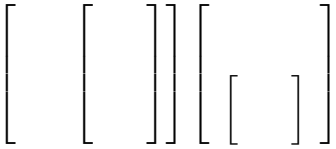
$$A = U\Sigma V',$$

donde $U \in \mathbb{R}^{m\times m}$ y $V \in \mathbb{R}^{n\times n}$ son ortogonales, y $\Sigma \in \mathbb{R}^{m\times n}$ es una matriz diagonal, de tal manera que si $diag(\Sigma) = [\mu_1; \mu_2; \ldots; \mu_p]$, $p := \min(m, n)$, entonces se cumple

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_p \geq 0.$$

Los elementos $diag(\Sigma)$ son los llamados valores singulares de A. Las columnas de U, $u_i := U(:, i)$, $1 \leq i \leq m$, son los llamados vectores singulares a izquierda, y las columnas de V, $v_i := V(:, i)$, $1 \leq i \leq n$, son los llamados vectores singulares a derecha.

Nota 2.18. *Para $m \geq n$, existe otra versión de la factorización SVD llamada por algunos autores versión económica ('economic size', 'thin'). Nótese que las submatrices recuadradas dentro de las matrices de la factorización SVD en el siguiente esquema, son en realidad innecesarias por el caracter diagonal de Σ:*



de ahí que se de una definición alternativa en la forma $U\Sigma V'$ en la que $U \in \mathbb{R}^{m\times n}$ cumple $U'U = I_n$, $\Sigma \in \mathbb{R}^{n\times n}$ es diagonal y $V \in \mathbb{R}^{n\times n}$ es ortogonal. Bastaría pues eliminar de la SVD la parte recuadrada para pasar a la versión económica. Y viceversa, desde la versión económica podemos pasar a la estándar a base de completar las columnas de U a una base ortonormal $\mathbb{R}^m$ y ampliando Σ de forma que $\Sigma(i,:)$ son el vector nulo para $n+1 \leq i \leq m$.

Si $A \in \mathbb{R}^{m\times n}$ con $m < n$, la factorización mediante valores singulares cuando se trabaja con la versión económica se suele establecer en términos de la correspondiente a A'.

Asumamos en los comentarios siguientes que existe una factorización SVD de A, $A = U\Sigma V'$, y analicemos que papel juega cada elemento de la factorización y que información aporta dicha factorización sobre la matriz A. Este análisis proporcionará las claves para demostrar la existencia de la SVD. En primer lugar nótese que

$$A = \sum_{j=1}^{p} \mu_j u_j v_j'. \tag{2.4}$$

Por lo tanto, para reproducir A sólo necesitamos las primeras p columnas de U y V junto con los valores singulares. En realidad, por la relación (2.4), sólo las primeras p columnas de U y V son las que nos interesan de la factorización, los restantes sólo se utilizan para conseguir las dimensiones requeridas (ver nota previa de nuevo).

Nótese que las matrices $\left\|u_j v_j'\right\|_2 = 1$, $1 \leq j \leq p$, por lo que la relación puede visualizarse como la de expresar la matriz A como superposición de 'capas' o matrices del mismo tamaño que la original, pero con pesos o 'importancia' dependiendo del correspondiente valor singular. En este sentido, las primeras 'capas' serían las más 'importantes'.

Además, tenemos

$$\|A\|_2 = \|\Sigma\|_2 = \|diag(\Sigma)\|_\infty = \mu_1.$$

Además,

$$A'A = V\Sigma'U'U\Sigma V' = V\Sigma'\Sigma V'.$$

Por lo tanto, las matrices $A'A$ y $\Sigma'\Sigma$ son semejantes, lo que implica que tienen los mismos valores propios. Lo mismo puede decirse de AA' y $\Sigma\Sigma'$. Si

$m \geq n = p$, como $\Sigma'\Sigma = diag([\mu_1^2; \ldots; \mu_p^2])$, entonces los valores singulares son las raíces cuadradas positivas de los valores propios de $A'A$, mientras que de forma análoga, si $m < n$ y los valores singulares son las raíces cuadradas positivas de AA'. En realidad, las primeras p, en orden decreciente, raíces cuadradas positivas de los valores propios de $A'A$ y de AA' son las mismas. Recuérdese además que $A'A \in \mathbb{R}^{n\times n}$ y $AA' \in \mathbb{R}^{m\times m}$ son simétricas semidefinidas positivas, y que, por lo tanto, sus valores propios son no negativos. Y ambas matrices son diagonalizables mediante matrices ortogonales.

Por otra parte, para , se tiene

$$Av_k = \mu_k u_k,\ 1 \leq k \leq p;\ Av_k = 0,\ p+1 \leq k \leq n,$$

y

$$A'u_k = \mu_k v_k,\ 1 \leq k \leq p;\ A'u_k = 0,\ p+1 \leq k \leq m.$$

En consecuencia,

$$A'Av_k = \mu_k A'u_k = \mu_k^2 v_k,\ 1 \leq k \leq p;\ A'Av_k = 0,\ p+1 \leq k \leq n,$$

luego los vectores singulares a derecha son vectores propios de la matriz $A'A$, y

$$AA'u_k = \mu_k Av_k = \mu_k^2 u_k,,\ 1 \leq k \leq p;\ AA'u_k = 0,\ p+1 \leq k \leq m,$$

es decir, los vectores a izquierda de A son vectores propios de la matriz $A'A$.

La existencia de una factorización SVD de A se traduciría en que existen bases ortonormales en $\mathbb{R}^n$ y $\mathbb{R}^m$ tales que la aplicación lineal determinada por A tiene matriz asociada respecto de dichas bases diagonal (Σ):

$$\begin{array}{rccc} A: & (\mathbb{R}^n, \{v_i\}_{i=1}^n) & \to & (\mathbb{R}^m, \{u_i\}_{i=1}^m) \\ & x = \sum_{i=1}^n x_i v_i & \rightsquigarrow & \sum_{j=1}^p \mu_j x_j u_j \end{array}$$

Si $s \in \{0, 1, \ldots, p\}$ representa el número de valores singulares no nulos, entonces es muy fácil comprobar que $rank(A) = s$. La definición del rango de una matriz en términos de la SVD es utilizada ampliamente dentro del ámbito del Análisis Numérico. Además, partiendo de una SVD de A se obtienen de forma inmediata bases de los espacios núcleo e imagen de A (Proposicion 2.15).

La diferencia con el problema de diagonalización de matrices cuadradas, estriba en que en este caso la base es la misma en ambos espacios, aunque no necesariamente es ortonormal; en contraposición, así como no toda matriz es diagonalizable, ni tan siquiera en $\mathbb{C}$, veremos a continuación que toda matriz sí admite una factorización SVD. Por supuesto, la SVD de una matriz, tal como se ha definido, no es en general única.

Teorema 2.4 (Existencia)**.** *Toda matriz $A \in \mathbb{R}^{m\times n}$ posee una SVD.*

Demostración. Sean $\{\lambda_j^2\}_{j=1}^n$, con $\lambda_j \geq 0$, $1 \leq j \leq n$, los valores propios de $A'A$ ordenados en orden decreciente, es decir,

$$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_n^2 \geq 0,$$

y sea $\{v_j\}_{j=1}^n$ una base ortonormal de $\mathbb{R}^n$ formada por vectores propios de $A'A$, de forma que

$$A'Av_j = \lambda_j^2 v_j \quad (1 \leq j \leq n).$$

Consideramos la matriz ortogonal $V = [v_1, \ldots, v_n]$. Comprobemos que los vectores de la familia $\{Av_j\}_{j=1}^n$ son ortogonales dos a dos. En efecto,

$$v_j' A'Av_i = \lambda_i^2 v_j' v_i = \lambda_i^2 \delta_{ij} \quad (1 \leq i, j \leq n).$$

En particular, $\|Av_j\|_2 = \lambda_j$, $1 \leq j \leq n$, por lo que Av_j es el vector nulo si, y sólo si, $\lambda_j = 0$. Como $rank(A) = \dim(\{Av_j\}_{j=1}^n) \leq p = \text{mín}(m, n)$, entonces $\lambda_j = 0$, $p+1 \leq j \leq n$. Sea $s \in \{0, 1, \ldots, p\}$ el número de valores propios de $A'A$ no nulos, es decir, $\lambda_j = 0$, $s+1 \leq j \leq n$. Nótese que $Av_j = 0$, $s+1 \leq j \leq n$. Definimos

$$u_j := Av_j/\lambda_j \quad (1 \leq j \leq s),$$

y, si fuera necesario por tenerse $s < m$, completamos $\{u_i\}_{i=1}^s$ a $\{u_i\}_{i=1}^m$, base ortonormal de $\mathbb{R}^m$. Definimos $U =: [u_1, \ldots, u_m]$ y $\Sigma := diag([\lambda_1; \lambda_2; \ldots; \lambda_p]) \in \mathbb{R}^{m \times n}$. Entonces

$$\begin{aligned} U\Sigma V' &= \sum_{j=1}^s \lambda_j u_j v_j' = \sum_{j=1}^s Av_j v_j' = \sum_{j=1}^n Av_j v_j' \\ &= A \sum_{j=1}^n v_j v_j' = AVV' = A, \end{aligned}$$

probando que $U\Sigma V'$ es una SVD de A. □

Nota 2.19 (extensión de un conjunto linealmente independiente a una base)**.** *Consideremos $\{u_i\}_{i=1}^s \subset \mathbb{R}^n$, $s < n$, linealmente independiente. Vamos a determinar vectores $\{u_i\}_{i=s+1}^n$ tal que $\{u_i\}_{i=1}^n$ es una base de $\mathbb{R}^n$. Supongamos $u_1 = \sum_{i=1}^n \lambda_{1i}\, e_i$ y que $\lambda_{1i_1} \neq 0$. Entonces $\{u_1\} \cup \{e_i\}_{i=1, i \neq i_1}^n$ es una base de $\mathbb{R}^n$. Entonces $u_2 = \lambda_{2i1} u_1 + \sum_{i=1, i \neq i_1}^n \lambda_{2i}\, e_i$ con algún $\lambda_{2i_2} \neq 0$, $i_2 \neq i_1$. Así, $\{u_1, u_2\} \cup \{e_i\}_{i=1, i \neq i_1, i_2}^n$ es una base de $\mathbb{R}^n$. Por un proceso de inducción sobre s se obtendría el resultado deseado.*

Nota 2.20. *La demostración del teorema anterior proporciona un método para obtener una descomposición SVD de una matriz, pero no es conveniente desde el punto de vista numérico, ya que los errores producidos en el cálculo de $A'A$ con aritmética en punto flotante pueden ser determinantes, es decir, es un proceso sensible a las perturbaciones.*

Nota 2.21. *También es cierto que toda matriz compleja admite una factorización SVD, sólo que las matrices de transformación en este caso son unitarias ($\overline{R}'R = I_n$, $R \in \mathbb{C}^{n \times n}$). Nuestro objetivo es tratar matrices reales, de ahí que hayamos evitado esa pequeña complicación adicional en la exposición.*

Ejemplo 2.1. *Obtengamos una factorización SVD de* $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \\ -1 & 1 \end{bmatrix}$. *Tenemos*

$$A'A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix},$$

uyos valores propios son $\{3, 1\}$*, por lo que* $\mu_1 = \sqrt{3}$ *y* $\mu_2 = 1$*, con vectores propios asociados* $[-1; 1]$ *y* $[1; 1]$ *respectivamente. Tomamos* $v_1 = [-1; 1]/\sqrt{2}$*,* $v_2 = [1; 1]/\sqrt{2}$*. Tenemos* $Av_1 = [1; 1; 2]/\sqrt{2}$ *y* $Av_2 = [1; -1; 0]/\sqrt{2}$*. Tomamos* $u_1 = [1; 1; 2]/\sqrt{6}$ *y* $u_2 = [1; -1; 0]/\sqrt{2}$*. Completamos a una base ortonormal tomando* $u_3 = [1; 1; -1]/\sqrt{3}$ *y tenemos*

$$\begin{aligned} & [u_1, u_2, u_3] \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} [v_1, v_2]' \\ = \; & \frac{1}{2\sqrt{3}} \begin{bmatrix} 1 & \sqrt{3} & \sqrt{2} \\ 1 & -\sqrt{3} & \sqrt{2} \\ 2 & 0 & -\sqrt{2} \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} = A. \end{aligned}$$

Ejemplo 2.2. *Obtengamos una factorización SVD de* $A = \begin{bmatrix} 0 & 1 & 1 \\ 0 & -1 & -1 \end{bmatrix}$. *Aunque sería más cómodo trabajar con* AA'*, vamos a realizar los cálculos tal y como se proponen en la primera demostración de existencia. Tenemos*

$$A'A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 2 & 2 \end{bmatrix},$$

cuyos valores propios son $\{4, 0, 0\}$*, por lo que* $\mu_1 = 2$ *y* $\mu_2 = 0$*, con vectores propios asociados* $[0; 1; 1]$*,* $[1; 0; 0]$ *y* $[0; -1; 1]$ *respectivamente. Como los vectores propios asociados al valor propio doble* 0 *son ortogonales, tomamos* $v_1 = [0; 1; 1]/\sqrt{2}$*,* $v_2 = [1; 0; 0]$ *y* $v_3 = [0; -1; 1]/\sqrt{2}$*. Tenemos* $Av_1 = [2; -2]/\sqrt{2}$*,* $Av_2 = Av_3 = [0; 0]$*. Tomamos* $u_1 = [1; -1]/\sqrt{2}$ *y completamos a una base ortonormal con* $u_2 = [-1; -1]/\sqrt{2}$*. Entonces*

$$\begin{aligned} & [u_1, u_2] \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} [v_1, v_2, v_3]' \\ = \; & \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ \sqrt{2} & 0 & 0 \\ 0 & -1 & 1 \end{bmatrix} = A. \end{aligned}$$

Listamos a continuación algunos resultados interesantes que relacionan la SVD con las propiedades de la matriz. Destacar el apartado (iii) que expresa $rank(A) = \dim(ran(A))$ en términos de los valores singulares. Está es la definición más útil desde el punto de vista numérico.

Proposición 2.15. *Sean* $A \in \mathbb{R}^{m \times n}$ *y* $p := \min\{m, n\}$*. Sea* $A = U\Sigma V'$ *una factorización SVD de A. Se cumplen:*

(i) *Los valores singulares de $A \in \mathbb{R}^{m\times n}$ son las raíces cuadradas positivas de los primeros p, en orden decreciente, valores propios de $A'A$ o de AA'.*

(ii) $\|A\|_F^2 := \sum_{i=1}^m \sum_{j=1}^n |A(i,j)|^2 = \sum_{i=1}^p \mu_i^2$.

(iii) $\|A\|_2 = \mu_1 = \sqrt{r(A'A)} = \sqrt{r(AA')} = \|A'\|_2$.

(iv) *Si $m = n$,* $|\det(A)| = \prod_{i=1}^n \mu_i$.

(v) *Sea $s \in \{0, 1, \ldots, p\}$ el número de valores singulares de A no nulos. Entonces $rank(A) = s$,* $\ker A = \left\langle \{v_i\}_{i=s+1}^n \right\rangle$ *e* $im\, A := \langle \{A(:,i)\}_{i=1}^n \rangle = \langle \{u_i\}_{i=1}^s \rangle$.

(vi) *Supongamos $m = n$. Entonces, A es regular si, y sólo si, $\mu_n > 0$. En este caso,*

$$\left\|A^{-1}\right\|_2 = \mu_n^{-1}$$

(o equivalentemente, $\min_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \mu_n$*). Por lo tanto,* $k_2(A) := \|A\|_2 \left\|A^{-1}\right\|_2 = \mu_1/\mu_n$.

Demostración. (ii)-(iii) Consecuencia de que la norma de Frobenius y $\|\|_2$ son invariantes por transformaciones ortogonales.

(iv) Obsérvese que $|\det(U)| = |\det(V)| = 1$.

(v) Téngase en cuenta que respecto de las bases formadas por vectores singulares la aplicación lineal representada por A es diagonal, junto con el teorema de la dimensión.

(vi) La primera parte es consecuencia de (iv) o de (v). Para la segunda

$$\left\|A^{-1}\right\|_2 = \left\|V\Sigma^{-1}U'\right\|_2 = \left\|\Sigma^{-1}\right\|_2 = \mu_n^{-1}.$$

□

Ejercicio 2.18. *Compruébese que $\sigma(A'A) \cap (0, +\infty) = \sigma(AA') \cap (0, +\infty)$.*

Estudiese como transforma una aplicación lineal del plano en el plano la circunferencia unidad.

La factorización de una matriz simétrica mediante sus valores propios es una SVD de la matriz? Y si la matriz es además semidefinida positiva?

Teorema 2.5 (Teorema de la mejor aproximación)**.** *Sean $A \in \mathbb{R}^{m\times n}$, con $rank(A) = s \geq 1$. Sea $A = U\Sigma V'$ una descomposición SVD de A. Sean*

$$A_k := \sum_{j=1}^k \mu_j u_j v_j' \quad (1 \leq k \leq s-1).$$

Entonces

$$\mu_{k+1} = \|A - A_k\|_2 = \min\left\{\|A - B\|_2 : B \in \mathbb{R}^{m\times n},\ rank(B) \leq k\right\} \quad (1 \leq k \leq s-1).$$

Demostración. Al igual que hemos razonado anteriormente con A, $rank(A_k) = k$, y también sabemos que $\mu_{k+1} = \|A - A_k\|_2$. Sea $B \in \mathbb{R}^{m\times n}$ con $rank(B) \leq k$. Como por el teorema de la dimensión $n = rank(B) + \dim(\ker B)$, entonces $\dim(\ker B) \geq n - k$. Consideremos el espacio vectorial $W := \left\langle \{v_i\}_{i=1}^{k+1} \right\rangle$ que evidentemente tiene dimensión $k+1$. Así tenemos

$$\begin{aligned} n + \dim(W \cap \ker B) &\geq \dim(W + \ker B) \\ &\quad + \dim(W \cap \ker B) \\ &\geq n - k + k + 1 = n + 1, \end{aligned}$$

luego existe algún vector unitario $y \in \ker B \cap W$ y se tiene

$$\begin{aligned} \|A - B\|_2^2 &\geq \|Ay\|_2^2 = \left\| \sum_{j=1}^{k+1} \mu_j (v_j' y) u_j \right\|_2^2 \\ &= \sum_{j=1}^{k+1} \mu_j^2 (v_j' y)^2 \geq \mu_{k+1}^2 \sum_{j=1}^{k+1} (v_j' y)^2 \\ &= \mu_{k+1}^2 \|y\|_2^2 = \mu_{k+1}^2, \end{aligned}$$

de ahí que el mínimo se alcanze en A_k. □

Nota 2.22. *El teorema anterior se enuncia para los casos donde realmente dice algo interesante. Analicemos el resto de casos. En el caso de μ_1, donde entenderíamos que estamos comparando con la matriz nula, $A_0 := 0$, tendríamos*

$$\mu_1 = \|A\|_2 = \min \left\{ \|A - B\|_2 : B \in \mathbb{R}^{m\times n},\ rank(B) = 0 \right\},$$

pues una matriz tiene rango cero si y sólo si es la matriz nula. En el caso de que haya índices $s + 1 \leq k \leq p$, se tiene que la matriz A tiene rango menor que k, $\mu_k = 0$ y $A_{k-1} = A$, de donde se deduce inmediatamente que

$$0 = \mu_k = \|A - A_{k-1}\|_2 = \min \left\{ \|A - B\|_2 : B \in \mathbb{R}^{m\times n},\ rank(B) \leq k \right\} = 0.$$

En definitiva, con la notación $A_0 := 0$, el teorema anterior se puede extender a todos los índices $1 \leq k \leq p$ de la siguiente forma:

$\mu_k = \|A - A_{k-1}\|_2 = \min \{\|A - B\|_2 : B \in \mathbb{R}^{m\times n},\ rank(B) < k\}$ $(1 \leq k \leq p)$.

Nota 2.23. *Del teorema anterior se desprende que si A es regular, entonces*

$$\frac{1}{\|A^{-1}\|_2} = \min \{\|A - B\|_2 : B \text{ es singular}\}.$$

Se puede probar, por otra vía, que el resultado es cierto para toda norma matricial subordinada.

Corolario 2.5. *Sean A, $B \in \mathbb{R}^{m\times n}$ y sea $p = \min\{m,n\}$. Entonces*

$$|\mu_k(A) - \mu_k(B)| \leq \|A - B\|_2 \quad (1 \leq k \leq p).$$

Demostración. Sea $1 \leq k \leq p$. Sea $C \in \mathbb{R}^{m\times n}$ de rango menor que k. Entonces por la Nota 2.22 y la desigualdad triangular se tiene

$$\mu_k(B) \leq \|B - C\|_2 \leq \|A - C\|_2 + \|A - B\|_2 ,$$

de donde tomando mínimos respecto de C se obtiene

$$\mu_k(B) \leq \mu_k(A) + \|A - B\|_2 .$$

Intercambiando los papeles de A y B llegamos a

$$\mu_k(A) \leq \mu_k(B) + \|A - B\|_2 ,$$

y esto prueba la tesis del enunciado. □

Obsérvese que el resultado anterior establece la 'estabilidad', o dependencia continua, del concepto de valor singular de una matriz: pequeña perturbación en el dato A se transmite en una perturbación en los valores singulares del dato incluso del mismo orden.

Ejercicio 2.19. *Sea $A \in \mathbb{R}^{m\times n}$ y sea $\{A_r\}_{r\in N} \subset \mathbb{R}^{m\times n}$ una sucesión convergente a A. Entonces*

$$\lim_{r\to\infty} \mu_k(A_r) = \mu_k(A) \quad (1 \leq k \leq p).$$

Ejercicio 2.20. *El conjunto de las matrices de $\mathbb{R}^{m\times n}$ de rango completo, ($rank(A) = \min(m,n)$), es abierto en $\mathbb{R}^{m\times n}$. En efecto, si $A^{(s)} \in \mathbb{R}^{m\times n} \to A$, cuando $s \to \infty$, y $rank(A^{(s)}) < p = \min(m,n)$, entonces*

$$\mu_p(A) \leq \left\|A - A^{(s)}\right\|_2 \to 0,$$

es decir, el límite A es también de rango deficiente, luego el conjunto indicado es abierto.

Ejercicio 2.21. *El conjunto de matrices de rango completo es denso en $\mathbb{R}^{m\times n}$. Sea $A \in \mathbb{R}^{m\times n}$ de rango deficiente. Consideramos una SVD de la matriz, $A = U\Sigma V'$, y tomamos $A^{(k)} = U\Sigma^{(k)}V'$ donde $\Sigma^{(k)}$ es la matriz diagonal obtenida al sustituir las entradas nulas de la diagonal principal de Σ por el valor $1/k$, para $k \in \mathbb{N}$. Sabemos que $A^{(k)}$ tiene rango completo y además $\left\|A - A^{(k)}\right\|_2 = 1/k$, lo que prueba la densidad mencionada.*

Ejercicio 2.22. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica. Sean $\lambda_{\min}(A) = \min\{\lambda : \lambda \in \sigma(A)\}$ y $\lambda_{\max}(A) = \max\{\lambda : \lambda \in \sigma(A)\}$. Entonces*

$$\lambda_{\min} \|x\|_2^2 \leq x'Ax \leq \lambda_{\max} \|x\|_2^2 , \quad \forall x \in \mathbb{R}^n.$$

Como consecuencia pruébese que para toda $A \in \mathbb{R}^{n\times n}$ se tiene

$$\|A\|_2 = \sqrt{r(A'A)}.$$

Si además A es simétrica, entonces

$$\|A\|_2 = r(A).$$

Ejercicio 2.23. *Para toda norma matricial $|\bullet|$ existe $\alpha > 0$ tal que $\alpha\,|\bullet|$ es una norma multiplicativa. Basta tener en cuenta la equivalencia de normas con una norma multiplicativa como por ejemplo $\|\|_2$, y exigir que se cumpla la multiplicatividad usando la multiplicatividad de $\|\|_2$.*

Ejercicio 2.24. *Para toda norma matricial multiplicativa, $\|\|$, existe una norma vectorial, $||$, tal que la norma matricial subordinada, que denotamos igual, cumple*

$$|A| \leq \|A\| \quad (A \in K^n).$$

(La idea es definir $|x| = \|[x\,0\,\ldots\,0]\|$, y observar que se cumple $|Ax| \leq \|A\|\,|x|$.)

Ejercicio 2.25. *El límite de matrices ortogonales es ortogonal.*

Capítulo 3

Solución Numérica de Sistemas de Ecuaciones Lineales

En lo que sigue asumiremos, si no decimos lo contrario, que $A \in \mathbb{R}^{n\times n}$ es regular y que $b \in \mathbb{R}^n$. Además, consideraremos la norma matricial subordinada a una norma vectorial real $\|\|$, y la denotaremos de la misma forma.

El objetivo de este tema es analizar la solución numérica, que no la exacta, del sistema $Ax = b$, estableciendo pautas que permitan decidir si es una solución numérica aceptable o no, a la par que analizaremos como viene afectada ésta bajo perturbaciones de los datos del problema: A y b.

Como consecuencia de trabajar con aritmética de precisión limitada, al aplicar por ejemplo un método directo como eliminación de Gauss, obtendremos una aproximación $\widehat{x}$ de la solución x. Una primera cuestión es decidir si el hecho de que el residuo $r_{\widehat{x}} := A\widehat{x} - b$ sea pequeño en relación con el tamaño de b, implica que $\widehat{x}$ y x están relativamente cerca. Y viceversa, si el error relativo cometido al aproximar la solución es pequeño, ¿podemos afirmar que el error relativo entre $A\widehat{x}$ y b es pequeño? En el caso de ecuaciones no lineales es conocido que no son ciertas las citadas implicaciones, pero podría resultar que el caso lineal fuera diferente. Veamos unos ejemplos que clarifican la situación.

Ejemplo 3.1. *Sea* $0 < \varepsilon << 1$. *Consideremos* $A = [1+\varepsilon, 1-\varepsilon; 1-\varepsilon, 1+\varepsilon]$ *y* $b = [2; 2]$. *La solución solución exacta de* $Ax = b$ *es* $x = [1; 1]$. *Notar que la matriz de coeficientes es simétrica definida positiva y, como veremos, ésa es una clase de matrices para la que los métodos de aproximación de la solución que iremos introduciendo presentan menos problemas. Sea* $\widehat{x} = [2; 0]$. *Entonces,* $\|\widehat{x} - x\|_\infty / \|x\|_\infty = 1$ *mientras que* $\|r\|_\infty / \|b\|_\infty = \varepsilon$.

Ejemplo 3.2. *Consideramos la matriz* $A = [1.2969, 0.8648; 0.2161, 0.1441]$ *cuyo determinante es* 10^{-8}. *Además,* $A^{-1} = 10^8[0.1441, -0.8648; -0.2161, 1.2969]$. *Sea* $b = [0.8642; 0.144]$. *La solución de* $Ax = b$ *es* $x = [2; -2]$. *Consideramos* $\widehat{x} = [0.9911; -0.487]$, *entonces* $r_{\widehat{x}} = A\widehat{x} - b = [-10^{-8}; 10^{-8}]$. *Por lo tanto,*

tenemos

$$\frac{\|x-\widehat{x}\|_\infty}{\|x\|_\infty} = 1.513/2 = 0.7565$$

y

$$\frac{\|r\|_\infty}{\|b\|_\infty} = 10^{-8}/0.8642 < 1.16 * 10^{-8}.$$

Por otra parte, si tomamos ahora $b = [1; 1]$*, entonces la solución exacta es* $x = 10^8 * [-0.7207; 1.0808]$*. Sin embargo,* $A * [-10^8; 10^8] = 10^8 * [-0.4321; -0.072]$*, teniendo entonces*

$$\frac{\|x-\widehat{x}\|_\infty}{\|x\|_\infty} = 0.2793/1.0808 < 0.26$$

y

$$\frac{\|r\|_\infty}{\|b\|_\infty} > 4 * 10^7.$$

Proposición 3.1. *Para todo* $\widehat{x} \in \mathbb{R}^n$*, si* $r_{\widehat{x}} := A\widehat{x} - b$*, se tiene*

$$\frac{1}{\|A\|\,\|A^{-1}\|}\frac{\|r\|}{\|b\|} \leq \frac{\|x-\widehat{x}\|}{\|x\|} \leq \|A\|\,\left\|A^{-1}\right\|\frac{\|r\|}{\|b\|}.$$

Demostración. De entrada tenemos

$$\|x\| = \left\|A^{-1}b\right\| \leq \left\|A^{-1}\right\|\|b\|,$$

y

$$\|b\| = \|Ax\| \leq \|A\|\,\|x\|.$$

Por una parte,

$$\begin{aligned}\|x-\widehat{x}\| &= \left\|A^{-1}(b-A\widehat{x})\right\| \leq \left\|A^{-1}\right\|\|r\| \\ &\leq \left\|A^{-1}\right\|\|r\|\,\|A\|\,\|x\|\,/\,\|b\|,\end{aligned}$$

y por otra,

$$\begin{aligned}\|r\| &= \|A(x-\widehat{x})\| \leq \|A\|\,\|x-\widehat{x}\| \\ &\leq \|A\|\,\|x-\widehat{x}\|\,\left\|A^{-1}\right\|\|b\|\,/\,\|x\|,\end{aligned}$$

lo que prueba las desigualdades buscadas tras despejar convenientemente. □

Ejemplo 3.3. *Para las matrices de los dos ejemplos previos se tiene respectivamente*

$$\|A\|_\infty \left\|A^{-1}\right\|_\infty = \varepsilon^{-1}$$

y

$$\|A\|_\infty \left\|A^{-1}\right\|_\infty = 2.1617 * 1.513 * 10^8 > 3.2 * 10^8.$$

3.1. El número de condición

Definición 3.1. *Sea $A \in \mathbb{R}^{n\times n}$ regular. El número de condición de A respecto de la norma matricial $\|\|\|$ es el número no negativo*

$$k(A) = \|A\| \left\|A^{-1}\right\|.$$

Usaremos la notación $k_p(A)$ para indicar el número de condición asociado a la norma $\|\|\|_p$, para $p = 1, 2, \infty$.

Nota 3.1. *Más tarde comprobaremos que el número de condición es una buena medida de lo próxima que está la matriz A a una matriz singular (de hecho esta propiedad se probó para $\|\|\|_2$ como corolario del teorema de la mejor aproximación.*

El determinante de A nos permite decidir si A es regular o no, pero no tiene relación directa con el concepto introducido, como ponen de manifiesto el siguientes ejemplos.

Ejemplo 3.4. *(i) Consideramos la matriz $A_n \in \mathbb{R}^{n\times n}$ siguiente*

$$A_n = \begin{bmatrix} 1 & -1 & \dots & -1 \\ 0 & 1 & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

Entonces, $\det(A_n) = 1$ *y* $\|A_n\|_\infty = n$. *Es fácil comprobar que*

$$A_n^{-1} = \begin{bmatrix} 1 & 1 & 2 & \dots & 2^{n-2} \\ 0 & 1 & 1 & \dots & 2^{n-3} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

por lo que $\left\|A_n^{-1}\right\|_\infty = 1+1+2+4+\ldots+2^{n-2} = 2^{n-1}$, *de ahí que* $k_\infty(A_n) = n2^{n-1}$.

(ii) Sea $A = \varepsilon I_n$. Entonces $\det(A) = \varepsilon^n$ *mientras que* $k_\infty(A) = 1$.

Proposición 3.2. *Sean $A, B \in \mathbb{R}^{n\times n}$ regulares y $P \in \mathbb{R}^{n\times n}$ ortogonal. Sean $\|\|\|_\alpha$, $\|\|\|_\beta$ dos normas matriciales. Entonces*

(i) $k(A) \geq 1$,

(ii) $k(AB) \leq k(A)\, k(B)$,

(iii) $k(\alpha A) = k(A)$, $\alpha \neq 0$, $\alpha \in \mathbb{R}$,

(iv) $k(A) = k(A^{-1})$,

(v) $k_2(P) = 1$,

(vi) $k_2(A) = k_2(AP) = k_2(PA)$,

(vii) $\exists c_1, c_2 > 0$, *dependiendo sólo de las normas matriciales consideradas, tales que*

$$c_1 k_\beta(A) \leq k_\alpha(A) \leq c_2 k_\beta(A).$$

Si $\alpha \in \mathbb{C}$ cumple que $|1-\alpha| < 1$, entonces sabemos que la serie geométrica $\sum_{k=0}^{\infty}(1-\alpha)^k$ es convergente y que su suma es α^{-1}. Esto sugiere el siguiente resultado análogo para matrices:

Teorema 3.1. *Si $\|I_n - A\| < 1$, entonces A es regular con*

$$A^{-1} = \sum_{k=0}^{\infty}(I_n - A)^k,$$

y

$$\left\|A^{-1}\right\| \leq \frac{1}{1 - \|I_n - A\|}.$$

En particular,

$$k(A) \leq \frac{\|A\|}{1 - \|I_n - A\|}.$$

Además,

$$\left\|A^{-1} - \sum_{k=0}^{m}(I_n - A)^k\right\| \leq \frac{\|I_n - A\|^{m+1}}{1 - \|I_n - A\|} \quad (m \in \mathbb{N}).$$

Demostración. Sea $S_m = \sum_{k=0}^{m}(I_n - A)^k$, $m \in \mathbb{N}$. Entonces para $m > m'$, $m, m' \in \mathbb{N}$, se tiene

$$\|S_m - S_{m'}\| = \left\|\sum_{k=m'+1}^{m}(I_n - A)^k\right\| \leq \sum_{k=m'+1}^{m}\|I_n - A\|^k$$

de donde $\{S_m\}_{m\in\mathbb{N}}$ es una sucesión de Cauchy en $\mathbb{R}^{n\times n}$ y, por tanto, es convergente. Sea S su límite. Además,

$$(I_n - A)S_m = S_{m+1} - I_n,$$

de donde,

$$\|AS_m - I_n\| = \left\|(I_n - A)^{m+1}\right\| \leq \|I_n - A\|^{m+1} \overset{m\to\infty}{\to} 0.$$

Por consiguiente,

$$\lim_{m\to\infty} AS_m = I_n,$$

y por la unicidad del límite, tenemos $AS = I_n$, lo que asegura que A es regular con $A^{-1} = S$. Por último,

$$\|S_m\| = \left\|\sum_{k=0}^{m}(I_n - A)^k\right\| \leq \sum_{k=0}^{m}\|I_n - A\|^k \leq \frac{1}{1 - \|I_n - A\|},$$

y al hacer $m \to \infty$ se obtiene la cota de $\left\|A^{-1}\right\|$ anunciada. Probada la existencia de A^{-1} entonces

$$S_m = A^{-1}(I_n - (I_n - A)^{m+1}),$$

luego

$$\|S_m - S_{m'}\| \leq \frac{\|I_n - A\|^{m+1} + \|I_n - A\|^{m'+1}}{1 - \|I_n - A\|}, \quad m, m' \in \mathbb{N},$$

de donde al hacer $m' \to \infty$ se obtiene la restante desigualdad. □

Nota 3.2. *Observar que la norma vectorial puede ser tanto real como compleja.*

Podemos aproximar A^{-1} mediante la suma parcial S_m. Sea $\varepsilon > 0$. Para que

$$\left\|A^{-1} - S_m\right\| \leq \varepsilon,$$

bastará con elegir

$$m = fix(\log((1 - \|I_n - A\|)\varepsilon)/\log(\|I_n - A\|).$$

Ejemplo 3.5. *Sea*

$$A = \begin{bmatrix} 1 & 0.2 & 0.1 \\ 0.2 & 1 & 0.1 \\ -0.1 & -0.1 & 1 \end{bmatrix}$$

cuya inversa obtenida con MATLAB como $inv(A)$ es

$$\begin{bmatrix} 1.0348 & -0.2152 & -0.0820 \\ -0.2152 & 1.0348 & -0.0820 \\ 0.0820 & 0.0820 & 0.9836 \end{bmatrix}$$

y que verifica $\|I_3 - A\|_\infty = 0.3$. Por tanto, A es regular. Si deseamos aproximar A^{-1} con un error en $\|\|_\infty$ menor o igual que 0.1, tomamos $m = fix(\log(0.7\varepsilon)/\log(0.3)) = 2$. Así,

$$\begin{aligned} A^{-1} &\approx I_3 + I_3 - A + (I_3 - A)^2 \\ &= \begin{bmatrix} 1.03 & -0.21 & -0.08 \\ -0.21 & 1.03 & -0.08 \\ 0.08 & 0.08 & 0.98 \end{bmatrix}. \end{aligned}$$

Tendríamos además que $k_\infty(A) \leq 1.3/0.7 < 1.86$.

Ejercicio 3.1. *Probar de nuevo que si A es estrictamente diagonalmente dominante, entonces es regular. Considerar $D = diag(diag(A))$ y aplicar el resultado anterior a la matriz $B = D^{-1}A$.*

Ejercicio 3.2. *Sea $|||$ una norma vectorial compleja, y denotemos de igual forma la norma matricial subordinada. Si $\lambda \in \mathbb{C}$ cumple $|\lambda| > \|A\|$, entonces $\lambda \notin \sigma(A)$ y*

$$(\lambda - A)^{-1} = \sum_{k=0}^{\infty} \lambda^{-k-1} A^k.$$

En particular, habríamos probado de nuevo que $r(A) \leq \|A\|$.

Ejercicio 3.3. *Probar que $\sigma(A)$ es un conjunto cerrado usando el resultado previo (la propiedad es evidente ya que $\sigma(A)$ es unión finita de cerrados; de hecho es compacto). Sea $\lambda \notin \sigma(A)$. Como*

$$\mu - A = (\lambda - A)((\mu - \lambda)(\lambda - A)^{-1} + I_n),$$

si $|\mu - \lambda| < \left\|(\lambda - A)^{-1}\right\|^{-1}$, entonces $\mu \notin \sigma(A)$.

Lema 3.1. *Sea $|||$ una norma vectorial y sean $u, v \in \mathbb{R}^n$ con u no nulo. Existe $C \in \mathbb{R}^{n\times n}$ tal que $Cu = v$ y $\|C\| = \|v\| / \|u\|$.*

Demostración. La demostración para una norma arbitraria se basa en resultados sobre conjuntos convexos y queda fuera de nuestro alcance. Daremos únicamente una prueba específica para la norma euclídea. Tomamos

$$C = \frac{vu'}{\|u\|_2^2} \in \mathbb{R}^{n\times n}.$$

Tenemos $Cu = v$ y además

$$\|C\|_2 = \max_{\|z\|_2=1} \frac{\|vu'z\|_2}{\|u\|_2^2} = \frac{\|v\|_2}{\|u\|_2^2} \max_{\|z\|_2=1} |u'z| = \frac{\|v\|_2}{\|u\|_2},$$

donde hemos utilizado la desigualdad de Cauchy-Schwarz que establece que $|u'z| \leq \|u\|_2 \|z\|_2$. □

Teorema 3.2. *Sean $|||$ una norma vectorial y $A \in \mathbb{R}^{n\times n}$ regular. Entonces*

$$\frac{1}{k(A)} = \min\{\frac{\|A - B\|}{\|A\|} : B \in \mathbb{R}^{n\times n} \text{ es singular}\}.$$

Demostración. En primer lugar, si $1 > \|A - B\| \left\|A^{-1}\right\| \geq \left\|I_n - A^{-1}B\right\|$, entonces $A^{-1}B$ es regular, por tanto también B sería singular. Así,

$$\frac{1}{\|A^{-1}\|} \leq \|A - B\|, \quad B \in \mathbb{R}^{n\times n} \text{ es singular.}$$

Sea $x \in \mathbb{R}^n$ con $\|x\| = 1$ tal que $\left\|A^{-1}\right\| = \left\|A^{-1}x\right\|$. Por el lema previo aplicado a $u = A^{-1}x / \left\|A^{-1}x\right\|$ y $v = x / \left\|A^{-1}x\right\|$, existe $C \in \mathbb{R}^{n\times n}$ tal que $Cu = v$ y $\|C\| = 1 / \left\|A^{-1}\right\|$. Tomamos $B = A - C$. Se tiene

$$\|A - B\| = 1 / \left\|A^{-1}\right\|$$

y

$$By = Ay - Cy = v - v = 0,$$

por lo que B es singular. □

Ejemplo 3.6. *Sea* $A = [1.2969, 0.8648; 0.2161, 0.1441]$ *para la que* $k_\infty(A) = 3.2706521 \times 10^8$. *Consideramos la matriz* $B = [1.2969, 0.8648; 0.21606354, 0.14407568]$ *que se obtiene al hacer en* A *el cambio* $A(2,:) = 0.1666 * A(1,:)$ *(*$0.1666 \approx 0.2161/1.2969$*). Tenemos* $\|A - B\|_\infty = \left\|[0, 0; 0.3646 * 10^{-4}, 0.2432 * 10^{-4}]\right\|_\infty = 0.6078 * 10^{-4}$. *Por el teorema anterior*

$$k_\infty(A) \geq \|A\|_\infty / \|A - B\|_\infty = 2.1617 * 10^4 / 0.6078 > 3.5 * 10^4.$$

Corolario 3.1. *Sea* $A \in \mathbb{R}^{n \times n}$ *triangular. Entonces*

$$k_p(A) \geq \frac{\|A\|_p}{\text{mín}_{1 \leq i \leq n} |A(i,i)|}, \quad p = 1, 2, \infty.$$

Demostración. Para $1 \leq k \leq n$ consideramos la matriz singular B_k determinada por $B_k(k,k) = 0$ y $B_k(i,j) = A(i,j)$ para el resto de índices. Por el teorema anterior,

$$\left\|A^{-1}\right\|_p^{-1} \leq \min_{1 \leq k \leq n} \|A - B_k\|_p = \min_{1 \leq k \leq n} |A(k,k)|,$$

lo que conduce a la acotación anunciada. □

Ejemplo 3.7. *Consideremos la matriz* A_n *del ejemplo 3.4. Por el corolario anterior*

$$n 2^{n-1} = k_\infty(A_n) \geq n.$$

Nota 3.3. *Vamos a introducir un algoritmo para aproximar el número de condición de matrices triangulares. Observar que las matrices la podemos factorizar en la forma* LU*, salvo un cambio de filas previo, y, por tanto, tiene interés este tipo de resultados. Si* $Ay = d$*, con* $\|d\| = 1$*, entonces* $\left\|A^{-1}\right\| \geq \|y\|$. *La idea es pues resolver el sistema triangular eligiendo convenientemente el término independiente* d. *Si nos centramos en el caso de* A *triangular superior, la solución del sistema se obtiene mediante*

$$y(i) = \left(d(i) - \sum_{k=i+1}^{n} A(i,k)\, y(k) \right) / A(i,i), \quad i = n-1, \ldots, 1.$$

Analicemos la situación para $\|\|_\infty$. *Para cada* i *elegiremos* $d(i) = -sign(\sum_{k=i+1}^{n} A(i,k)\, y(k))$, *entendiendo en el caso de que la suma sea cero que elegimos* $d(i) = \pm 1$. *Notar que* $\|d\|_\infty = 1$. *La elección* $d(n) = \pm 1$ *puede influir en el resultado obtenido: aplicar el algoritmo a la matriz*

$$A = [1, 0, 10, 10; 0, 1, -10, 10; 0, 0, 1, 0; 0, 0, 0, 1].$$

Para A_n *del ejemplo 3.4 se obtiene* $\widehat{k_\infty(A_n)} = k_\infty(A_n)$.

3.2. Perturbación de los datos de un sistema de ecuaciones lineales

Teorema 3.3. *Sean $A \in \mathbb{R}^{n\times n}$ regular y $\triangle A \in \mathbb{R}^{n\times n}$ tal que $\|\triangle A\| \left\|A^{-1}\right\| < 1$. Sean $b \in \mathbb{R}^n$, no nulo, $\triangle b \in \mathbb{R}^n$. Sea $x \in \mathbb{R}^n$, con $Ax = b$, y sea $y \in \mathbb{R}^n$, con $(A + \triangle A)y = b + \triangle b$. Entonces*

$$\frac{\|x-y\|}{\|x\|} \leq \frac{k(A)}{1 - \|\triangle A\| \, \|A^{-1}\|} \left(\frac{\|\triangle A\|}{\|A\|} + \frac{\|\triangle b\|}{\|b\|} \right).$$

Demostración. En primer lugar $A + \triangle A$ es regular ya que

$$A + \triangle A = A(I_n + A^{-1}\triangle A),$$

y por la Proposición 3.1 la matriz $I_n + A^{-1}\triangle A$ es regular, gracias a la condición requerida a $\triangle A$. Además,

$$\left\|(A + \triangle A)^{-1}\right\| \leq \frac{\left\|A^{-1}\right\|}{1 - \|\triangle A\| \, \|A^{-1}\|}.$$

Por tanto,

$$\begin{aligned} x - y &= A^{-1}b - (A + \triangle A)^{-1}(b + \triangle b) = (A^{-1} - (A + \triangle A)^{-1})b - (A + \triangle A)^{-1}\triangle b \\ &= (A + \triangle A)^{-1}\triangle A x - (A + \triangle A)^{-1}\triangle b = (A + \triangle A)^{-1}(\triangle A x - \triangle b), \end{aligned}$$

y en consecuencia, como $\|b\| \leq \|A\| \, \|x\|$, se tiene

$$\begin{aligned} \frac{\|x-y\|}{\|x\|} &\leq \frac{\left\|A^{-1}\right\|}{1 - \|\triangle A\| \, \|A^{-1}\|}(\|\triangle A\| + \frac{\|\triangle b\|}{\|x\|}) \\ &\leq \frac{k(A)}{1 - \|\triangle A\| \, \|A^{-1}\|}(\frac{\|\triangle A\|}{\|A\|} + \frac{\|\triangle b\|}{\|b\|}). \end{aligned}$$

□

Ejemplo 3.8. *Sea $0 < \varepsilon << 1$. Consideramos $A = [1, 0; 0, \varepsilon]$. Se tiene $k_\infty(A) = \varepsilon^{-1}$. Tomamos $b = [1, \varepsilon]'$. Entonces $A * [1, 1]' = b$. Es evidente que $A * [1 + \varepsilon, 1]' = b + [\varepsilon, 0]'$. Por el teorema anterior tendríamos*

$$\varepsilon = \frac{\|x-y\|_\infty}{\|x\|_\infty} \leq \varepsilon^{-1}\varepsilon = 1,$$

mostrándose que la acotación obtenida del error relativo en la solución puede ser muy burda. Por otra parte, $A * [1, 2]' = b + [0, \varepsilon]'$, y ahora el teorema de perturbación anterior nos proporciona

$$1 = \frac{\|x-y\|_\infty}{\|x\|_\infty} \leq \varepsilon^{-1}\varepsilon = 1,$$

es decir, se alcanza la cota.

Recordemos que un sistema numérico en punto flotante F en base β con t dígitos es el conjunto de números reales

$$F = \{\pm 0.d_1\, d_2\ d_t * \beta^e : 0 \leq d_i < \beta,\ 1 \leq i \leq t;\ d_1 \neq 0;\ e \in [e_{\text{mín}}, e_{\text{máx}}]\} \cup \{0\}.$$

El rango de F es $\pm[\beta^{e_{\text{mín}}-1}, (1-\beta^{-t})\beta^{e_{\text{máx}}}] \cup \{0\}$. La unidad máquina para este sistema si se utiliza redondeo es $u = \beta^{1-t}/2$ (β^{1-t} en el caso de utilizar truncamiento).

Si un número real x está en el rango de F, éste se representa en la máquina mediante $fl(x) \in F$, cumpliéndose

$$fl(x) = x(1+\delta) \text{ con } |\delta| \leq u.$$

Las operaciones aritméticas, que representamos en forma general como $\odot \in \{+, -, *, /\}$, se implementan como $\odot^*$ de modo que se cumpla

$$x \odot^* y = fl(x \odot y).$$

En el tema siguiente comenzaremos a introducir técnicas de aproximación de la solución de $Ax = b$ mediante el uso de computadoras. Éstas requeriran el almacenamiento de los datos A y b, lo que conllevará que, en el mejor de los casos, estemos tratando de resolver con aritmética exacta un sistema perturbado próximo al original en términos de la unidad máquina. Tendríamos como matriz de coeficientes $A + \triangle A$ con $|\triangle A(i,j)| \leq u\,|A(i,j)|$, $1 \leq i, j \leq n$. Por tanto,

$$\|\triangle A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |\triangle A(i,j)| \leq u\,\|A\|_\infty\,.$$

De igual modo tendríamos un término independiente aproximado $b + \triangle b$, con $\|\triangle b\|_\infty \leq u\,\|b\|_\infty$. Supongamos que podemos obtener de forma exacta y cumpliendo el sistema perturbado. Si $uk_\infty(A) < 1$, entonces

$$\|\triangle A\|_\infty \left\|A^{-1}\right\|_\infty \leq u\,\|A\|_\infty \left\|A^{-1}\right\|_\infty < 1,$$

luego por el teorema anterior,

$$\frac{\|x-y\|_\infty}{\|x\|_\infty} \leq \frac{2uk_\infty(A)}{1-uk_\infty(A)}.$$

Si para simplificar la cota asumimos que $uk_\infty(A) \leq 1/2$, entonces obtenemos que en el caso ideal descrito

$$\frac{\|x-y\|_\infty}{\|x\|_\infty} \leq 4uk_\infty(A).$$

Parece pues sensato hablar de matriz mal o bien condicionada en términos del tamaño, grande o pequeño, de $uk_\infty(A)$. Así, no tiene sentido criticar un algoritmo que produzca una solución aproximada alejada de la real si

la matriz está mal condicionada en relación con la precisión de la máquina ($uk_\infty(A) \approx 1$).

Al aplicar un método directo para aproximar la solución del sistema usando un sistema numérico en punto flotante, lo que encontraremos es $\widehat{x}$ que verifica exactamente un sistema perturbado $(A + \triangle A)y = b$ (véase Teorema siguiente), pero no podemos esperar tener la situación anterior $\|\triangle A\|_\infty \leq u \|A\|_\infty$, debido a la aparición y posterior propagación de errores de redondeo. En general, sólo podemos esperar tener una cota del tipo $\|\triangle A\|_\infty \leq \phi(n)u \|A\|_\infty$.

Teorema 3.4. *Sea $r = A\widehat{x} - b$. Existe $\triangle A \in \mathbb{R}^{n\times n}$ tal que $(A+\triangle A)\widehat{x} = b$ y $\|\triangle A\| = \|r\| / \|\widehat{x}\|$. No existe otra matriz $\triangle A$ de menor norma cumpliéndose $(A + \triangle A)\widehat{x} = b$.*

Demostración. La condición $(A + \triangle A)\widehat{x} = b$ se cumple si, y sólo si, $\triangle A\widehat{x} = -r$. Esta relación implica que $\|\triangle A\| \geq \|r\| / \|\widehat{x}\|$. Para probar su existencia aplicamos el lema 3.1 con $x = \widehat{x}$ e $y = -r$. □

Nota 3.4. *Usando la matriz anterior$\triangle A$ tenemos $\widehat{x} - x = -A^{-1}\triangle A\widehat{x}$, de donde*

$$\|x - \widehat{x}\| \leq \left\|A^{-1}\right\| \|\triangle A\| \|\widehat{x}\| . = \left\|A^{-1}\right\| \|r\| ,$$

que coincide con la cota ya obtenida.

Capítulo 4

Métodos directos

En este tema introduciremos algunas técnicas directas para aproximar la solución de un sistema de ecuaciones lineales, todas ellas basadas en esencia en el conocido método de eliminación de Gauss, aunque veremos alguna alternativa mediante el uso de transformaciones ortogonales en el tema de sistemas sobredeterminados. El adjetivo de directas hace referencia a que si pudieramos trabajar con precisión ilimitada, obtendríamos en un número finito de pasos la solución exacta. Una forma de mejorar un método directo consistirá en idear una alternativa que reduzca el número de operaciones aritméticas que debemos realizar.

Por contra, los métodos iterativos, que veremos en un tema posterior, consistirán en generar una sucesión de vectores que converja, teóricamente, a la solución exacta, por lo que no podemos esperar nunca obtener la solución exacta al quedarnos con un término de la sucesión como aproximación. Estos últimos suelen utilizarse para sistemas dispersos de gran tamaño, es decir, con un gran número de ceros.

4.1. Uso de determinantes. Sistemas triangulares

Regla de Cramer

Se obtiene la solución de $Ax = b$ mediante

$$x(i) = \frac{\det([A(:,1), \ldots, A(:,i-1),\, b,\, A(:,i+1), \ldots, A(:,n)])}{\det(A)}, \quad 1 \leq i \leq n.$$

Tenemos pues una fórmula de indudable valor teórico que nos proporciona la solución, pero que se desvelará como totalmente inservible para aproximar al solución por el elevado número de operaciones aritméticas que involucra.

El cálculo del determinante de $A \in \mathbb{R}^{n\times n}$ mediante la fórmula que recordamos en el Tema 2, requiere el cálculo de $n! - 1$ sumas y $n!n$ productos, es decir, $n!(n+1)-1$ flops. Así, el uso de la fórmula de Cramer, cálculo de $n+1$

determinantes y n cocientes, requiere un total de $(n!(n+1)-1)(n+1)+n = n!(n+1)^2-1$ flops. Por ejemplo, para $n=10$ alrededor de 10^8 flops (MATLAB no cálcula determinantes mediante la fórmula comentada, si no triangulizando previamente). Veremos que eliminación de Gauss requerirá en este caso menos de 10^3 flops.

Obtención de A^{-1}

Como ya se recordó, podemos obtener A^{-1} mediante determinantes, pero por los comentarios previos es ya evidente que dicha fórmula es sólo interesante desde el punto de vista teórico. Por otra parte, determinar la inversa es equivalente a la resolución de n sistemas de ecuaciones. En efecto, como $AA^{-1} = I_n$, entonces

$$AA^{-1}(:,j) = e_j, \quad 1 \leq j \leq n.$$

Por tanto, no tiene sentido práctico intentar hallar A^{-1} y luego tomar $x = A^{-1}b$, a menos de que se disponga ya de A^{-1}, pues en dicho caso sólo necesitaríamos realizar $2n^2 - n$ flops para calcular $A^{-1}b$.

Sistemas triangulares

La situación más sencilla corresponde al caso de matrices diagonales en el que obtenemos la solución realizando n divisiones. Nos centraremos en los sistemas que tienen por matriz de coeficientes una matriz banda con ancho de banda $[p,0]$, $0 \leq p \leq n-1$, para los que introducimos el método de remonte (descenso para triangulares inferiores).

Algoritmo 4.1 (Método de remonte para matrices banda $[p,0]$)**.** *Datos:* $A, b, p(opcional)$

$b(n) = b(n)/A(n,n)$
$for\ k = n-1:-1:n-p$
  $b(k) = (b(k) - A(k, k+1:n) * b(k+1:n))/A(k,k) \quad (*)$
end
$for\ k = n-p-1:-1:1$
  $b(k) = (b(k) - A(k, k+1:k+p) * b(k+1:k+p))/A(k,k) \quad (**)$
end

Notar que la parte correspondiente a los dos for se puede compactar escribiendo

$for\ k = n-1:-1:1$
  $s = \text{mín}\{k+p, n\}$
  $b(k) = (b(k) - A(k, k+1:s) * b(k+1:s))/A(k,k)$
end

Para contar el número de operaciones aritméticas a realizar con el anterior algoritmo, necesitamos el siguiente resultado ya conocido.

Lema 4.1. *Se cumplen:*

(i) $\sum_{k=1}^{n} k = n(n+1)/2$.

(ii) $\sum_{k=1}^{n} k^2 = n(n+1)(2n+1)/6$.

Demostración. Se puede probar el resultado por inducción sobre n. Es interesante recordar que las fórmulas (i)-(ii) se basan en el hecho de que $\sum_{k=1}^{n}(a_{k+1} - a_k) = a_{n+1} - a_1$. Basta ahora aplicar el valor de la suma telescópica a $(k+1)^2 - k^2 = 2k+1$ para (i)- y $(k+1)^3 - k^3 = 3k^2 + 3k + 1$ para (ii).

Para probar (i) tambien se pueden contar los elementos de una matriz de tamaño n, n^2, sumando los que hay en las diagonales. □

Hemos de realizar los siguientes flops: para la parte (*), $n-k$ productos, $n-k$ sumas y 1 división, es decir

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

y para la parte (**), p productos, p sumas y 1 división, esto es

$$(n-p-1)(2p+1) = 2np - 2p^2 - 3p + n - 1$$

lo que da un total de $2np - p^2 + n - p$ si tenemos en cuenta el cociente aún no contabilizado. En particular, para el caso de $p = n-1$ necesitaríamos un total de n^2 flops.

Observar como se ha reducido drásticamente el número de operaciones necesarias para encontrar la solución aprovechando la estructura especial de la matriz.

4.2. El método de eliminación de Gauss (MEG)

Sobre las ecuaciones del sistema $Ax = b$ vamos a realizar dos tipos de operaciones básicas que transforman el sistema en uno nuevo equivalente, queriendo indicar con ello que ambos tienen una única solución y ésta coincide. Sobre la matriz ampliada $[A, b]$ realizaremos las siguientes operaciones:

(i) Intercambiar dos filas entre si.

(ii) Sustituir una fila por ella misma más un múltiplo de otra fila.

Por las propiedades del determinante la operación (i) cambia el signo del determinante y la operación (ii) no lo modifica, por lo que la nueva matriz de coeficientes sigue siendo regular, pudiendo asegurarse entonces la existencia y unicidad de solución, y está claro que la solución del sistema original verifica el nuevo sistema.

Antes de dar la descripción detallada del método de eliminación de Gauss en el que realizamos las operaciones descritas de forma apropiada para conseguir que la nueva matriz de coeficientes sea triangular superior y poder

resolver el sistema equivalente utilizando el método del remonte. Notar que el proceso permite calcular el determinante de una matriz si multiplicamos el deteminante de la matriz de coeficientes final por -1 elevado al número de cambio de filas realizado.

Ejemplo 4.1. *Consideremos el sistema cuya matriz ampliada es*

$$\begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 2 & -1 & 2 \\ 2 & 5 & 0 & 7 \end{bmatrix}$$

y cuya solución es $[\ 1 \quad 1 \quad 1\ ]'$.

$$\begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 2 & -1 & 2 \\ 2 & 5 & 0 & 7 \end{bmatrix} \overset{1^{aa}\longleftrightarrow 2^{aa}}{\rightarrow} \begin{bmatrix} 1 & 2 & -1 & 2 \\ 0 & 1 & 1 & 2 \\ 2 & 5 & 0 & 7 \end{bmatrix} \overset{3^{aa}-2*1^{aa}}{\rightarrow} \begin{bmatrix} 1 & 2 & -1 & 2 \\ 0 & 1 & 1 & 2 \\ 0 & 1 & 2 & 3 \end{bmatrix}$$

$$\overset{3^{aa}-2^{aa}\longleftrightarrow 3^{aa}}{\rightarrow} \begin{bmatrix} 1 & 2 & -1 & 2 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

de donde el determinante de la matriz de coeficientes es -1 *y se obtiene la solución ya adelantada aplicado remonte el último sistema obtenido.*

Describamos el método con más detalle. Introducimos la notación $A^{(1)} = A$ y $b^{(1)} = b$. Para simplificar la notación asumimos que $A^{(1)}(1,1) \neq 0$; en otro caso haríamos previamente un cambio de filas en $[A^{(1)}, b^{(1)}]$ para conseguir esta situación. Esto es factible gracias a la regularidad de $A^{(1)}$. Construimos ahora el sistema equivalente al anterior, $[A^{(2)}, b^{(2)}]$, mediante las operaciones ya descritas de manera que

$$A^{(2)}(i,1) = 0, \quad 2 \leq i \leq n.$$

Calculamos los multiplicadores

$$m_{i1} = -\frac{A^{(1)}(i,1)}{A^{(1)}(1,1)}, \quad 2 \leq i \leq n,$$

para luego tomar

$$A^{(2)}(i,:) = \begin{cases} A^{(1)}(1,:), \ i = 1, \\ A^{(1)}(i,:) + m_{i1}A^{(1)}(1,:), \ 2 \leq i \leq n, \end{cases}$$

y

$$b^{(2)}(i) = \begin{cases} b^{(1)}(1), \ i = 1, \\ b^{(1)}(i) + m_{i1}b^{(1)}(1), \ 2 \leq i \leq n. \end{cases}$$

Supongamos construido un sistema equivalente al original, $[A^{(k)}, b^{(k)}]$, con la propiedad deseada

$$A^{(k)}(i,j) = 0, \quad 1 \leq j \leq k-1, \ j+1 \leq i \leq n,$$

es decir, que ya se han realizado $k-1$ etapas del MEG. De nuevo supondremos que el llamado pivote de la etapa es no nulo, es decir, $A^{(k)}(k,k) = 0$, lo se puede conseguir en otro caso mediante un cambio de filas sin alterar la estructura de partida de la matriz de coeficientes. El proceso continua como si comenzaramos el proceso con el sistema $[A^{(k)}(k:n, k:n),\, b^{(k)}(k:n)]$ que es regular ya que

$$0 \neq \det(A^{(k)}) = \det(A^{(k)}(k:n, k:n)) \prod_{i=1}^{k-1} A^{(k)}(i,i).$$

Calculamos los multiplicadores

$$m_{ik} = -\frac{A^{(k)}(i,k)}{A^{(k)}(k,k)}, \quad k+1 \leq i \leq n,$$

para luego tomar

$$A^{(k+1)}(i,:) = \begin{cases} A^{(k)}(i,:), \ 1 \leq i \leq k, \\ A^{(k)}(i,:) + m_{ik} A^{(k)}(k,:), \ k+1 \leq i \leq n, \end{cases}$$

y

$$b^{(k+1)}(i) = \begin{cases} b^{(k)}(1), \ 1 \leq i \leq k, \\ b^{(k)}(i) + m_{ik} b^{(k)}(k), \ k+1 \leq i \leq n. \end{cases}$$

Se obtiene así un sistema equivalente al anterior, $[A^{(k+1)},\, b^{(k+1)}]$, y por tanto equivalente al inicial, con la propiedad

$$A^{(k+1)}(i,j) = 0, \quad 1 \leq j \leq k,\ j+1 \leq i \leq n.$$

En definitiva, tras $n-1$ etapas, con la mecánica ya descrita, obtenemos un sistema equivalente triangular superior, $[A^{(n)},\, b^{(n)}]$, que podemos resolver por remonte.

Algoritmo 4.2 (MEG sin pivote)**.** *Datos:* A, b

$W = [A, b]$
for $k = 1 : n-1$
  for $i = k+1 : n$
    $W(i,k) = W(i,k)/W(k,k)$
    $W(i, k+1 : n+1) = W(i, k+1 : n+1) - W(i,k) * W(k, k+1 : n+1)$
  end
end
$W(:, n+1) =$*remonte(*$W(:, 1:n)$,$W(:, n+1)$*)*

Contemos ahora el número de flops requerido por el algoritmo propuesto. En la etapa k-ésima realizamos

$$(n-k)(2(n-k+1)+1) = 2(n-k)^2 + 3(n-k)$$

flops, lo que da lugar a un total de

$$2\sum_{k=1}^{n-1} k^2 + 3\sum_{k=1}^{n-1} k + n^2 = \frac{2}{3}n^3 + \frac{3}{2}n^2 - \frac{7}{6}n.$$

Podemos resolver sistemas de ecuaciones con la misma matriz de coeficientes mediante el MEG. Como la parte más costosa es la de triangulizar, es conveniente considerar una matriz ampliada con todos los términos independientes y repetir el proceso descrito, finalizando aplicando remonte a cada uno de los correspondientes términos independientes. Con esta idea podemos obtener la inversa de A. El algoritmo, sin contemplar el posible intercambio de filas para elegir el pivote, lo escribimos a continuación.

Algoritmo 4.3 (Sistemas simultáneos sin pivote)**.** *Datos:* $A, b_1, \ldots, b_s$

$W = [A, b_1, \ldots, b_s]$
for $k = 1 : n-1$
  for $i = k+1 : n$
    $W(i,k) = W(i,k)/W(k,k)$
    $W(i, k+1 : n+s) = W(i, k+1 : n+s) - W(i,k) * W(k, k+1 : n+s)$
  end
end
for $k = 1 : s$
  $W(:, n+s) =$*remonte*$(W(:, 1:n), W(:, n+s))$
end

Calculemos el número de flops:

$$\begin{aligned}\sum_{k=1}^{n-1}(n-k)(2(n-k+s)+1) + sn^2 &= 2\sum_{k=1}^{n-1} k^2 + (2s+1)\sum_{k=1}^{n-1} k + sn^2 \\ &= \frac{4n^3 + (12s-3)n^2 - (6s+1)n}{6}.\end{aligned}$$

Por tanto, el cálculo de $A^{-1}(s = n)$ requiere $(16n^3 - 9n^2 - n)/6$ flops.

Ahora veremos que la situación se puede mejorar para las matrices banda. Si no es necesario el intercambio de filas, como veremos sucede con ciertos tipos de matrices, el proceso de triangularización conserva la estructura de la banda superior, por lo que podemos aplicar remonte específico para matrices banda, y a la vez para anular la banda inferior podemos realizar un número menor de operaciones que en el caso general.

Algoritmo 4.4 (Matrices banda con ancho de banda [p,q])**.** *Datos:* A, b, p, q

$W = [A, b]$
for $k = 1 : n-1$
  for $i = k+1 : min(k+q, n)$
    $W(i,k) = W(i,k)/W(k,k)$

$t = min(k+p, n)$
$W(i, k+1:t) = W(i, k+1:t) - W(i,k) * W(k, k+1:t)$
$W(i, n+1) = W(i, n+1) - W(i,k) * W(k, n+1)$
end
end
$W(:, n+1) =$*remonte(*$W(:, 1:n)$,$W(:, n+1)$*)*

Vamos a contar el número de flops necesario en el caso de matrices tridiagonales ($p = q = 1$). La triangularización requiere $5(n-1)$ flops y el remonte específico $3n-2$, lo que da un total de $8n-7$, que se compara muy favorablemente con la situación general.

4.2.1. Elección del pivote

Comencemos profundizando sobre la situación en la que el pivote 'natural' de la etapa k-ésima, $1 \leq k \leq n-1$, del MEG es nulo, es decir, $A^{(k)}(k,k) = 0$. Vamos a justificar que existe un elemento no nulo $A^{(k)}(i(k), k) = 0$ para algún $k+1 \leq i(k) \leq n$. Suponemos que en las etapas anteriores no nos hemos encontrado con esta situación. Como

$$A^{(k)}(i,j) = 0, \quad 1 \leq j \leq k-1,\ j+1 \leq i \leq n,$$

es sencillo mostrar, desarrollando el determinante por columnas y un proceso de inducción, que

$$\det(A^{(k)}) = \det(A^{(k)}(k:n, k:n)) \prod_{i=1}^{k-1} A^{(k)}(i,i),$$

de donde $\det(A^{(k)}(k:n, k:n)) \neq 0$ ya que $\prod_{i=1}^{k-1} A^{(k)}(i,i) \neq 0$ y $A^{(k)}$ es regular. Desarrollando el último determinante por los elementos de su primera columna se obtiene el resultado deseado por redución al absurdo.

Ejercicio 4.1. $\det(\begin{bmatrix} A & B \\ 0 & C \end{bmatrix}) = \det(A)\det(C)$, *siendo A y C matrices cuadradas.*

Veamos que el MEG puede presentar problemas cuando trabajamos con aritmética de precisión limitada.

Ejemplo 4.2. *Sean $\varepsilon \neq 1$, $A = \begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix}$ y $b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Apliquemos MEG:*

$$\begin{bmatrix} \varepsilon & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix} \overset{m_{21}=\varepsilon^{-1}}{\rightarrow} \begin{bmatrix} \varepsilon & 1 & 1 \\ 0 & 1-\varepsilon^{-1} & 2-\varepsilon^{-1} \end{bmatrix}$$

y por tanto la solución del sistema lineal es $x = \begin{bmatrix} 1/(1-\varepsilon) & (2\varepsilon-1)/(\varepsilon-1) \end{bmatrix}'$. Tomamos $\varepsilon = 10^{-9}$ y suponemos que trabajamos en un sistema en punto flotante decimal con mantisa de 8 dígitos. La solución exacta representada

en el sistema es $\overline{x} = \begin{bmatrix} 1 & 1 \end{bmatrix}'$ *que es lo que desearíamos obtener, pero el sistema equivalente al que llegaríamos es*

$$\begin{bmatrix} \varepsilon & 1 & 1 \\ 0 & -10^9 & -10^9 \end{bmatrix}$$

cuya solución es $\begin{bmatrix} 0 & 1 \end{bmatrix}'$. *(Notar que* $k_\infty(A) = 4/(1-\varepsilon) \doteq 4$*, luego no es éste el problema)*

Pivote parcial (MEGPP): En la k-ésima etapa del MEG elegimos $k \leq i_0 \leq n$ tal que

$$\left|A^{(k)}(i_0, k)\right| = \max_{k \leq i \leq n} \left|A^{(k)}(i, k)\right|,$$

e intercambiamos las filas k-ésima e i_0-ésima de $A^{(k)}$, es decir, $A^{(k)}(i_0, k)$ es nuestra elección del pivote de la k-ésima etapa.

Ejemplo 4.3. *Volvemos al ejemplo anterior y aplicamos esta técnica. Obtenemos*

$$\begin{bmatrix} \varepsilon & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 2 \\ \varepsilon & 1 & 1 \end{bmatrix} \overset{m_{21}=\varepsilon}{\rightarrow} \begin{bmatrix} 1 & 1 & 2 \\ 0 & 1-\varepsilon & 1-2\varepsilon \end{bmatrix}$$
$$\doteq \begin{bmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \end{bmatrix},$$

obteniéndose por remonte $\overline{x}$.

Destaquemos que el MEGPP requiere un esfuerzo adicional de comparación. En concreto, en la etapa k*-ésima hay que comparar* $n-k$ *números, lo que da un total de*

$$n - 1 + \ldots + 1 = O(n^2/2)$$

comparaciones.

Ejemplo 4.4. *Aquí ponemos de manifiesto que no siempre el pivote parcial resuelve el problema de trabajar con aritmética no exacta. Aplicando MEGPP tenemos*

$$\begin{bmatrix} 1 & 10^9 & 10^9 \\ 1 & 1 & 2 \end{bmatrix} \overset{m_{21}=1}{\rightarrow} \begin{bmatrix} 1 & 10^9 & 10^9 \\ 0 & -10^9 & -10^9 \end{bmatrix},$$

cuya solución es $\begin{bmatrix} 0 & 1 \end{bmatrix}'$*, lejana a la solución exacta,* $\begin{bmatrix} 1/(1-10^{-9}) & (2*10^{-9}-1)/(10^{-9}-1) \end{bmatrix}'$*, que se representa en el sistema en punto flotante como* $\begin{bmatrix} 1 & 1 \end{bmatrix}'$.

Pivote total (MEGPT): En la k-ésima etapa del MEG elegimos $k \leq i_0, j_0 \leq n$ tales que

$$\left|A^{(k)}(i_0, j_0)\right| = \max_{k \leq i,j \leq n} \left|A^{(k)}(i, j)\right|,$$

e intercambiamos las filas k-ésima e i_0-ésima de $A^{(k)}$ y también las columnas k-ésima y j_0-ésima de la matriz que acabamos de obtener, es decir, $A^{(k)}(i_0, j_0)$ es nuestra elección del pivote de la k-ésima etapa.

Notar que el cambio de columnas transforma el sistema en otro equivalente, siempre y cuando intercambiemos también las correspondientes componentes de la solución.

Ejemplo 4.5. *Aplicando MEGPT al ejemplo anterior tenemos*

$$\begin{bmatrix} 1 & 10^9 & 10^9 \\ 1 & 1 & 2 \end{bmatrix} \to \begin{bmatrix} 10^9 & 1 & 10^9 \\ 1 & 1 & 2 \end{bmatrix} \overset{m_{21}=10^{-9}}{\to} \begin{bmatrix} 10^9 & 1 & 10^9 \\ 0 & 1 & 1 \end{bmatrix},$$

cuya solución es $\begin{bmatrix} 1 & 1 \end{bmatrix}'$.

Como suele suceder, la mejora que puede suponer el MEGPT conlleva un mayor esfuerzo, ya que el número de comparaciones requerido en este caso es

$$n^2 - 1 + \ldots + 2^2 - 1 = O(2n^3/3).$$

4.2.2. Descripción matricial del MEG

Comenzaremos expresando matricialmente las dos operaciones que realizamos sobre las filas de la matriz ampliada del sistema W.

(i) Intercambio de filas $i-$ésima y $j-$ésima: consideramos la permutación σ_{ij} transposición de i y j. Si consideramos la matriz

$$T(i,j) = P'_{\sigma_{ij}} = P_{\sigma_{ij}^{-1}} = P_{\sigma_{ij}},$$

ya sabemos que $Z = T(i,j)W$ es la matriz que obtiene al realizar el cambio de filas mencionado sobre W. Recordar que $\det(Z) = -\det(W)$.

(ii) Sustitución de la fila $i-$ésima por ella más un múltiplo de la fila $k-$ésima: sea $\alpha \in \mathbb{R}$; introducimos la matriz

$$E(i,k,\alpha) = I_n + \alpha e^{(i)} e^{(k)'},$$

que no es más que la matriz identidad en la que el elemento de la posición k, i se ha sustituido por α. Por tanto, $\det(E(i,k,\alpha)) = 1$. Además, es inmediato comprobar que

$$(E(i,k,\alpha)W)(s,:) = \begin{cases} W(s,:), & s \neq i, \\ W(i,:) + \alpha W(k,:), & s = i. \end{cases}$$

En particular, $E(k,i,\alpha)^{-1} = E(k,i,-\alpha)$. Además, si $s \neq i$ o siendo $s = i$ tenemos $j \neq k$, entonces

$$\begin{aligned} E(j,s,\beta)E(i,k,\alpha) &= I_n + \alpha e^{(i)} e^{(k)'} + \beta e^{(j)} e^{(s)'} + \alpha\beta e^{(j)} e^{(s)'} e^{(i)} e^{(k)'} \\ &= I_n + \alpha e^{(i)} e^{(k)'} + \beta e^{(j)} e^{(s)'}. \end{aligned}$$

Ahora ya estamos en disposición de describir el MEG:
1^a ^{a}etapa: Consideramos

$$P_1 = \begin{cases} I_n, & \text{si } W^{(1)}(1,1) \neq 0, \\ T(1,i_1), & \text{si } W^{(1)}(1,1) = 0 \text{ y } W^{(1)}(i_1,1) \neq 0 \ (2 \leq i_1 \leq n). \end{cases}$$

que permite elegir el pivote al hacer $B^{(1)} = P_1 W^{(1)}$. Ahora anulamos los elementos por debajo de la diagonal principal en la primera columna, es decir, tomamos

$$m_{i1} = -B^{(1)}(i,1)/B^{(1)}(1,1), \quad 2 \leq i \leq n,$$

y consideramos

$$E_1 = E(n,1,m_{n1}) \dots E(3,1,m_{31}) E(2,1,m_{21}).$$

Entonces el sistema equivalente

$$W^{(2)} = E_1 B^{(1)} = E_1 P_1 W^{(1)}$$

goza de la propiedad deseada.

Supongamos realizadas $k-1$ etapas del MEG, es decir, tenemos un sistema equivalente al original, con matriz ampliada $W^{(k)}$, cumpliendo

$$W^{(k)}(i,j) = 0, \quad 1 \leq j \leq k-1, \ j+1 \leq i \leq n.$$

k^a ^{a}etapa: Consideramos

$$P_k = \begin{cases} I_n, & \text{si } W^{(k)}(k,k) \neq 0, \\ T(k,i_k), & \text{si } W^{(k)}(k,k) = 0 \text{ y } W^{(k)}(i_k,k) \neq 0 \ (k+1 \leq i_k \leq n). \end{cases}$$

que permite elegir el pivote al hacer $B^{(k)} = P_k W^{(k)}$. Ahora consideramos

$$m_{ik} = -B^{(k)}(i,k)/B^{(k)}(k,k), \quad k+1 \leq i \leq n,$$

y tomamos

$$E_k = E(n,k,m_{nk}) \dots E(k+1,k,m_{k+1,k}).$$

Entonces el sistema equivalente al anterior

$$W^{(k+1)} = E_k B^{(k)} = E_k P_k W^{(k)},$$

presenta la característica

$$W^{(k+1)}(i,j) = 0, \quad 1 \leq j \leq k, \ j+1 \leq i \leq n.$$

En definitiva, tras realizar $n-1$ etapas del MEG tenemos un sistema triangular superior equivalente al original

$$W^{(n)} = E_{n-1} P_{n-1} \dots E_1 P_1 W^{(1)}.$$

Ejemplo 4.6. *Consideramos el sistema cuya matriz ampliada es*

$$W = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 1 & 2 & 0 & 1 \\ 3 & 2 & 1 & 1 \end{bmatrix}$$

y su solución es $\begin{bmatrix} 0 & 1/2 & 0 \end{bmatrix}'$. *Tendríamos*

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 & 1 \\ 1 & 2 & 0 & 1 \\ 3 & 2 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 0 & -4 & -2 & -2 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

de donde en particular obtenemos que el determinante de la matriz de coeficientes es -4.

Teorema 4.1. *Sea* $A \in \mathbb{R}^{n\times n}$. *Existe una matriz* $M \in \mathbb{R}^{n\times n}$ *regular tal que* MA *es triangular superior.*

Demostración. Aplicar el MEG tal y como se ha descrito, con la salvedad de que cuando en la etapa k, $1 \leq k \leq n-1$, suceda que

$$W^{(k)}(i,k) = 0, \quad k \leq i \leq n,$$

entonces esa etapa 'la saltaríamos', es decir, tomaríamos $E_k = P_k = I_n$. □

Nota 4.1. *Es muy fácil comprobar que la anterior matriz no es única, ya que por ejemplo podemos tener diferentes opciones de elegir el pivote. Otro razonamiento consistiría en tener en cuenta que el producto de matrices triangulares superiores es estable.*

Ejemplo 4.7. $\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

Ejercicio 4.2. *Proporcionar una alternativa del MEG para la matriz*

$$\begin{bmatrix} * & * & * & \dots & * \\ * & * & & & \\ * & & \ddots & & \\ \vdots & & & \ddots & \\ * & & & & * \end{bmatrix},$$

en la que los únicos elementos no nulos son los asteriscos, de forma que reduzcamnos drásticamente el número de flops necesario, suponiendo siempre que no es necesario pivotaje.

Intercambiar primera y última fila y después primera y última columna (con el consiguiente cambio de incógnitas), para aplicar MEG al sistema equivalente correpondiente.

Ejercicio 4.3. *Considerar la sucesión de Fibonacci*

$$y_{n+1} = y_n + y_{n-1}, \quad n = 1, 2, \dots ; \quad y_0 = 0, \quad y_1 = 1.$$

(La sucesión es : $0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, \dots$*). Para cada* $n = 0, 1, \dots$ *consideramos el sistema*

$$\begin{bmatrix} y_n & y_{n+1} \\ y_{n+1} & y_{n+2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} y_{n+2} \\ y_{n+3} \end{bmatrix}$$

El sistema tiene solución única $x = y = 1$ *ya que* $y_n y_{n+2} - y_{n+1}^2 = (-1)^{n+1}$ *(inducción). Comentar si el sistema está mal condicionado para* n *grande (los elementos de la matriz de coeficiente aumentan mientras que el determinante siempre tiene valor absoluto* 1*). Comprobar que para* $n = 10$*, si le sumamos al elemento* y_{n+2} *de la segunda ecuación* $\varepsilon = 0.018$ *la solución del nuevo sistema es* $[-159.2, 100]'$*, muy lejana de la original. Para* $\varepsilon = 1/55$ *el sistema resultante no tiene unicidad de solución.*

Ejercicio 4.4. *¿Hay dependencia continua respecto de los datos independientemente que la matriz esté bien o mal condicionada? Sí, aplicar el teorema de perturbación.*

Capítulo 5

Factorización LU de una matriz

La cuestión que nos planteamos consiste en la posibilidad de reescribir una matriz $A \in \mathbb{R}^{n\times n}$ en la forma LU con $L \in \mathbb{R}^{n\times n}$ triangular inferior y $U \in \mathbb{R}^{n\times n}$ triangular superior. De lograrse, resolver el sistema $Ax = b$ sería equivalente a resolver el sistema $Ly = b$ por remonte, seguido del sistema $Ux = y$ resuelto por descenso. El hecho de disponer de una factorización conveniente de la matriz de coeficientes, será útil en situaciones en las que haya que resolver varios sistemas con la misma matriz de coeficientes, ya que, como se verá, la parte más costosa de resolver el sistema recae en la obtención de la factorización.

Nota 5.1. *Veamos aplicaciones de tener factorizada la matriz en la forma LU. Básicamente, son situaciones en las que se requiere resolver sistemas con la misma matriz de coeficientes, y, como ya sabemos, es más económico utilizar la factorización en la forma ya indicada que aplicar reiteradamente MEG. De hecho, veremos que el costo de obtener la factorización es similar al de aplicar MEG, mientras que tras obtener la factorización resolver un sistema sólo conlleva $2n^2$ flops.*

(i) ***Mejora iterativa.***

Supongamos que hemos obtenido una aproximación x_0 de la solución del sistema $Ax = b$. El sistema se puede interpretar como el problema de obtener un cero de la función vectorial $f(x) \equiv Ax - b$, y por tanto podemos aplicar el método de Newton para intentar mejorar el valor conocido. En el caso escalar, el método consiste aproximar la función f por la recta tangente a la mismo en el punto $(x_0, f(x_0))$ y hallar su cero, lo que se corresponde con tomar $x_1 = x_0 - f(x_0)/f'(x_0)$ como nueva aproximación. En nuestro caso haríamos

$$r_0 = Ax - b, \quad Ay_0 = r,\ x_1 = x_0 - y_0,$$

es decir, en cada etapa del método de Newton tendríamos que resolver un sistema de ecuaciones con la misma matriz A de coeficientes. En la

práctica el cálculo del residuo debería realizarse en doble precisión. El proceso descrito tiene sentido cuando no somos capaces de aproximar bien la solución como consecuencia de la aritmética de precisión limitada y del mal condicionamiento del sistema. Podría pensarse también en aumentar la precisión en todo el MEG, pero esto sería más costoso.

Veamos un ejemplo. Si consideramos el sistema con matriz ampliada

$$[1, 10^{20}, 10^{20}; 1, 1, 2],$$

cuya solución exacta representada en el sistema en punto flotante del MATLAB es $[1; 1]$ *y aplicamos el operador* $\backslash$ *obtenemos como solución* $[0; 1]$*. Es inmediato comprobar que al aplicar mejora iterativa una sóla vez encontramos como nueva aproximación* $[1; 1]$*.*

(ii) *Sistemas del tipo* $A^p x = b$ *(*$p \in \mathbb{N}$*).*

Calcular un producto de dos matrices cuadradas de tamaño n requiere el cálculo de n^2 *productos escalares y que cada uno de ellos involucra* $2n-1$ *flops, luego en total necesitaríamos realizar* $2n^3 - n^2$ *flops. Así, para construir la matriz* A^p *se necesitan un total de* $(2n^3 - n^2)(p-1)$ *flops. Luego aplicaríamos MEG para resolver el sistema.*

Un alternativa consiste en resolver los p sistemas con la misma matriz de coeficientes

$$Ax_1 = b,\ Ax_2 = x_1, \ldots,\ Ax_p = x_{p-1},$$

con un total de $2n^2 p$ *flops, si partimos de una factorización LU de A. En este caso la solución es* x_p*.*

(iii) *Si* $A = LU$*,* $\|\|$ *es una norma matricial multiplicativa y* $k(A)$ *representa el número de condición de A respecto de* $\|\|$*, entonces*

$$k(A) \leq k(L)k(U),$$

por lo que podríamos acotar el tamaño de $k(A)$ *mediante buenos estimadores del número de condición para matrices triangulares, que es un problema en general más cómodo.*

5.1. Existencia y unicidad de la factorización LU

Comencemos probando un sencillo lema sobre matrices triangulares.

Lema 5.1. *Sean* $A, B \in \mathbb{R}^{n\times n}$ *triangulares inferiores (superiores). Entonces*

(i) $C = AB$ *es triangular inferior (superior) y*

$$C(i,i) = A(i,i)B(i,i), \quad 1 \leq i \leq n.$$

(ii) *Si A es regular, entonces A^{-1} es triangular inferior (superior) con*

$$A^{-1}(i,i) = 1/A(i,i), \quad 1 \leq i \leq n.$$

Demostración. Al ser A triangular inferior,

$$A(i,j) = 0, \quad 1 \leq i < j \leq n,$$

y lo mismo tenemos sobre los elementos de B, por lo que

$$C(i,j) = \sum_{k=1}^{n} A(i,k)B(k,j) = \sum_{k=1}^{i} A(i,k)B(k,j) = \begin{cases} 0, & 1 \leq i < j \leq n, \\ A(i,i)B(i,i), & i = j, \end{cases}$$

lo que confirma (i).

Por otra parte, cuando $A(i,i) \neq 0$, $1 \leq i \leq n$, existe A^{-1} y $AA^{-1}(:,j) = ej$, $1 \leq j \leq n$, sistemas de ecuaciones que podemos resolver por el método del descenso, obteniendo para todo $1 \leq j \leq n$ que $A^{-1}(k,j) = 0$, $1 \leq k \leq j-1$, $A^{-1}(j,j) = 1/A(j,j)$, lo que constituye el resultado deseado. □

Más tarde estudiaremos la existencia de una tal factorización. Respecto de la unicidad, es evidente que no se cumple de no imponer alguna condición adicional:

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

Definición 5.1. *Una factorización LU de $A \in \mathbb{R}^{n\times n}$ es aquella de la forma*

$$A = RS,$$

donde $R, S \in \mathbb{R}^{n\times n}$, R es triangular inferior con $R(i,i) = 1$, $1 \leq i \leq n$, y S es triangular superior.

Teorema 5.1. *La factorización LU de una matriz regular es única.*

Demostración. Supongamos que tenemos dos factorizaciones de A, es decir $L_1U_1 = A = L_2U_2$. Como A es regular, entonces U_i es regular para $i = 1, 2$. Así, tenemos $L_1^{-1}L_2 = U_1U_2^{-1}$ y por el lema anterior la primera matriz es triangular inferior con unos en la diagonal principal, y la segunda triangular superior, de ahí que al ser iguales coincidan con la matriz identidad, lo que se traduce en que $L_1 = L_2$ y $U_1 = U_2$. □

Nota 5.2. *Notar que no hay unicidad si la matriz A no es regular. Evidentemente, en esta situación U no podría ser regular. Por ejemplo tenemos*

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

Además, la existencia no está asegurada, basta considerar la matriz regular

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

En efecto, si

$$A = \begin{bmatrix} r & 0 \\ s & t \end{bmatrix} \begin{bmatrix} u & v \\ 0 & w \end{bmatrix},$$

entonces $ru = 0$, $su = 1$ y $rv = 1$, lo que es imposible al mismo tiempo.

Supongamos ahora que el MEG para la matriz regular A no ha requerido cambio de filas gracias a que los pivotes obtenidos son no nulos. Entonces obtenemos

$$A^{(n)} = E_{n-1} \dots E_1 A$$

con $A^{(n)}$ triangular superior y, para todo $k = 1, \dots, n-1$,

$$E_k = E(k, n, m_{nk}) \dots E(k, k+1, m_{k+1,k}),$$

donde

$$m_{ik} = -A^{(k)}(i,k)/A^{(k)}(k,k), \quad k+1 \leq i \leq n.$$

Si definimos $U =$ y $L = E_1^{-1} \dots E_{n-1}^{-1}$, entonces $A = LU$, y la existencia en este caso de la factorización estaría probada a falta de comprobar que L verifica las propiedades requeridas. Por el Lema previo sabemos que L tiene la forma requerida si recordamos las propiedades de las matrices $E(i,k,\alpha)$. No obstante, recordando que post-multiplicar una matriz por $E(i,k,\alpha)$ cambiaba la columna k-ésima de la misma por ella más α veces su columna i-ésima, es fácil comprobar que

$$\begin{aligned} L &= E(2,1,-m_{21}) \cdots E(n,1,-m_{n1}) E(3,2,-m_{32}) \cdots E(n,2,-m_{n2}) \cdots E(n,n-1,-m_{n,n-1}) \\ &= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -m_{21} & 1 & 0 & \cdots & 0 & 0 \\ -m_{31} & -m_{32} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -m_{n-1,1} & -m_{n-1,2} & -m_{n-1,3} & \cdots & 1 & 0 \\ -m_{n1} & -m_{n2} & -m_{n3} & \cdots & -m_{n,n-1} & 1 \end{bmatrix} \end{aligned}$$

lo que nos da idea de como obtener L y permite asegurar que L es triangular inferior con unos en la diagonal principal.

Lema 5.2. *Sean $A, B, C \in \mathbb{R}^{n \times n}$tales que $A = BC$ y B es triangular inferior o C es triangular superior. Entonces*

$$A_k = B_k C_k, \quad 1 \leq k \leq n.$$

Demostración. Bajo las condiciones indicadas tenemos

$$A_k(i,j) = \sum_{s=1}^{n} B(i,s)C(s,j) = \sum_{s=1}^{k} B(i,s)C(s,j) = \sum_{s=1}^{k} B_k(i,s)C_k(s,j) = (B_k C_k)(i,j),$$

para $1 \leq i, j \leq k \leq n$. □

Nota 5.3. *Evidentemente, el Lema previo no es cierto en general. Basta, por ejemplo, con plantearselo para matrices* 2×2.

Teorema 5.2 (Existencia factorización *LU*)**.** *Sea* $A \in \mathbb{R}^{n\times n}$ *regular. Entonces A es factorizable en la forma LU si, y sólo si, las submatrices*

$$A_k = A(1:k, 1:k), \quad 1 \leq k \leq n-1,$$

son regulares.

Demostración. Veamos que la condición es suficiente. Por el comentario previo, bastará con mostrar que podemos aplicar MEG sin intercambio de filas para elegir el pivote, y lo probaremos por inducción sobre la etapa. Como $A(1,1) = \det(A_1) \neq 0$, entonces en la primera etapa no es necesario pivotaje. Supongamos que lo propio sucede hasta la etapa $k-1$, es decir, hemos construido

$$A^{(k)} = E_{n-1} \dots E_1 A \equiv R^{(k)} A,$$

con

$$A^{(k)}(s,s) \neq 0, \quad 1 \leq s \leq k-1, \text{ y } A^{(k)}(i,j) = 0, \quad 1 \leq j \leq k-1, \ j+1 \leq i \leq n,$$

y veamos que sucede con la siguiente. Es fácil comprobar que

$$A_k^{(k)} := A^{(k)}(1:k, 1:k) = R_k^{(k)} A_k,$$

ya que $R^{(k)}$ es triangular inferior, y en consecuencia, puesto que $A_k^{(k)}$ es triangular superior, tenemos

$$\prod_{s=1}^{k} A^{(k)}(s,s) = \det(R_k^{(k)}) \det(A_k) = \det(A_k) \neq 0, \tag{5.1}$$

lo que permite concluir que $A^{(k)}(k,k) \neq 0$, haciendo innecesario un intercambio de filas en la etapa $k-$ésima.

Mostremos ahora que la condición es necesaria. Supongamos que $A = LU$ con las condiciones ya apuntadas. Tenemos

$$0 \neq \det(A) = \det(L) \det(U) = \prod_{s=1}^{n} U(s,s),$$

por lo que $U(s,s) \neq 0$, $1 \leq s \leq n$. Además, por ser L triangular inferior, $A_k = L_k U_k$, de donde

$$\det(A_k) = \det(L_k) \det(U_k) = \prod_{s=1}^{k} U(s,s) \neq 0, \quad 1 \leq k \leq n-1.$$

□

Nota 5.4. *Por todo lo visto, podemos afirmar que una matriz es factorizable LU si, y sólo si, podemos aplicar MEG sin necesidad de intercambiar filas. En consecuencia, disponemos de un criterio práctico para determinar si una matriz es factorizable o no: aplicar directamente MEG atendiendo a si es necesario o no algún intercambio de filas. En concreto, tenemos las equivalencias deducidas del siguiente esquema de demostración:*

$$\begin{array}{ccc} \textit{Factorizable LU} & \longleftarrow & \textit{MEG sin pivotaje} \\ \downarrow & \nearrow & \\ A_k \textit{ regulares} (1 \leq k \leq n-1) & & \end{array}$$

Ejemplo 5.1. *Las matrices simétricas definidas positivas son factorizables LU ya que* $\det(A_k) > 0$, $1 \leq k \leq n$.

De la demostración anterior se desprende un criterio práctico para determinar si una matriz, siendo simétrica, es defnida positiva, que consiste en aplicar el MEG a la matriz y comprobar que toda etapa el pivote obtenido es positivo (úsese 5.1).

Ejemplo 5.2. *Vamos a resolver el sistema*

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{bmatrix} x = \begin{bmatrix} 3 \\ 3 \\ 4 \end{bmatrix}$$

obteniéndo previamente la factorización LU de la matriz de coeficientes. Aplicamos MEG con la precaución de guardar en la parte triangular inferior de la matriz de coeficientes la matriz L y en la parte superior la matriz U:

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

Así,

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

y resolvemos el sistema

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 1 & 1 \end{bmatrix} y = \begin{bmatrix} 3 \\ 3 \\ 4 \end{bmatrix}$$

cuya solución es $[\ 3 \quad -3 \quad -2\ ]'$ *para resolver finalmente*

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

cuya solución es $[\ 1 \quad 1 \quad 1\ ]'$ *y es, como se observa sin dificultad, la solución del sistema planteado.*

Ejemplo 5.3. *Aplicando MEG a la matriz inicial para transformarla en una triangular superior, vemos que evidentemente no es factorizable LU, pero que si hacemos un cambio previo de filas, la nueva matriz si es factorizable, hecho que estableceremos en el teorema siguiente para cualquier matriz.*

$$A \equiv \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \overset{1^{aa} \Leftrightarrow 2^{aa}}{\rightarrow} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \overset{2^{aa} \Leftrightarrow 3^{aa}}{\rightarrow} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \equiv U,$$

luego existe una matriz permitación P tal que

$$PA \equiv \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} A = U = I_n U.$$

Analizemos ahora la situación general. Sea $A \in \mathbb{R}^{n \times n}$, regular o no. Por MEG sabemos que podemos conseguir

$$A^{(n)} = E_{n-1} P_{n-1} \dots E_1 P_1 A^{(1)},$$

en los términos ya establecidos. Recordar que si en una etapa todos los elementos por debajo de la diagonal principal son nulos, entonces saltamos la etapa, es decir, tomamos las correspondientes $E = P = I_n$, lo que incluye la situación que se presenta cuando A es singular. Vamos a intentar reescribir el producto $P_k E_j$, $k > j$, en la forma $\widetilde{E_j} P_k$. Para situarnos, supongamos que $P_k = T(k, i)$, $i > k$. La matriz $P_k E_j$ es la matriz E_j en la que se han intercambiado las filas $k-$ésima y la $i-$ésima. Por tanto, respecto de E_j simplemente hemos de intercambiar en la columna $j-$ésima los elementos m_{ij} y m_{kj}, anular los elementos de las posiciones (i, i) y (k, k), y convertir en unos los ceros de las posiciones (i, k) y (k, i). Postmultiplicar por una matriz transposición se traduce en intercambiar columnas, por lo tanto, la matriz $\widetilde{E_j}$ que resuelve el problema es la matriz E_j en la que se han intercambiado los elementos m_{ij} y m_{kj}, por lo que conserva la estructura original de matriz triangular inferior con unos en la diagonal principal. Nótese que

$$\widetilde{E_j}(:, s) = \begin{cases} E_j(:, s),\ 1 \leq s \leq n,\ s \neq j, \\ \widetilde{E_j}(:, j) = P_k E_j(:, j),\ s = j. \end{cases}$$

En resumidas cuentas, podemos cambiar el orden de los factores con facilidad cuando surja la necesidad de elegir el pivote.

Todo lo que acabamos de comentar conduce al siguiente resultado:

Teorema 5.3 (Factorización $PA = LU$)**.** *Sea $A \in \mathbb{R}^{n \times n}$. Existe una matriz permutación $P \in \mathbb{R}^{n \times n}$ tal que PA es factorizable LU.*

Demostración. Comprobemos por inducción sobre la etapa del MEG, con las ideas antes expuestas, que podemos obtener, tras las $n - 1$ etapas,

$$A^{(n)} = S_{n-1} \dots S_1 P_{n-1} \dots P_1 A,$$

donde $A^{(n)}$ es triangular superior y para todo $1 \leq i \leq n$, P_i es una matriz permutación, y S_i es una matriz triangular inferior con unos en la diagonal principal de forma que los elementos no nulos se pueden presentar sólo en la columna $i-$ésima. Entonces, el resultado queda probado tomando $P \equiv P_{n-1} \dots P_1$, $U \equiv A^{(n)}$ y $L = S_1^{-1} \dots S_{n-1}^{-1}$.

En la segunda etapa tenemos

$$\begin{aligned} A^{(3)} &= E_2 P_2 E_1 P_1 A^{(1)} = E_2 E_1^{(3)} P_2 P_1 A^{(1)} \\ &= E_2^{(3)} E_1^{(3)} P_2 P_1 A^{(1)}, \end{aligned}$$

donde $E_1^{(3)}$ es la matriz E_1 en la que hemos cambiado en su primera columna las filas que intercambia P_2 y $E_2^{(3)} = E_2$. Realizadas $k-1$ etapas de esta forma y reeescribiendo los factores de la forma indicada obtenemos

$$A^{(k)} = E_{k-1}^{(k)} \dots E_2^{(k)} E_1^{(k)} P_{k-1} \dots P_2 P_1 A^{(1)},$$

cumpliendo las condiciones requeridas en esta fase.

Realicemos la etapa k-ésima. Elegimos el pivote mediante una matriz permutación P_k. Tendríamos

$$\begin{aligned} B^{(k)} &\equiv P_k E_{k-1}^{(k)} \dots E_2^{(k)} E_1^{(k)} P_{k-1} \dots P_2 P_1 A^{(1)} \\ &= E_{k-1}^{(k+1)} \dots E_2^{(k+1)} E_1^{(k+1)} P_k P_{k-1} \dots P_2 P_1 A^{(1)}, \end{aligned}$$

donde $E_j^{(k+1)}$ se obtiene intercambiando en la columna j-ésima de $E_j^{(k)}$ las filas que intercambie P_k $(1 \leq j \leq k-1)$. Ahora anulamos los correspondientes elementos de $B^{(k)}$ en su columna k-ésima mediante

$$A^{(k+1)} \equiv E_k B^{(k)} \equiv E_k^{(k+1)} E_{k-1}^{(k+1)} \dots E_2^{(k+1)} E_1^{(k+1)} P_k P_{k-1} \dots P_2 P_1 A^{(1)}.$$

siendo $E_k \equiv E_k^{(k+1)}$ y $E_j^{(k+1)}$ la matriz $E_j^{(k)}$ en cuya columna j-ésima hemos aplicado la permutación P_k. Así, tras $n-1$ etapas obtenemos

$$A^{(n)} = E_{n-1}^{(n)} \dots E_2^{(n)} E_1^{(n)} P_n \dots P_2 P_1 A,$$

cumpliendo las condiciones requeridas. □

Nota 5.5. *Si tenemos $PA = LU$, entonces* $\det(A) = (-1)^k \prod_{i=1}^n U(i,i)$, *donde k es el número de transposiciones que hay en la correspondiente permutación (cambios de columnas que hay en P respecto de la identidad, o el número de cambios de filas realizado si hemos obtenido nosotros la factorización). Además, resolver el sistema $Ax = b$ se traduciría en resolver*

$$Ly = Pb, \quad Ux = y.$$

Insistamos de nuevo en que Pb no lo calculamos mediante un producto si no realizando los cambios en las componentes que provoca P.

En la práctica simplemente hay que realizar en la etapa $k-$ésima el cambio de filas que requiera la elección del pivote en la matriz en la que hemos ido guardando los multiplicadores, con el signo adecuado, en la parte estrictamente triangular inferior correspondiente a la $k-1$ primeras columnas y en cuya parte restante tenemos la matriz $A^{(k)}$. Respecto como obtener la matriz permutación, lo que haremos es anotar en cada etapa, en un vector que originalemente representa la identidad $(1:n)$, los cambios de filas que hemos realizado en dicha etapa, simplemente cambiando las componentes del vector correspondientes a las posiciones de las filas a cambiar. Finalmente la matriz P se obtiene realizando los cambios por filas anotados en el vector, es decir, si el vector es p, y tomamos $I = eye(n)$, entonces $P = I(p,:)$. Nótese que obtener P como matriz de permutación asociada a una permutación σ, es mucho más complicado puesto que es composición de transposiciones. Es fácil observar que en realidad $\sigma = p^{-1}$, entendiendo que ahora p representa una permutación.

Ejercicio 5.1. *Crear una función cuyo argumento sea una permutación s y que devuelva su inversa r.*

Sol.: $function \quad r = pinvn(s)$
$for \quad i = 1 : length(s)$
$r(s(i)) = i;$
end

Algoritmo 5.1 ($PA = LU$)**.** *Datos: A*

$p = 1 : n$
for $k = 1 : n-1$
elección del pivote: $p(k) \longleftrightarrow p(i_k)$ $(n \geq i_k \geq k)$
$A(k,:) \longleftrightarrow A(i_k,:)$ *(Intercambiar filas)*
for $i = k+1 : n$
$A(i,k) = A(i,k)/A(k,k)$
$A(i,k+1:n) = A(i,k+1:n) - A(i,k) * A(k,k+1:n)$
end
end

Salida: p, A (como alternativa podemos devolver P; en realidad, si luego queremos hacer Pb haríamos $b = b(p)$ de forma que no necesitamos P)

Ejemplo 5.4. *Consideremos el sistema de matriz ampliada*

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

Obtengamos la factorizacion de la matriz de coeficientes

$$\begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & 0 \\ 3 & 2 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 1 \\ 1 & 0 & -1 \\ 3 & -4 & -2 \end{bmatrix} \overset{2^{aa} \Leftrightarrow 3^{aa}}{\underset{p=(1,3,2)}{\rightarrow}} \begin{bmatrix} 1 & 2 & 1 \\ 3 & -4 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

En definitiva,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & 0 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 0 & -4 & -2 \\ 0 & 0 & -1 \end{bmatrix}.$$

Notar que $\det(A) = -4$. *Ahora resolvemos*

$$\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} y = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

cuya solución es $\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$, *y finalmente*

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & -4 & -2 \\ 0 & 0 & -1 \end{bmatrix} x = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

cuya solución es $\begin{bmatrix} 1/2 & -1/4 & 0 \end{bmatrix}$.

Por último contemplemos la opción de obtener una factorización LU empleando el MEG con pivote total. Tendríamos

$$A^{(n)} = E_{n-1}P_{n-1}\dots E_1P_1A^{(1)}Q_1\dots Q_{n-1},$$

donde la novedad es la aparición de las matrices transposición que permiten elegir el pivote en cada etapa intercambiando columnas si fuera necesario. Al igual que el caso anterior, podemos conmutar las matrices P_j y E_i realizando los oportunos cambios en E_i, lo que nos conduciría a

$$A^{(n)} = E_{n-1}^{(n-1)}\dots E_1^{(n-1)}P_{n-1}\dots P_1A^{(1)}Q_1\dots Q_{n-1},$$

de donde llegamos a la factorización

$$PAQ = LU,$$

si tomamos $P = P_{n-1}\dots P_1$, $Q = Q_1\dots Q_n$, $L^{-1} = E_{n-1}^{(n-1)}\dots E_1^{(n-1)}$ y $U = A^{(n)}$.

Si almacenamos todos los multiplicadores, sin signo, en la parte triangular inferior de A y vamos realizando los cambios necesarios de filas y de columnas (observar que estos últimos no afectan a la parte ya calculada de L), al final en la parte triangular inferior tendremos L y en la parte triangular superior U.

Nota 5.6. *Si tenemos* $PAQ = LU$, *entonces* $\det(A) = (-1)^{k+s}\prod_{i=1}^{n}U(i,i)$, *donde* k *es el número de cambios de fila y* s *el número de cambios de columna que hemos realizado. Además, resolver el sistema* $Ax = b$ *se traduciría en resolver*

$$Lz = Pb, \quad Uy = z,$$

y tomar $x = Qy$.

Algoritmo 5.2 ($PAQ = LU$)**.** *Datos: A*

$p = 1 : n$
$q = p$
for $k = 1 : n - 1$
Elección del pivote: $p(k) \longleftrightarrow p(i_k)$, $q(k) \longleftrightarrow q(j_k)$ $(n \geq i_k, j_k \geq k)$
$A(k,:) \longleftrightarrow A(i_k,:)$ *(Intercambiar filas)*
$A(:,k) \longleftrightarrow A(:,j_k)$ *(Intercambiar columnas)*
for $i = k + 1 : n$
$A(i,k) = A(i,k)/A(k,k)$
$A(i, k+1:n) = A(i, k+1:n) - A(i,k) * A(k, k+1:n)$
end
end
Salida: p, A

Ejemplo 5.5. *Vamos a obtener una factorización de la matriz*

$$A = \begin{bmatrix} 1 & -4 & 2 \\ 2 & -1 & -3 \\ 4 & -4 & -8 \end{bmatrix}$$

aplicando el MEG con pivote total:

$$\begin{bmatrix} 1 & -4 & 2 \\ 2 & -1 & -3 \\ 4 & -4 & -8 \end{bmatrix} \underset{p=(3,2,1)}{\rightarrow} \begin{bmatrix} 4 & -4 & -8 \\ 2 & -1 & -3 \\ 1 & -4 & 2 \end{bmatrix} \underset{q=(3,2,1)}{\rightarrow} \begin{bmatrix} -8 & -4 & 4 \\ -3 & -1 & 2 \\ 2 & -4 & 1 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} -8 & -4 & 4 \\ 3/8 & 1/2 & 1/2 \\ -1/4 & -5 & 2 \end{bmatrix} \underset{p=(3,1,2)}{\rightarrow} \begin{bmatrix} -8 & -4 & 4 \\ -1/4 & -5 & 2 \\ 3/8 & 1/2 & 1/2 \end{bmatrix} \rightarrow \begin{bmatrix} -8 & -4 & 4 \\ -1/4 & -5 & 2 \\ 3/8 & -1/10 & 7/10 \end{bmatrix}$$

Por tanto,

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} A \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1/4 & 1 & 0 \\ 3/8 & -1/10 & 1 \end{bmatrix} \begin{bmatrix} -8 & -4 & 4 \\ 0 & -5 & 2 \\ 0 & 0 & 7/10 \end{bmatrix}.$$

5.2. Análisis del error en la factorización LU

En esta sección vamos a considerar un sistema en punto flotante F como al final del tema 3 cuya unidad máquina denotamos por u. Recordemos que las operaciones aritméticas, que representamos en forma general como $\odot \in \{+, -, *, /\}$, se implementan como $\odot^*$ de modo que se cumpla

$$x \odot^* y = fl(x \odot y), \quad \forall x, y \in F,$$

es decir, $x \odot^* y = (x \odot y)(1 + \delta)$ con $|\delta| \leq u$.

Sea $A \in \mathbb{R}^{n \times n}$ una matriz regular factorizable LU. Recordar de la sección anterior que esta condición siempre se tiene salvo un cambio previo de filas.

Nuestro objetivo es analizar como se propaga el error de redondeo cuando obtenemos la factorización LU y resolvemos mediante esta el sistema $Ax = b$. Obtendremos que la solución obtenida, $\overline{x}$, es la solución exacta de un sistema perturbado

$$(A + \triangle A)\overline{x} = b$$

y que el tamaño de la perturbación, $\triangle A$, depende del tamaño de los valores obtenidos en las matrices L y U, que habrá que comparar con el tamaño de las entradas de A, como ya se hizo al final del tema 3 cuando nos planteamos la situación 'ideal' en la que sólo cometíamos error de redondeo al almacenar los datos del problema. El hecho anunciado permitirá entender con más profundidad porqué las técnicas de pivote parcial o total son en general estables, siendo más estable la última.

Comencemos obteniendo la factorización LU de forma compacta mediante el *algoritmo de Crout-Doolittle*: se determina la fila k-ésima de ambas matrices una vez que conocemos las anteriores filas y que usamos que $A = LU$. De entrada

$$L(1,:) = [1, zeros(1, n-1)] \text{ y } U(1,:) = A(1,:).$$

Supongamos conocidas las primeras $k-1$ filas de ambas matrices. Observar que al ser U triangular superior en realidad conocemos también las primeras $k-1$ columnas. Determinemos la fila k-ésima de L. Para $1 \leq j \leq k-1$ tenemos

$$A(k,j) = \sum_{s=1}^{n} L(k,s)U(s,j) = \sum_{s=1}^{j} L(k,s)U(s,j), \tag{5.2}$$

de donde

$$L(k,j) = \left(A(k,j) - \sum_{s=1}^{j-1} L(k,s)U(s,j) \right) / U(j,j),$$

de donde obtenemos de forma recursiva los elementos desconocidos de la fila buscada, ya que en las fórmula intervienen los elementos de las primeras $k-1$ columnas de U. Determinemos ahora la fila k-ésima de U. Para $k \leq j \leq n$ tenemos

$$A(k,j) = \sum_{s=1}^{k-1} L(k,s)U(s,j) + U(k,j),$$

de donde

$$U(k,j) = A(k,j) - \sum_{s=1}^{k-1} L(k,s)U(s,j).$$

Algoritmo 5.3 (Crout-Doolittle)**.** *Datos:A*

for $k = 2 : n$

for $j = 1 : k-1$

$A(k,j) = A(k,j) - A(k:1:j-1) * A(1:j-1,j))/A(j,j)$

```
        end
        for j = k : n
            A(k, j) = A(k, j) − A(k, 1 : k − 1) ∗ A(1 : k − 1, j)
        end
    end
    Salida:A
```

Nota 5.7. *Del método de Crout-Doolittle se desprende que tenemos que analizar como se propaga el error de redondeo al evaluar expresiones del tipo*

$$y = (c - \sum_{i=1}^{k-1} a_i b_i)/b_k.$$

Notar que, por ejemplo, la suma en punto flotante no es asociativa, por lo que el error final puede depender del orden en el que se realizen las operaciones. No obstante, nosotros buscaremos una cota que será independiente del orden de las operaciones. Notar también que si deseamos realizar por ejemplo $s = a + b + c$, en realidad obtenemos

$$\begin{aligned} \overline{s} &= ((a+b)(1+\delta_1)+c)(1+\delta_2) = (a+b)(1+\delta_1)(1+\delta_2)+c(1+\delta_2) \\ &= (a+b+c)(1+\varepsilon) \end{aligned}$$

donde ε, que sería el error relativo cometido, tiene una expresión complicada. Por ello será conveniente trabajar con unas cantidades auxiliares que introducimos en el Lema siguiente.

Lema 5.3. *Sea $m \in \mathbb{N}$ tal que $mu < 1$. Introducimos los números positivos*

$$\gamma_k = \frac{ku}{1-ku}, \quad 1 \le k \le m.$$

Entonces

(i) $\gamma_k \le \gamma_{k'}$, $1 \le k, k' \le m$,

(ii) $\gamma_k + \gamma_{k'} + \gamma_k\gamma_{k'} \le \gamma_{k+k'}$, $k + k' \le m$.

Demostración. La prueba de (i) es sencilla. Para (ii) sólo realizamos las operaciones indicadas y tenemos

$$\gamma_k + \gamma_{k'} + \gamma_k\gamma_{k'} = \frac{(k+k')u - kk'u^2}{1-(k+k')u+kk'u^2} \le \gamma_{k+k'}.$$

□

Lema 5.4. *Sea $k \in \mathbb{N}$ tal que $ku < 1$. Sean $\delta_i \in \mathbb{R}$ con $|\delta_i| \le u$, $1 \le i \le k$, y sean $\rho_i = \pm 1$, $1 \le i \le k$. Entonces*

$$\prod_{i=1}^{k}(1+\delta_i)^{\rho_i} = 1 + \mu_k$$

con $|\mu_k| \le \gamma_k$.

Demostración. Por inducción sobre el número de factores. Para $k = 1$ tenemos que si $\rho_1 = 1$, entonces tomaríamos $\mu_1 = \delta_1$ con

$$|\mu_1| \leq u \leq \gamma_1,$$

mientras que si $\rho_1 = -1$, entonces

$$\frac{1}{1+\delta_1} = 1 - \frac{\delta_1}{1+\delta_1} =: 1 + \mu_1,$$

con

$$|\mu_1| \leq \frac{|\delta_1|}{|1+\delta_1|} \leq \frac{|\delta_1|}{1-|\delta_1|} \leq \gamma_1.$$

Supongamos el resultado cierto para $k-1$ factores y comprobémoslo para k. Si $\rho_k = 1$, entonces por hipótesis de inducción tenemos

$$\prod_{i=1}^{k}(1+\delta_i)^{\rho_i} = (1+\mu_{k-1})(1+\delta_k) = 1 + \mu_{k-1} + \delta_k + \mu_{k-1}\delta_k =: 1 + \mu_k$$

donde

$$|\mu_k| \leq \gamma_{k-1} + u(1+\gamma_{k-1}) = \frac{ku}{1-(k-1)u} \leq \gamma_k.$$

Si $\rho_k = -1$, entonces por hipótesis de inducción tenemos

$$\prod_{i=1}^{k}(1+\delta_i)^{\rho_i} = (1+\mu_{k-1})(1-\frac{\delta_k}{1+\delta_k}) = 1+\mu_{k-1}-\frac{\delta_k}{1+\delta_k}(1+\mu_{k-1}) =: 1+\mu_k,$$

donde

$$\begin{aligned} |\mu_k| &\leq \gamma_{k-1} + \frac{u}{1-u}(1+\gamma_{k-1}) = \frac{(k-1)u(1-u)+u}{(1-u)(1-(k-1)u)} \\ &= \frac{ku - u^2(k-1)}{1-ku+u^2(k-1)} \leq \gamma_k. \end{aligned}$$

□

Lema 5.5. *Sea $k \in \mathbb{N}$ tal que $ku < 1$. Sean $a_i, b_i, c, d \in F$, $1 \leq i \leq k-1$. Calculamos en nuestro sistema en punto flotante*

$$y = (c - \sum_{i=1}^{k-1} a_i b_i)/d,$$

entendiendo que para $k = 1$ no hay sumatorio, mediante el algoritmo

$s = c$
for $i = 1 : k-1$
  $s = s - a_i b_i$
end
$s = s/d$

Sea $\overline{y}$ el valor obtenido. Entonces

$$\overline{y}d(1+\mu_k) = c - \sum_{i=1}^{k-1} a_i b_i (1+\mu_i),$$

donde $|\mu_i| \leq \gamma_i$, $1 \leq i \leq k$. (Si $d = 1$, entonces $|\mu_k| \leq \gamma_{k-1}$, sobreentendiendo que $\gamma_0 = 0$)

Demostración. En el caso $k = 1$, en el que no tenemos sumatorio, se tiene

$$\overline{y} = (1+\delta_1)c/d$$

con lo que $\overline{y}d(1+\delta_1)^{-1} = c$, y por el lema anterior $(1+\delta_1)^{-1} = 1+\mu_1$ con $|\mu_1| \leq \gamma_1$.

Analicemos el caso $k > 1$. Sea $\overline{s}$ el valor calculado de $c - \sum_{i=1}^{k-1} a_i b_i$ por la parte correspondiente del algoritmo propuesto. Comprobemos por inducción sobre k que

$$\overline{s} = c\prod_{i=1}^{k-1}(1+\delta_i) - \sum_{i=1}^{k-1} a_i b_i (1+\varepsilon_i)\prod_{j=i}^{k-1}(1+\delta_j),$$

donde $|\delta_i|, |\varepsilon_i| \leq u$, $1 \leq i \leq k-1$. En el caso $k = 2$ tenemos

$$\overline{s} = (c - a_1 b_1 (1+\varepsilon_1))(1+\delta_1) = c(1+\delta_1) - a_1 b_1 (1+\varepsilon_1)(1+\delta_1).$$

Supongamos cierta la relación para $k-1$ y comprobémoslo para k. Por hipótesis de inducción tenemos

$$\overline{s} = \left((c\prod_{i=1}^{k-1}(1+\delta_i) - \sum_{i=1}^{k-1} a_i b_i (1+\varepsilon_i)\prod_{j=i}^{k-1}(1+\delta_j)) - a_k b_k (1+\varepsilon_k) \right)(1+\delta_k)$$

que se corresponde con el resultado anunciado. Por tanto,

$$\overline{y}d = \left(c\prod_{i=1}^{k-1}(1+\delta_i) - \sum_{i=1}^{k-1} a_i b_i (1+\varepsilon_i)\prod_{j=i}^{k-1}(1+\delta_j) \right)(1+\delta_k),$$

y en consecuencia, despejando c (el fin será evitar perturbaciones en el término independiente del sistema de ecuaciones lineales) se tiene

$$\overline{y}d\prod_{i=1}^{k}(1+\delta_i)^{-1} = c - \sum_{i=1}^{k-1} a_i b_i (1+\varepsilon_i)\prod_{j=1}^{i-1}(1+\delta_j).$$

Ahora aplicamos el lema anterior para reescribir $\prod_{i=1}^{k}(1+\delta_i)^{-1} = 1+\mu_k$ y $(1+\varepsilon_i)\prod_{j=1}^{i-1}(1+\delta_j) = 1+\mu_i$, $1 \leq i \leq k-1$, con las condiciones requeridas. (Cuando no hay división, $d = 1$, no aparece el factor $(1+\delta_k)$ de ahí que se pueda mejorar la cota de la forma indicada). □

Definición 5.2. *Sean $A, B \in \mathbb{R}^{n\times n}$. Diremos que $A \leq B$ si $A(i,j) \leq B(i,j)$, $1 \leq i,j \leq n$.*

La matriz $|A| \in \mathbb{R}^{n\times n}$ es la determinada por $|A|(i,j) = |A(i,j)|$, $1 \leq i,j \leq n$.

Ejercicio 5.2. *Mostrar que si $|A| \leq |B|$, entonces*

$$\|A\|_p \leq \|B\|_p$$

para $p = 1, \infty, F$. Analizar el caso $p = 2$.(En este caso no puede ser cierta la propiedad, ya que $\|\|_2$ es una norma dependiente del signo de las entradas de la matriz. Por ejemplo, tomar $A = [1, -1; -1, 1]$ con $\|A\|_2 = 2$ y $B = [1, 1; -1, 1]$ con $\|B\|_2 = \sqrt{2}$, mientras que $|A| = |B|$)

Teorema 5.4. *Sea $n \in \mathbb{N}$ tal que $nu < 1$. Sea $U \in F^{n\times n}$ regular y triangular superior, y $b \in F^n$. Sea $\overline{x}$ el valor obtenido por el método del remonte con aritmética de precisión limitada aplicado al sistema $Ux = b$. Entonces,*

$$(U + \triangle U)\overline{x} = b$$

con $|\triangle U| \leq \gamma_n |U|$.

Demostración. Recordemos que

$$\begin{aligned}
\overline{x}(k) &= fl\left(\left(b(k) - \sum_{s=k+1}^{n} U(k,s)\overline{x}(s)\right)/U(k,k)\right) \\
&= fl\left(\left(b(k) - \sum_{i=1}^{n-k} U(k,k+i)\overline{x}(k+i)\right)/U(k,k)\right), \quad k = n, \ldots, 1.
\end{aligned}$$

Aplicando lema anterior tenemos

$$\overline{x}(k)U(k,k)(1+\mu_{n-k+1}) = b(k) - \sum_{i=1}^{n-k} U(k,k+i)\overline{x}(k+i)(1+\mu_i),$$

con $|\mu_i| \leq \gamma_i$, $1 \leq i \leq n-k+1$ (en realidad los μ_i dependerían de k pero no lo indicaremos ya que lo único que nos importa es que están acotados de la forma indicada). Por consiguiente,

$$\overline{x}(k)U(k,k)(1+\mu_{n-k+1}) + \sum_{i=1}^{n-k} U(k,k+i)\overline{x}(k+i)(1+\mu_i) = b(k),$$

es decir, $\overline{x}$ se puede interpretar como la solución exacta de un sistema obtenido por perturbación de la matriz de coeficientes del original, que sigue siendo triangular y cuya ecuación k-ésima tiene por coeficientes

$$(U+\triangle U)(k,:) = [zeros(1,k-1), U(k,k)(1+\mu_{n-k+1}), \ldots U(k,n)(1+\mu_{n-k}),].$$

Así,

$$|\triangle U(k,j)| \leq \begin{cases} \gamma_{n-k+1}|\triangle U(k,k)|, & k = j, \\ \gamma_{j-k}|\triangle U(k,j)|, & k+1 \leq j \leq n, \end{cases}$$

por lo que podemos acotar todas las entradas de $|\triangle U|$ utilizando el γ más grande que nos aparece, γ_n, obteniéndose la cota del enunciado. □

Nota 5.8. *Del teorema anterior se desprende que resolver sistemas triangulares es un proceso estable. De hecho, por un resultado de perturbación comentado en el tema 3 se tiene que*

$$\frac{\|x-\overline{x}\|_\infty}{\|x\|_\infty} \leq \frac{\gamma_n k_\infty(U)}{1-\gamma_n k_\infty(U)}$$

ya que $\|\triangle U\|_\infty \leq \gamma_n \|U\|_\infty$*, siempre que* $\gamma_n k_\infty(U) < 1$*.*

Si consideramos el caso más real de cometer además errores al almacenar los datos del problema, lo que hemos evitado tmando datos en F*, obtendríamos* $\overline{x}$ *que sería la solución exacta del sistema*

$$(U + \triangle U + \triangle(U + \triangle U))\overline{x} = b + \triangle b,$$

con

$$\|\triangle U\|_\infty \leq u\|U\|_\infty\,,\ \|\triangle b\|_\infty \leq u\|b\|_\infty$$

y

$$\|\triangle(U+\triangle U)\|_\infty \leq \gamma_n \|U+\triangle U\|_\infty \leq \gamma_n \|U\|_\infty (1+u)$$

de donde

$$\begin{aligned}\|\triangle U + \triangle(U+\triangle U)\|_\infty &\leq (u+\gamma_n(1+u))\|U\|_\infty \\ &= \frac{(n+1)u}{1-nu}\|U\|_\infty \leq \gamma_{n+1}\|U\|_\infty\,,\end{aligned}$$

lo que muestra que esta última fuente de errores la podemos practicamente despreciar o considerar ya incluida. Si $\gamma_{n+1}k_\infty(U) < 1$*, entonces*

$$\frac{\|x-\overline{x}\|_\infty}{\|x\|_\infty} \leq \frac{k_\infty(U)}{1-\gamma_{n+1}k_\infty(U)}(\gamma_{n+1}+u) \leq \frac{\gamma_{n+2}k_\infty(U)}{1-\gamma_{n+1}k_\infty(U)}.$$

(Notar que $nu \leq \gamma_n k_\infty(U) < 1$*).*

Observar que al trabajar con cotas y no con los errores relativos introducidos durante el proceso, el resultado obtenido no depende del orden en que realizemos las operaciones. Además, es evidente que el resultado es cierto también para sistemas triangulares inferiores a los que les aplicamos el método del descenso.

Teorema 5.5. *Sea* $n \in \mathbb{N}$ *tal que* $nu < 1$*. Sea* $A \in F^{n\times n}$ *regular y factorizable LU. Sean* $\overline{L}$ *y* $\overline{U}$ *los factores obtenidos mediante el algoritmo de Crout-Doolittle con aritmética de precisión limitada. Entonces,*

$$\overline{L}\,\overline{U} = A + \triangle A$$

con $|\triangle A| \leq \gamma_n \left|\overline{L}\right|\left|\overline{U}\right|$*.*

Demostración. Recordemos que el citado algoritmo se calculan de manera recursiva las filas de L y U. Calculemos las filas k-ésimas. Tenemos

$$\overline{L}(k,j) = fl\left(\left(A(k,j) - \sum_{s=1}^{j-1}\overline{L}(k,s)\overline{U}(s,j)\right)/\overline{U}(j,j)\right), \quad 1 \leq j \leq k-1.$$

Por el último lema,

$$\overline{L}(k,j)\overline{U}(j,j)(1+\mu_j) = A(k,j) - \sum_{s=1}^{j-1} \overline{L}(k,s)\overline{U}(s,j)(1+\mu_s),$$

con $|\mu_i| \leq \gamma_i$, $1 \leq i \leq j$ (en realidad los μ_i dependerían de (k,j) pero no lo indicaremos para simplificar la notación, ya que lo único que nos importa es que están acotados de la forma indicada). Por tanto,

$$\begin{aligned} \left|A(k,j) - (\overline{LU})(k,j)\right| &\leq \left|\overline{L}(k,j)\overline{U}(j,j)\mu_j + \sum_{s=1}^{j-1} \overline{L}(k,s)\overline{U}(s,j)\mu_s\right| \\ &\leq \gamma_j \sum_{s=1}^{j} \left|\overline{L}(k,s)\right| \left|\overline{U}(s,j)\right| = \gamma_j(\left|\overline{L}\right| \left|\overline{U}\right|)(k,j) \end{aligned}$$

para $1 \leq j \leq k-1$. Ahora para $k \leq j \leq n$, tenemos

$$\overline{U}(k,j) = fl\left(A(k,j) - \sum_{s=1}^{k-1} \overline{L}(k,s)\overline{U}(s,j)\right), \quad 1 \leq j \leq k-1,$$

y por el lema antes mencionado

$$\overline{U}(k,j)(1+\mu_k) = A(k,j) - \sum_{s=1}^{k-1} \overline{L}(k,s)\overline{U}(s,j)(1+\mu_s),$$

con $|\mu_s| \leq \gamma_s$, $1 \leq s \leq k$ (seguimos cometiendo el abuso de notación ya comentado). Por tanto, puesto que $\overline{L}(k,k) = 1$,

$$\begin{aligned} \left|A(k,j) - (\overline{LU})(k,j)\right| &\leq \left|\overline{U}(k,j)\mu_k + \sum_{s=1}^{k-1} \overline{L}(k,s)\overline{U}(s,j)\mu_s\right| \\ &\leq \gamma_k \sum_{s=1}^{k} \left|\overline{L}(k,s)\right| \left|\overline{U}(s,j)\right| = \gamma_k(\left|\overline{L}\right| \left|\overline{U}\right|)(k,j), \end{aligned}$$

para $k \leq j \leq n$. El resultado anunciado se obtiene al considerar el mayor de los γ que nos aparecen, γ_n, para tener una cota común para todas las entradas de $|\triangle A|$. □

Nota 5.9. *Como en cada fila el mayor γ que aparece es consecuencia de calcular $\overline{U}(k,j)$, y este no requiere de división, en realidad podríamos obtener en la cota γ_{n-1} en lugar de γ_n.*

Ejemplo 5.6. *Volvamos a los ejemplos que motivaron la introducción de las técnicas de pivote parcial y total.*

(i) *Sea*

$$A = \begin{bmatrix} 10^{-17} & 1 \\ 1 & 1 \end{bmatrix}$$

La factorización que obtendríamos utilizando MATLAB con nuestro algoritmo es

$$\overline{L} = \begin{bmatrix} 1 & 0 \\ 10^{17} & 1 \end{bmatrix}, \overline{U} = \begin{bmatrix} 10^{-17} & 1 \\ 0 & -10^{17} \end{bmatrix}$$

y por tanto,

$$A - \overline{L}\,\overline{U} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

observándose un error relativo alto en la factorización de A. Como

$$|\overline{L}|\,|\overline{U}| = \begin{bmatrix} 10^{-17} & 1 \\ 1 & 2 * 10^{17} \end{bmatrix}$$

nuestro teorema nos dice que

$$\frac{1}{2} = \frac{\left\|A - \overline{L}\,\overline{U}\right\|_\infty}{\|A\|_\infty} \leq \gamma_2 \frac{\left\||\overline{L}|\,|\overline{U}|\right\|_\infty}{\|A\|_\infty} \leq 4.5 * 10^{-16} \frac{1 + 2 * 10^{17}}{2} \doteq 45.$$

Notar que $k_\infty(A) = 4/(1-\varepsilon) \doteq 4$, es decir, el problema no es consecuencia de que la matriz A esté mal condicionada.

(ii) *Retomamos el ejemplo anterior utilizando pivote parcial, es decir, factorizamos*

$$PA = \begin{bmatrix} 1 & 1 \\ 10^{-17} & 1 \end{bmatrix}.$$

Se obtiene

$$\overline{L} = \begin{bmatrix} 1 & 0 \\ 10^{-17} & 1 \end{bmatrix}, \overline{U} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Por tanto,

$$PA - \overline{LU} = \begin{bmatrix} 0 & 0 \\ 0 & 10^{-17} \end{bmatrix}$$

y $\left\||\overline{L}|\,|\overline{U}|\right\|_\infty / \|A\|_\infty = 1$, mientras que $\frac{1}{2}10^{-17} = \left\|A - \overline{L}\,\overline{U}\right\|_\infty / \|A\|_\infty$.

(iii) NO *En este ejemplo se pone de manifiesto que el pivote parcial, aún evitando multiplicadores grandes, ya que son menores que la unidad en módulo, no evitan un error relativo grande. Sea*

$$A = \begin{bmatrix} 1 & 10^{17} \\ 1 & 1 \end{bmatrix}$$

para la que obtendríamos los siguientes factores

$$\overline{L} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \overline{U} = \begin{bmatrix} 1 & 10^{17} \\ 0 & -10^{17} \end{bmatrix}.$$

Notar que en realidad estamos utilizando pivote parcial. Cometemos pues un error relativo en $\|\|_\infty$ al factorizar la matriz de $1/(1+10^{17}) \doteq 10^{-17}$. En este caso, $\left\||\overline{L}|\,|\overline{U}|\right\|_\infty / \|A\|_\infty = (1+2*10^{17})/(1+10^{17}) \doteq 2$.

Corolario 5.1. *Sea $n \in \mathbb{N}$ tal que $3nu < 1$. Sea $A \in F^{n\times n}$ regular y factorizable LU. Sea $b \in F^n$. Sean $\overline{L}$ y $\overline{U}$ los factores obtenidos mediante el algoritmo de Crout-Doolittle con aritmética de precisión limitada. Sea $\overline{x}$ el valor obtenido al resolver*

$$\overline{L}y = b \text{ y } \overline{U}z = y$$

por descenso y remonte respectivamente con aritmética de precisión finita. Entonces,

$$(A + \triangle A)\overline{x} = b$$

con $|\triangle A| \leq \gamma_{3n} \left|\overline{L}\right| \left|\overline{U}\right|$.

Demostración. Por el teorema anterior $\overline{L}\,\overline{U} = A + \delta A$ con $|\delta A| \leq \gamma_n \left|\overline{L}\right| \left|\overline{U}\right|$. Por el teorema correspondiente a la propagación del error de redondeo en el método del descenso tenemos que $\overline{y}$, solución calculada de $\overline{L}y = b$, verifica

$$(\overline{L} + \triangle\overline{L})\overline{y} = b$$

con $\left|\triangle\overline{L}\right| \leq \gamma_n \left|\overline{L}\right|$. Análogamente, $\overline{x}$, solución calculada de $\overline{U}y = \overline{y}$, verifica

$$(\overline{U} + \triangle\overline{U})\overline{x} = \overline{y}$$

con $\left|\triangle\overline{U}\right| \leq \gamma_n \left|\overline{U}\right|$.

En definitiva,

$$\begin{aligned} b &= (\overline{L} + \triangle\overline{L})(\overline{U} + \triangle\overline{U})\overline{x} = (\overline{L}\overline{U} + \triangle\overline{L}\overline{U} + \overline{L}\triangle\overline{U} + \triangle\overline{L}\triangle\overline{U})\overline{x} \\ &= (A + \delta A + \triangle\overline{L}\overline{U} + \overline{L}\triangle\overline{U} + \triangle\overline{L}\triangle\overline{U})\overline{x} =: (A + \triangle A)\overline{x}, \end{aligned}$$

con

$$|\triangle A| \leq (3\gamma_n + \gamma_n\gamma_n)\left|\overline{L}\right|\left|\overline{U}\right| \leq (\gamma_n + \gamma_{2n})\left|\overline{L}\right|\left|\overline{U}\right| \leq \gamma_{3n}\left|\overline{L}\right|\left|\overline{U}\right|,$$

donde hemos usado el apartado (ii) del lema del primer lema en las dos últimas desigualdades. □

Nota 5.10. *En el teorema previo se pone de manifiesto que la estabilidad del método, medida en términos de cometer un error relativo pequeño en la solución, no depende de si los multiplicadores son en módulo de tamaño controlado, como podríamos pensar y sucede al utilizar técnicas de elección del pivote, si no del tamaño de la matriz $\left|\overline{L}\right|\left|\overline{U}\right|$.*

En las condiciones del citado teorema y con la notación allí introducida, si $\gamma_{3n}k_\infty(A)\frac{\left\|\left|\overline{L}\right|\left|\overline{U}\right|\right\|_\infty}{\|A\|_\infty} =: \alpha < 1$, se tiene que

$$\frac{\|x - \overline{x}\|_\infty}{\|x\|_\infty} \leq \frac{\alpha}{1-\alpha}.$$

En la práctica, es en general difícil calcular a priori $\left\|\left|\overline{L}\right|\left|\overline{U}\right|\right\|_\infty$. Sólo para matrices con ciertas propiedades es relativamente comodo hacerlo,

como por ejemplo las matrices positivas (?). Por ello, se trabaja con un nuevo concepto que introducimos a continuación, que es mucho más cómodo y proporciona una idea de lo que sucede, aunque no puede usarse de forma rigurosa como cota del error relativo.

Estimemos $|L|\,|U|$ *en lugar de* $\left|\overline{L}\right|\left|\overline{U}\right|$. *Usaremos la notación del MEG. Sea* $1 \leq k \leq n-1$. *Para* $k+1 \leq i,j \leq n$ *tenemos*

$$A^{(k+1)}(i,j) = A^{(k)}(i,j) - L(i,k)U(k,j),$$

de donde

$$|L(i,k)|\,|U(k,j)| \leq 2 \max_{1\leq k\leq n} \left|A^{(k)}(i,j)\right|.$$

Además, como $L(i,k) = 0$ *para* $1 \leq i \leq k-1$ *y* $L(k,k) = 1$, *entonces*

$$|L(i,k)|\,|U(k,j)| \leq |U(k,j)| = \left|A^{(k+1)}(k,j)\right|, \quad 1 \leq j \leq n,$$

donde hemos tenido en cuenta la posible elección de pivote. Por último, como en el razonamiento anterior,

$$|L(i,n)|\,|U(n,j)| \leq |U(n,j)| = \left|A^{(n)}(n,j)\right|, \quad 1 \leq i,j \leq n.$$

De todo lo expuesto,

$$|L(i,k)|\,|U(k,j)| \leq 2 \max_{1\leq k,i,j\leq n} \left|A^{(k)}(i,j)\right|, \quad 1 \leq i,j,k \leq n.$$

Así,

$$(|L|\,|U|)(i,j) = \sum_{k=1}^{n} |L(i,k)|\,|U(k,j)| \leq 2n \max_{1\leq i,j,k\leq n} \left|A^{(k)}(i,j)\right|,$$

y como consecuencia,

$$\begin{aligned} \||L|\,|U|\|_\infty &= \max_{1\leq i\leq n} \sum_{j=1}^{n} (|L|\,|U|)(i,j) \leq 2n^2 \max_{1\leq i,j,k\leq n} \left|A^{(k)}(i,j)\right| \\ &\leq \|A\|_\infty\, 2n^2 \frac{\max_{1\leq i,j,k\leq n} \left|A^{(k)}(i,j)\right|}{\|A\|_{\max}}, \end{aligned}$$

ya que $\|A\|_{\max} \leq \|A\|_\infty$.

Definición 5.3. *Sea* $A \in \mathbb{R}^{n\times n}$ *regular. Se define el factor de crecimiento de A como*

$$\rho_n(A) = \frac{\max_{1\leq i,j,k\leq n} \left|A^{(k)}(i,j)\right|}{\|A\|_\infty}.$$

Estudiemos el factor de crecimiento cuando aplicamos MEGPP a $A \in \mathbb{R}^{n\times n}$ regular, que denotaremos mediante $\rho_n^P(A)$. Analicemos como se modifican los tamaños de las entradas de la matriz $A^{(k)}$ en la etapa $k-$ésima para

pasar a $A^{(k+1)}$, $1 \leq k \leq n-1$. En primer lugar notar que $A^{(k)}$ y $A^{(k+1)}$ coinciden en sus primeras $k-1$ filas y columnas. Si tenemos que elegir pivote, deberíamos intercambiar la fila $k-$ésima con una posterior $i(k)-$ésima. Sea $B^{(k)}$ la matriz así obtenida. Notar que

$$\max_{1\leq i,j\leq n} \left|B^{(k)}(i,j)\right| = \max_{1\leq i,j\leq n} \left|A^{(k)}(i,j)\right|.$$

Para $k+1 \leq i,j \leq n$ tenemos

$$\begin{aligned} \left|A^{(k+1)}(i,j)\right| &= \left|B^{(k)}(i,j) - \frac{B^{(k)}(i,k)}{B^{(k)}(k,k)} B^{(k)}(k,j)\right| \\ &\leq \left|B^{(k)}(i,j)\right| + \left|B^{(k)}(k,j)\right| \leq 2 \max_{1\leq r,s\leq n} \left|A^{(k)}(r,s)\right|, \end{aligned}$$

donde hemos usado que los multiplicadores en módulo son menores que la unidad (notar que esto no podemos asegurarlo con aritmética de precisión limitada). En resumidas cuentas como mucho duplicamos en cada etapa el tamaño de los elementos de los que partíamos. Por lo tanto,

$$\rho_n^P(A) \leq 2^{n-1}.$$

En el ejemplo que sigue, tenemos un caso 'académico' en el que se alcanza dicha cota. No obstante, recientemente se han encontrado ejemplos en los que MEGPP sufre crecimiento exponencial como el mostrado en el ejemplo (problemas de contorno de dos puntos a los que se les aplica el método del disparo (Wright,1993)). Sin embargo, en general se puede decir que el método es estable en la práctica ya que es estadísticamente muy dificíl encontrar ejemplos en los que suceda esto.

Ejemplo 5.7. *Se tiene la siguiente factorización de la matriz*

$$\begin{aligned} A_n &= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ -1 & 1 & 0 & \cdots & 0 & 1 \\ -1 & -1 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & -1 & -1 & \cdots & 1 & 1 \\ -1 & -1 & -1 & \cdots & -1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ -1 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & -1 & -1 & \cdots & 1 & 0 \\ -1 & -1 & -1 & \cdots & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & 0 & \cdots & 0 & 2 \\ 0 & 0 & 1 & \cdots & 0 & 4 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 2^{n-2} \\ 0 & 0 & 0 & \cdots & 0 & 2^{n-1} \end{bmatrix} \end{aligned}$$

con lo que $\rho_n^P(A) = 2^{n-1}$.

Para MEGPT J. H. Wilkinson (1961) probó que el correspondiente factor de crecimiento, $\rho_n^T(A)$, satisface

$$\rho_n^T(A) \leq \sqrt{n\, 2\, 3^{1/2}\, n^{1/(n-1)}} =: f(n) \approx n^{1/2+\log n/4}.$$

Se tiene por ejemplo que $f(100) = 3570$ mientras que $2^{99} > 10^{29}$. Podemos afirmar que esta técnica es más estable que MEGPP, pero como ya se vio es significativamente más costosa y lenta, por lo que no se suele elegir.

Wilkinson conjeturo que $\rho_n^T(A) \leq n$, y no ha sido hasta muy recientemente (N. Gould, 1991) cuando se ha probado que era falsa. Sigue siendo un problema abierto determinar una buena cota superior de $\rho_n^T(A)$ y se espera que sea $O(n)$.

5.3. Sobre la factorización de algunas matrices especiales

En este capítulo analizamos el MEG-factorización LU para ciertas clases de matrices que aparecen en algunas aplicaciones interesantes como las que vimos en la introducción: matrices diagonalmente dominantes, matrices simétricas definidas positivas y matrices tridiagonales.

5.3.1. Matrices diagonalmente dominantes

Teorema 5.6. *Sea $A \in \mathbb{R}^{n\times n}$ regular y diagonalmente dominante. Entonces A es factorizable LU y $\rho_n(A) \leq 2$.*

Si A es diagonalmente dominante por columnas, entonces

$$|L(i,j)| \leq 1, \quad 2 \leq i \leq n,\ 1 \leq j \leq i-1. \tag{5.3}$$

Demostración. Haremos la prueba para matrices diagonalmente dominante por columnas, dejando para el lector las oportunas modificaciones para el otro caso. Probaremos por inducción sobre $n \geq 2$ que el MEG aplicado a A no requiere elección del pivote, que se cumple (5.3) y que

$$\sum_{i=k}^{n} \left|A^{(k)}(i,j)\right| \leq \sum_{i=1}^{n} |A(i,j)|\,, \quad 2 \leq k \leq n,\ k \leq j \leq n. \tag{5.4}$$

que nos permitirá posteriormente probar la propiedad relativa al factor de crecimiento.

Para $n = 2$ tenemos

$$|A(2,1)| \leq |A(1,1)|\,,$$

lo que al ser A regular implica que $A(1,1) \neq 0$. Por tanto, el MEG no requiere elección del pivote. Además,

$$|L(2,1)| = |A(2,1)/A(1,1)| \leq 1.$$

En cuanto a (5.4) tenemos

$$\left|A^{(2)}(2,2)\right| = |A(2,2) - A(1,2)A(2,1)/A(1,1)| \leq |A(1,2)| + |A(2,2)|\,.$$

Supongamos el resultado válido para $n-1$ y comprobémoslo para el siguiente natural. Tenemos

$$\sum_{i=2}^{n} |A(i,1)| \leq |A(1,1)|\,,$$

por lo que al ser A regular se tiene que $A(1,1) \neq 0$ y puede elegirse como pivote. Además,

$$|L(i,1)| = |A(i,1)/A(1,1)| \leq \sum_{i=2}^{n} |A(i,1)| \,/\, |A(1,1)| \leq 1, \quad 2 \leq i \leq n.$$

Consideremos la matriz $B := A^{(2)}(2:n, 2:n) \in \mathbb{R}^{(n-1)\times(n-1)}$ y comprobemos que es regular y diagonalmente dominante por columnas. La regularidad de B es ya conocida, aunque se puede deducir de la relación $\det(A) = A(1,1)\,\det(B)$. En cuanto a la dominancia, para $1 \leq j \leq n-1$ tenemos

$$\begin{aligned}
\sum_{i=1,\, i\neq j}^{n-1} |B(i,j)| &= \sum_{s=2,\, s\neq r}^{n} \left|A^{(2)}(s,r)\right| \\
&= \sum_{s=2,\, s\neq r}^{n} |A(s,r) - A(1,r)A(s,1)/A(1,1)| \\
&\leq \sum_{s=2,\, s\neq r}^{n} |A(s,r)| + \sum_{s=2,\, s\neq r}^{n} |A(s,1)|\,|A(1,r)|\,/\,|A(1,1)| \\
&\leq |A(r,r)| - |A(1,r)| + (|A(1,1)| - |A(r,1)|)\,|A(1,r)|\,/\,|A(1,1)| \\
&= |A(r,r)| - |A(r,1)A(1,r)/A(1,1)| \\
&\leq |A(r,r) - A(r,1)A(1,r)/A(1,1)| = |B(j,j)|\,.
\end{aligned}$$

Por la hipótesis de inducción„ el MEG aplicado a B no requiere elección del pivote y los multiplicadores son en módulo menores que la unidad, concluyéndose que lo mismo es cierto para la matriz original A. Además, por la propiedad (5.4) para B tenemos, para $2 \leq k \leq n-1$ y $k+1 \leq r \leq n$, que

$$\begin{aligned}
\sum_{s=k+1}^{n} \left|A^{(k+1)}(s,r)\right| &\leq \sum_{s=2}^{n} \left|A^{(2)}(s,r)\right| \\
&= \sum_{s=2}^{n} |A(s,r) - A(1,r)A(s,1)/A(1,1)| \\
&\leq \sum_{s=2}^{n} |A(s,r)| + |A(1,r)|\,,
\end{aligned}$$

lo que junto con

$$\begin{aligned}\sum_{s=2}^{n}\left|A^{(2)}(s,r)\right| &= \sum_{s=2}^{n}|A(s,r)-A(1,r)A(s,1)/A(1,1)| \\ &\leq \sum_{s=1}^{n}|A(s,r)|\,, \quad 2\leq r\leq n,\end{aligned}$$

prueba (5.4) para A.

Así, para $2\leq k\leq n$ tenemos

$$\begin{aligned}\left|A^{(k)}(s,r)\right| &\leq \sum_{s=k}^{n}\left|A^{(k)}(s,r)\right| \leq \sum_{s=1}^{n}|A(s,r)| \\ &\leq 2\,|A(r,r)| \leq 2\,\|A\|_{\text{máx}}\,,\end{aligned}$$

para $k\leq s,r\leq n$, si $k\leq s\leq n$ y $1\leq r\leq k-1$, entonces

$$\left|A^{(k)}(s,r)\right| = 0 \leq 2\,\|A\|_{\text{máx}}\,,$$

y para $1\leq s\leq k-1$,

$$\left|A^{(k)}(s,r)\right| = \left|A^{(s)}(s,r)\right| \leq 2\,\|A\|_{\text{máx}}\,,$$

de lo que se deduce que $\rho_n(A)\leq 2$. □

Ejemplo 5.8. *Este ejemplo pone de manifiesto que la propiedad (5.3) no es válida para matrices diagonalmente dominantes por filas. Tenemos*

$$\begin{bmatrix}\varepsilon & 0\\ \frac{1}{2} & 1\end{bmatrix} = \begin{bmatrix}1 & 0\\ \frac{1}{2\varepsilon} & 1\end{bmatrix}\begin{bmatrix}\varepsilon & 0\\ 0 & 1\end{bmatrix},$$

por lo que basta con elegir $\varepsilon << 1$ para convencerse de la afirmación realizada. Notar que

$$\left|\overline{L}\right|\left|\overline{U}\right| = |L|\,|U| = |LU| = |A|\,,$$

pese a que los multiplicadores son grandes.

Nota 5.11. *La parte del teorema anterior para filas referente a la existencia de la factorización LU se puede probar trabajando con la transpuesta y usando el resultado ya probado:*

$$A = u'l' = u'diag([1/u(1,1),\ldots,1/u(n,n)])\,diag([u(1,1),\ldots,u(n,n)])l' =: LU.$$

5.3.2. Matrices simétricas

Teorema 5.7 (factorización LDM')**.** *Sea $A\in\mathbb{R}^{n\times n}$ regular y factorizable LU. Entonces existen $L,M\in\mathbb{R}^{n\times n}$ triangulares inferiores con unos en la diagonal principal, y $D\in\mathbb{R}^{n\times n}$ diagonal, únicas, tales que*

$$A = LDM'.$$

Demostración. Sea $A = LU$. Consideramos $D = diag(diag(U))$ que es regular por serlo U. Definimos $M = (D^{-1}U)'$ que cumple claramente $M(i,i) = 1$, $1 \leq i \leq n$.

Para la unicidad de la factorización, si

$$L_1 D_1 M_1' = A = L_2 D_2 M_2',$$

entonces por la unicidad de la factorización LU, $L_1 = L_2$ y $D_1 M_1' = D_2 M_2'$, de donde al ser D_i regular, $i = 1, 2$, se tiene

$$M_1'(M_2')^{-1} = D_1^{-1} D_2 = I_n$$

siendo la última igualdad consecuencia de la igualdad entre una matriz diagonal, $D_1^{-1}D_2$, y una triangular inferior con unos en la diagonal principal, $M_1'(M_2')^{-1}$. Por tanto, $D_1 = D_2$ y $M_1 = M_2$. □

Ejercicio 5.3. *Sea A regular que se puede escribir en la forma del teorema anterior $A = LDM'$. Entonces,* $\det(A_k) \neq 0$, $1 \leq k \leq n$.

En efecto, tendríamos que A es regular y factorizable LU, y ya se probó que entonces se verificaba la propiedad anterior.

Corolario 5.2 (factorización LDL')**.** *Sea $A \in \mathbb{R}^{n\times n}$ regular y simétrica. Si tenemos la factorización $A = LDM'$ entonces $M = L$.*

Demostración. Es consecuencia de la unicidad del teorema anterior, ya que

$$LDM' = A = A' = MDL'.$$

□

Nota 5.12. *Si tenemos $A = LDL'$ entonces el sistema $Ax = b$ se resolvería mediante*

$$Ly = b, \quad L'x = D^{-1}y,$$

entendiendo que $D^{-1}y$ no se realiza mediante producto de matrices, si no reescalando las componentes de y convenientemente.

Nota 5.13 (Obtención de la factorización LDL')**.** *Tenemos $A = LDL'$, de donde*

$$A(i,j) = (LD)(i,:)L'(:,j) = \sum_{k=1}^{\min(i,j)} L(i,k)D(k,k)L(j,k),$$

luego

$$A(i,i) = \sum_{k=1}^{i-1} L(i,k)^2 D(k,k) + D(i,i), \quad 1 \leq i \leq n, \qquad ((\mathrm{I}))$$

y

$$A(i,j) = \sum_{k=1}^{i} L(i,k)D(k,k)L(j,k), \quad 1 \leq i < j \leq n. \qquad ((\mathrm{II}))$$

Vamos a obtener las entradas de L y D de estas relaciones, columna por columna. Supuestas conocidas hasta la columna r-ésima de ambas matrices obtenemos de (I) el elemento

$$D(r+1,r+1) = A(r+1,r+1) - \sum_{k=1}^{r} L(r+1,k)^2 D(k,k),$$

y luego de (II), con $i = r+1$ $y r+1 < j \leq n$, determinamos

$$L(j,r+1) = \left[A(r+1,j) - \sum_{k=1}^{r} L(r+1,k)D(k,k)L(j,k) \right] / D(r+1,r+1).$$

Algoritmo 5.4 (LDL')**.** *Guardaremos L en la parte inferior de A, y D en la diagonal principal.*

Datos: A
for $i = 1 : n$
  for $k = 1 : i-1$
    $r(k) = A(i,k) * A(k,k)$
    $A(i,i) = A(i,i) - A(i,k) * r(k)$
  end
  for $j = i+1 : n$
    for $k = 1 : i-1$
      $A(j,i) = A(j,i) - A(j,k) * r(k)$ *(Usamos que* $A(j,i) = A(j,i)$*)*
    end
  $A(j,i) = A(j,i)/A(i,i)$
  end
end

Nota 5.14. *El número de flops que requiere el anterior algoritmo es*

$$\sum_{i=1}^{n} (3(i-1) + (n-i)(2(i-1)+1)) = \frac{1}{3}n^3 + n^2 - \frac{4}{3}n.$$

5.3.3. Matrices simétricas definidas positivas

Teorema 5.8. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica. Equivalen:*

(i) *A es definida positiva,*

(ii) $\det(A_k) > 0$, $1 \leq k \leq n$,

(iii) *existe B triangular inferior con $B(i,i) > 0$, $1 \leq i \leq n$, tal que $A = BB'$ (Factorización de Cholesky)*

Demostración. (i)$\Longrightarrow$(ii) Ya se probó, viendo que las matrices A_k son simétricas definidas pòsitivas, y usando que los valores propios de esta clase de matrices son positivos, y por lo tanto también su determinante es positivo.

(ii)$\Longrightarrow$(iii) Por (ii) tenemos que A es una matriz regular, factorizable LU y simétrica. Por lo tanto, podemos escribir $A = LDL'$ en las condiciones expresadas en la sección anterior. Por un lema anterior tenemos

$$A_k = L_k D_k L_k',$$

de donde

$$0 < \det(A_k) = \det(D_k) = \prod_{i=1}^{k} D(i,i), \quad 1 \leq k \leq n,$$

por lo que $D(i,i) > 0$, $1 \leq i \leq n$. Definimos $H = diag(sqrt(diag(D)))$, con lo que evidentemente tenemos

$$A = LDL' = LHHL' = BB',$$

si elegimos $B = LH$. Es claro que B es triangular inferior; además,

$$B(i,i) = L(i,i)\sqrt{D(i,i)} > 0, \quad 1 \leq i \leq n.$$

(iii)$\Longrightarrow$(i) Es consecuencia de que $A = BB'$ con B es regular. □

Nota 5.15. *En las condiciones del apartado (iii) del teorema, si tenemos*

$$BB' = A = CC',$$

entonces $C^{-1}B = C'B'^{-1}$ *siendo la primera matriz de la igualdad triangular inferior y la segunda triangular superior. Obtendremos la unicidad si comprobamos que los elementos de la diagonal de* $C^{-1}B$ *son todos unos. Tenemos*

$$\begin{aligned} (C^{-1}B)(i,i) &= C^{-1}(i,i)B(i,i) = C(i,i)^{-1}B(i,i) \\ &= B(i,i)^{-1}C(i,i) = (C'B'^{-1})(i,i), \end{aligned}$$

luego

$$B(i,i)^2 = C(i,i)^2,$$

de donde, al ser positivos, se deduce que

$$B(i,i) = C(i,i), \quad 1 \leq i \leq n.$$

Nota 5.16. *El número de flops para obtener la factorización de Cholesky es del mismo orden que el de la factorización* LDL'*, con la diferencia de requerir el cálculo de raíces cuadradas. Para evitar ese cálculo de n raíces cuadradas, puede ser interesante considerar la factorización* LDL' *en lugar de la de Cholesky.*

Nota 5.17. *Recordemos que los elementos más grandes en módulo de una matriz A simétrica definida positiva, están en la diagonal principal. Por otra parte, si $A = BB'$, entonces para $1 \leq i, j \leq n$ se tiene*

$$A(i,i) = \sum_{k=1}^{n} B(i,k)^2 \geq B(i,j)^2,$$

de donde

$$|B(i,j)| \leq \sqrt{A(i,i)},$$

lo que, por analogía con lo que sucedía en el caso de la factorización LU, conduce a la intuición de que el proceso de obtener la factorización de Cholesky de A va a ser estable. En relación con esta idea tenemos el siguiente resultado.

Teorema 5.9. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica y definida positiva. Entonces $\rho_n(A) = 1$.*

Demostración. La idea de la prueba se basa en que los elementos más grandes en módulo de una matriz simétrica definida positiva están en la diagonal principal, de hecho estos son positivos. Comencemos probando por inducción sobre k, $1 \leq k \leq n$, que $A^{(k)}(k:n, k:n)$ es simétrica definida positiva. Esto permitirá tener que comparar tan sólo los elementos de la diagonal principal de estas matrices con los de la de A. Para $k = 1$ es cierto ya que $A^{(1)} = A$. Supongamos que $A^{(k-1)}(k-1:n, k-1:n)$ es de la clase de matrices indicada y la reescribimos en la forma

$$\begin{bmatrix} \alpha & c' \\ c & D \end{bmatrix}$$

con $\alpha > 0$, $c \in \mathbb{R}^{n-k+1}$, $D \in \mathbb{R}^{(n-k+1)\times(n-k+1)}$. Por una parte

$$D(i,j) = A^{(k-1)}(i+1, j+1) = A^{(k-1)}(j+1, i+1) = D(j,i)$$

y por otra, para $z \in \mathbb{R}^{n-k+1}$ no nulo, tenemos

$$z'Dz = [0, z']\, A^{(k-1)}(k-1:n, k-1:n)\, [0; z] > 0,$$

luego D es simétrica definida positiva. Por el MEG tenemos que $A^{(k)}(k-1:n, k-1:n) = D - cc'/\alpha$, que por simplicidad llamaremos S. Esta matriz es simétrica ya que D lo es. Además, para $z \in \mathbb{R}^{n-k+1}$ no nulo y $x \in \mathbb{R}$, tenemos

$$\begin{aligned} 0 &< [1, xz']A^{(k-1)}(k-1:n, k-1:n)[1; xz] \\ &= \alpha + 2c'z\, x + z'Dz\, x^2, \end{aligned}$$

luego el polinomio de segundo grado en x no puede tener raíces reales, y por tanto,

$$0 < \alpha z'Dz - (c'z)^2 = \alpha z'Sz,$$

y así, $z'Sz > 0$.

Mostremos ahora que

$$A^{(k)}(i,i) \leq A(i,i), \quad k \leq i \leq n,\ 1 \leq k \leq n,$$

de nuevo usando inducción sobre k. Para $k = 1$ es evidente. Si es cierto para $k-1$ entonces para $k \leq i \leq n$ tenemos

$$\begin{aligned} A^{(k)}(i,i) &= A^{(k-1)}(i,i) - \frac{A^{(k-1)}(i,k)}{A^{(k-1)}(k,k)} A^{(k-1)}(k,i) \\ &= A^{(k-1)}(i,i) - \frac{A^{(k-1)}(i,k)^2}{A^{(k-1)}(k,k)} \\ &\leq A^{(k-1)}(i,i) \leq A(i,i), \end{aligned}$$

ya que $A^{(k-1)}(k,k) > 0$ y $A^{(k-1)}(i,k) = A^{(k-1)}(k,i)$ por ser $A^{(k)}(k:n, k:n)$ definida positiva y simétrica.

Por la ultima propiedad probada

$$\left\| A^{(k)}(k:n, k:n) \right\|_{\text{máx}} \leq \|A\|_{\text{máx}}, \quad 1 \leq k \leq n$$

y por consiguiente $\rho_n(A) \leq 1$. □

Ejemplo 5.9. *Aquí también puede suceder que los multiplicadores sean grandes: para $0 < \varepsilon << 1$ se tiene*

$$\begin{bmatrix} \varepsilon^2 & \varepsilon \\ \varepsilon & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1/\varepsilon & 1 \end{bmatrix} \begin{bmatrix} \varepsilon^2 & \varepsilon \\ 0 & 1 \end{bmatrix}.$$

Ejemplo 5.10. *Si no exigimos la simetría de la matriz, el teorema previo puede no verificarse: para $0 < \varepsilon << 1$ se tiene*

$$\begin{bmatrix} \varepsilon & 1 \\ -1 & \varepsilon \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1/\varepsilon & 1 \end{bmatrix} \begin{bmatrix} \varepsilon & 1 \\ 0 & \varepsilon + 1/\varepsilon \end{bmatrix},$$

luego $\rho_n(A) = 1 + 1/\varepsilon >> 1$.

5.3.4. Matrices tridiagonales

Teorema 5.10. *Sea $A \in \mathbb{R}^{n\times n}$ la matriz tridiagonal regular*

$$\begin{bmatrix} d_1 & e_1 & & & \\ c_2 & d_2 & e_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-1} & d_{n-1} & e_{n-1} \\ & & & c_n & d_n \end{bmatrix}$$

La matriz A es factorizable LU si, y sólo si, $u_1 := d_1 \neq 0$ y la fórmula de recurrencia

$$u_i = d_i - c_i e_{i-1}/u_{i-1}, \quad 2 \leq i \leq n,$$

determina números reales no nulos $\{u_i\}_{i=2}^n$. *En dicho caso, la factorización de* A *se corresponde con*

$$L = \begin{bmatrix} 1 & & & & \\ l_2 & 1 & & & \\ & \ddots & \ddots & & \\ & & l_{n-1} & 1 & \\ & & & l_n & 1 \end{bmatrix} \text{ y } U = \begin{bmatrix} u_1 & e_1 & & & \\ & u_2 & e_2 & & \\ & & \ddots & \ddots & \\ & & & u_{n-1} & e_{n-1} \\ & & & & u_n \end{bmatrix},$$

siendo $l_i = c_i/u_{i-1}$, $2 \le i \le n$.

Demostración. Supongamos en primer lugar que A es factorizable LU, es decir, $\det(A_k)0$ $(1 \le k \le n)$. Veamos por inducción sobre k, $2 \le k \le n$, que

$$\det(A_k)/\det(A_{k-1}) = u_k,$$

lo que permite asegurar la condición requerida sobre los números $\{u_i\}_{i=1}^n$. Como

$$\det(A_2) = d_1 d_2 - c_2 e_1 = (d_2 - c_2 e_1/u_1)u_1 = u_2 \det(A_1),$$

el resultado es válido para $k = 2$. Supongamos cierta la relación para $k-1$. En particular podemos construir el número u_k pues u_{k-1} es no nulo. Tenemos,

$$\begin{aligned} \det(A_k) &= d_k \det(A_{k-1}) - c_k e_{k-1} \det(A_{k-2}) = \\ &= (d_k - c_k e_{k-1}/u_{k-1}) \det(A_{k-1}) = u_k \det(A_{k-1}), \end{aligned}$$

luego es cierto para k.

Para la otra implicación, como por la hipótesis es posible construir las matrices L y U, bastará con comprobar que $A = LU$, si bien se puede deducir del razonamiento anterior. En efecto,

$$(LU)(i,i) = \begin{cases} e_i, & i = 1, \\ l_i e_{i-1} + u_i = d_i, & 2 \le i \le n, \end{cases}$$

y

$$(LU)(i,j) = \begin{cases} e_i, & j = i+1,\ 1 \le i \le n-1, \\ l_i u_{i-1} = c_i, & j = i-1,\ 2 \le i \le n, \\ 0, & |i-j| > 1, \end{cases}$$

luego se da la igualdad buscada. □

Algoritmo 5.5 ($Ax = b$, A tridiagonal y factorizable LU)**.** *Guardaremos* L *y* U *en los correspondientes vectores que determinan* A*:* c, d *y* e. *La solución la almacenamos en* b.

Datos: c, d, e, b
(factorización; $3(n-1)$ *flops)*
for $i = 2 : n$
    $c(i-1) = c(i-1)/d(i-1)$
    $d(i) = d(i) - c(i-1) * e(i-1)$

end
(resolución del sistema; $2*(n-1)+1+3(n-1)$ *flops)*
for $i = 2:n$
    $b(i) = b(i) - c(i-1)*b(i-1)$
end
    $b(n) = b(n)/d(n)$
for $i = n-1:-1:1$
    $b(i) = (b(i) - e(i)*b(i-1))/d(i)$
end
Salida: b, c, d, e

Nota 5.18. *El anterior algoritmo requiere* $8n-7$ *flops.*

Ejercicio 5.4. *Probar que utilizando el anterior algoritmo y aritmética de precisión limitada, obtenemos* $\overline{x}$ *que verifica*

$$(A + \triangle A)\overline{x} = b$$

donde

$$|\triangle A| \leq f(u)\left|\overline{L}\right|\left|\overline{U}\right|,$$

donde $f(u)$ *no depende de* n.

Por ejemplo, tenemos

$$\overline{l_i}(1+\delta_1) = c_i/\overline{u_{i-1}}, \quad (1+\nu_1)\overline{u_i} = d_i - \overline{l_i}e_{i-1}(1+\varepsilon_1),$$

con $|\delta_1|, |\nu_1|, |\varepsilon_1| < \gamma_1 = u/(1-u)$. *Así, se obtiene*

$$A = \overline{LU} + \triangle A,$$

con

$$|\triangle A| \leq \gamma_1\left|\overline{L}\right|\left|\overline{U}\right|.$$

Capítulo 6

Métodos iterativos

6.1. Introducción

Consideremos el sistema $Ax = b$ con $A \in \mathbb{R}^{n\times n}$ regular y $b \in \mathbb{R}^n$, cuya solución denotamos por α. Nos planteamos la posibilidad de encontrar $B \in \mathbb{R}^{n\times n}$ y $c \in \mathbb{R}^n$, dependientes de A y b, tales que el sistema de ecuaciones $By + c = y$ tenga solución, única, y coincida con α. Notar que para ello $B - I_n$ debe ser regular. Por ejemplo, podemos reescribir el sistema original como $(A + I_n)x - b = x$, que se correspondería con tomar $B = A + I_n$ y $c = -b$.

En la situación planteada, α sería un punto fijo de la aplicación

$$\begin{array}{rccc} \phi: & \mathbb{R}^n & \longmapsto & \mathbb{R}^n \\ & y & \rightsquigarrow & By + c \end{array}$$

y por tanto, tendríamos un método, el método de punto fijo, para aproximar la solución, que consiste en generar la sucesión

$$x_0 \in \mathbb{R}^n; \quad x_{k+1} = Bx_k + c, \quad k = 0, 1, \ldots \tag{6.1}$$

Definición 6.1. *Un método iterativo asociado al sistema $Ax = b$ consiste en determinar un par (B, c), con $B \in \mathbb{R}^{n\times n}$ y $c \in \mathbb{R}^n$, tales que el sistema de ecuaciones $By + c = y$ sea equivalente al original, y generar la sucesión determinada por (6.1).*

Diremos que el método iterativo determinado por (B, c) es convergente, si para todo $x_0 \in \mathbb{R}^n$ existe el límite de la sucesión generada mediante la fórmula de recurrencia (6.1).

Las cuestiones que nos vamos a plantear en torno a los métodos iterativos son las naturales: como generar diferentes métodos iterativos, si éstos son convergentes o cuando lo son, y comparar métodos convergentes midiendo la rapidez con que en cada caso la sucesión que generamos converge al punto fijo α.

Nota 6.1. *Recordemos el siguiente resultado general.*

Sea (X, d) un espacio métrico completo. Sea $\phi : X \to X$. Si existe $0 < k < 1$ tal que

$$d(\phi(x), \phi(y)) \leq k\, d(x, y), \quad \forall x, y \in X,$$

entonces ϕ tiene un y sólo un punto fijo. Además, la sucesión

$$x_0 \in X; \quad x_{k+1} = \phi(x_k), \quad k = 0, 1, \ldots,$$

converge al punto fijo, para cualquier dato inicial x_0.

Como consecuencia, si para alguna norma matricial subordinada $|||$ se cumple que $\|B\| < 1$, el método iterativo (B, c) será convergente. En el siguiente teorema se precisa este resultado.

Teorema 6.1. *Consideremos el método iterativo (B, c). Equivalen:*

(i) *El método iterativo es convergente,*

(ii) *$r(B) < 1$,*

(ii) *existe una norma matricial subordinada $|||$ sobre $\mathbb{R}^{n\times n}$ tal que $\|B\| < 1$.*

Demostración. (i)⇒(ii) En primer lugar observemos que por inducción es sencillo probar que

$$x_k - \alpha = B^k(x_0 - \alpha), \quad k \geq 1,$$

y por lo tanto, el método es convergente si y sólo si $y \in \mathbb{R}^n$ se cumple

$$\lim_{k\to\infty} B^k y = 0.$$

Supongamos ahora que $r(B) \geq 1$. Entonces existe $\lambda \in \mathbb{R}$ y $v \in \mathbb{R}^n$, no nulo, tales que $Bv = \lambda v$, de donde, como

$$\left\| B^k v \right\| = |\lambda|^k \, \|v\| \geq \|v\|,$$

podemos afirmar que de existir el límite, $\lim_{k\to\infty} B^k v \neq 0$. Si

$$\lim_{k\to\infty} B^k \operatorname{Re} v = \lim_{k\to\infty} B^k \operatorname{Im} v = 0,$$

entonces $\lim_{k\to\infty} B^k v = 0$, por lo que alguno de los anteriores dos límites o no existe o es no nulo, lo que contradice la convergencia del método iterativo.

(ii)⇒(iii) Por reducción al absurdo. Por un resultado ya probado en el tema 2, para todo $\varepsilon > 0$, existe una norma matricial, $|||_\varepsilon$, subordinada a una norma vectorial real tal que

$$1 \leq \|B\|_\varepsilon \leq r(B) + \varepsilon,$$

de donde $1 \leq r(B) + \varepsilon$, y en consecuencia, $1 \leq r(B)$, que contradice (ii).

(iii)⇒(i) Simplemente hay que tener en cuenta que

$$\left\| B^k y \right\| \leq \|B\|^k \, \|y\|,$$

y así, $\lim_{k\to\infty} B^k y = 0$, $\forall y \in \mathbb{R}^n$. □

Creada una sucesión para aproximar la solución de un problema, interesa medir de alguna forma la rapidez con que dicha sucesión se acerca al límite para poder comparar las diferentes sucesiones o métodos y decantarse por uno de ellos. En el caso de sucesiones de números reales $\{r_k\}_{k\geq 1}$ se habla de convergencia a r con velocidad p cuando existe el siguiente límite y se cumple

$$0 < \lim_{k\to\infty} \frac{|r_{k+1} - r|}{|r_k - r|^p}.$$

En el caso de sucesiones de vectores de $\mathbb{R}^n$, el concepto análogo dependería de la norma. Por ello, se introduce la siguiente forma de medir velocidades de sucesiones o más concretamente de un método iterativo.

Definición 6.2. *El factor asintótico de convergencia de un método iterativo convergente (B, c) es el número no negativo*

$$fa(B,c) = \sup_{x_0\in\mathbb{R}^n} \{\limsup_{k\to\infty} \|x_k - \alpha\|^{1/k} : \{x_k\}_{k\geq 1} \text{ generada mediante (6.1)}\}.$$

(Se sobreentiede que x_0 no es α o que $0^{1/k} = 0$)

Nota 6.2. *Recordemos el concepto de límite superior y límite superior de una sucesión de números reales acotada y algunas de sus propiedades elementales. Sea $\{r_k\}_{k\geq 1} \subset \mathbb{R}$.Entonces la sucesiones $\{\text{ínf}_{m\geq k}\, r_m\}_{k\geq 1}$ y $\left\{\sup_{m\geq k} r_m\right\}_{k\geq 1}$ son monótona creciente acotada superiormente y monótona decreciente acotada inferiormente, respectivamente, por lo que existe su límite y se define*

$$\lim_{k\to\infty} \text{ínf}\, r_k = \lim_{k\to\infty} \left(\inf_{m\geq k} r_m\right), \quad \limsup_{k\to\infty} r_k = \lim_{k\to\infty} \left(\sup_{m\geq k} r_m\right).$$

En adelante dejaremos implícito que $k \to \infty$. Se cumple:

(i) *existe* $\lim r_k = r$ *si y sólo si* $\lim \text{ínf}\, r_k = \lim \sup r_k = r$,

(ii) $\lim \sup(r_k + t_k) \leq \lim \sup(r_k) + \lim \sup(t_k)$,

(iii) $\lim \sup(r_k t_k) \leq \lim \sup(r_k) \lim \sup(t_k)$, $r_k, t_k \geq 0$,

(iv) *si existe* $\lim(r_k)$, *entonces* $\lim \sup(r_k t_k) = \lim(r_k) \lim \sup(t_k)$,

(v) *si* $r_k \leq t_k$, *entonces* $\lim \sup(r_k) \leq \lim \sup(t_k)$.

(Si tomamos la sucesión $-1, 1/2, -1, 1/2...$ el límite superior es $1/2$, mientras que el de la sucesión de cuadrados es 1, así que por ejemplo (iii) no es siempre cierto).

Volviendo a nuestra definición, como $\lim \|x_k - \alpha\| = 0$, entonces a partir de cierto índice se cumple que $0 < \|x_k - \alpha\|^{1/k} < 1$, luego existe $\lim \sup \|x_k - \alpha\|^{1/k} \in [0, 1]$, por lo que $0 \leq fa(B, c) \leq 1$.

Analicemos el significado de l concepto introducido. Sea x_0 y $\beta = \lim\sup \|x_k - \alpha\|^{1/k}$. Se tiene que

$$\forall \varepsilon > 0,\, \exists k_0 \in \mathbb{N} : \|x_k - \alpha\| \leq (\beta + \varepsilon)^k,\, \forall k \geq k_0.$$

Por tanto si $\beta + \varepsilon < 1$, lo que se puede conseguir si $\beta < 1$, entonces la convergencia de a cero sería al menos tan rápida como la de la sucesión geométrica $(\beta + \varepsilon)^k$ a cero. El supremo de nuestra definición se toma para reflejar la peor de las situaciones al variar el dato inicial.

Proposición 6.1. *El factor asintótico de convergencia es independiente de la norma escogida en* $\mathbb{R}^n$.

Demostración. Sean $\|\|\|_a$ y $\|\|\|_b$ dos normas sobre $\mathbb{R}^n$. Sabemos que existen constantes $m, M > 0$, tales que

$$m \|x\|_a \leq \|x\|_b \leq M \|x\|_a, \quad \forall x \in \mathbb{R}^n,$$

luego si tenemos $\{x_k\}_{k\geq 1}$ generada mediante (6.1), entonces

$$m^{1/k} \|x_k - \alpha\|_a^{1/k} \leq \|x_k - \alpha\|_b^{1/k} \leq M^{1/k} \|x_k - \alpha\|_a^{1/k},$$

y como $\lim r^{1/k} = 1$, $r > 0$, entonces tomando límites superiores se obtiene

$$\lim\sup \|x_k - \alpha\|_a^{1/k} = \lim\sup \|x_k - \alpha\|_b^{1/k}.$$

□

Teorema 6.2. *Consideremos el método iterativo* (B, c) *con* $r(B) < 1$. *Entonces*

$$fa(B, c) = r(B).$$

Demostración. Por un resultado ya mencionado, para todo $\varepsilon > 0$, existe una norma matricial, $\|\|\|_\varepsilon$, subordinada a una norma vectorial real tal que

$$\|B\|_\varepsilon \leq r(B) + \varepsilon,$$

luego si tenemos $\{x_k\}_{k\geq 1}$ generada mediante (6.1), entonces

$$\|x_k - \alpha\|_\varepsilon^{1/k} \leq \|B\|_\varepsilon \|x_0 - \alpha\|_\varepsilon^{1/k} \leq (r(B) + \varepsilon) \|x_0 - \alpha\|_\varepsilon^{1/k},$$

de donde al tomar límites superiores se obtiene

$$fa(B, c) \leq r(B) + \varepsilon,$$

y en consecuencia, $fa(B, c) \leq r(B)$.

Comprobemos ahora que existe un dato inicial x_0 tal que para la correspondiente sucesión se verifica

$$\lim\sup \|x_k - \alpha\|^{1/k} = r(B). \tag{6.2}$$

(Notar que esta propiedad en particular nos dice que en realidad $fa(B,c)$ es un máximo). Supongamos en primer lugar que existe $\lambda \in \mathbb{R} \cap \sigma(B)$ tal que $|\lambda| = r(B)$. Si v es un vector propio no nulo asociado a λ, tomamos $x_0 = \alpha + v$, y tenemos

$$\|x_k - \alpha\| = \left\|B^k(x_0 - \alpha)\right\| = \left\|B^k v\right\| = |\lambda|^k \|v\|,$$

de donde es ya evidente que se cumple (6.2). Si no se da la situación anterior entonces existe $\lambda \in (\mathbb{C} \setminus \mathbb{R}) \cap \sigma(B)$ tal que $|\lambda| = r(B)$. Sea $v^{(1)}$ un vector propio no nulo asociado a λ, y sea $v^{(2)} = \overline{v^{(1)}}$. Sea $\{v^{(i)}\}_{i=1}^n$ una base de $\mathbb{C}^n$. Para cualquier, $w \in \mathbb{C}^n$, escribimos éste como combinación lineal de la base, $w = \sum_{i=1}^n w_i v^{(i)}$, de forma única, lo que permite introducir una norma sobre $\mathbb{C}^n$, y por tanto sobre $\mathbb{R}^n$, del modo siguiente

$$\|w\| = \sum_{i=1}^{n} |w_i|.$$

Denotamos de igual forma a la correspondiente norma matricial subordinada. Tomamos $x_0 = \alpha + \operatorname{Re} v^{(1)} = \alpha + (v^{(1)} + v^{(2)})/2$. Entonces,

$$\begin{aligned} \|x_k - \alpha\| &= \left\|B^k(v^{(1)} + v^{(2)})/2\right\| \leq \left\|\lambda^k v^{(1)} + \overline{\lambda}^k v^{(2)}\right\|/2 \\ &= (|\lambda|^k + \left|\overline{\lambda}\right|^k)/2 = |\lambda|^k, \end{aligned}$$

de donde se obtiene (6.2). □

Nota 6.3 (Criterios de parada)**.** *Mediante un método iterativo esperamos obtener una aproximación al quedarnos con un término de la sucesión generada. La pregunta natural es cuando aceptamos x_k como aproximación buena de α. Un primer criterio podría ser comprobar que x_k satisface 'bastante' bien el sistema de ecuaciones, es decir,*

$$\|Ax_k - b\| < tol,$$

pero en primer lugar es costoso hacer este cálculo en cada iteración y en segundo ya se vio que no es representativo de que la x_k esté próximo a la solución.

Parece pues más natural considerar el criterio de quedarse con x_k si

$$\|x_{k+1} - x_k\| < tol * \|x_k\|.$$

Analicemos hasta que punto $\|x_{k+1} - x_k\|$ pequeño implica $\|x_k - \alpha\|$ pequeño. Tenemos

$$B(x_k - \alpha) = x_{k+1} - \alpha = x_{k+1} - x_k + x_k - \alpha = B(x_k - x_{k-1}) + x_k - \alpha,$$

de donde

$$(1 - \|B\|) \|x_k - \alpha\| \leq \|B\| \|x_k - x_{k-1}\|,$$

y si $\|B\| < 1$*, entonces*

$$\|x_k - \alpha\| \leq \frac{\|B\|}{1 - \|B\|} \|x_k - x_{k-1}\|,$$

llegándose a

$$\|x_k - \alpha\| \leq \|x_k - x_{k-1}\|,$$

si $\|B\| < 1/2$.

Nota 6.4 (Generación de métodos iterativos)**.** *La idea general consiste en descomponer* $A = M - N$ *con* M *regular, y reescribir el sistema como*

$$x = M^{-1}Nx + M^{-1}b,$$

es decir, tomar $B = M^{-1}N$ *y* $c = M^{-1}b$*. Se requeriría que* $r(M^{-1}N) < 1$*, y se construiría la sucesión mediante*

$$x_0 \in \mathbb{R}^n; \quad Mx_{k+1} = Nx_k + b, \quad k = 0, 1, \ldots,$$

donde en cada iteración se resolvería un sistema de ecuaciones con matriz M*, por lo que además se buscaría* M *de forma que el correspondiente sistema se resolviera con cierta facilidad, por ejemplo, se elegiría triangular.*

En las siguientes secciones introduciremos los métodos iterativos más clásicos como Jacobi, Gauss-Seidel y de relajación. Destacar por último, que los métodos iterativos (existen una gran variedad que no vamos a analizar aquí como métodos basados en subespacios de Krylov, entre los que destaca el método del gradiente conjugado, o la transformada rápida de Fourier, etc.) suelen elegirse frente a los métodos directos para matrices muy grandes con gran número de ceros (sparse) que surgen por ejemplo en la discretización de ecuaciones diferenciales. Una situación típica sería un tamaño de 10^5 *y* 10 *elementos no nulos por fila, estructura que algunos métodos iterativos consiguen explotar.*

6.2. Métodos de Jacobi y Gauss-Seidel

En primer lugar, si A es regular es fácil convencerse de que podemos reorganizar las filas de A de forma que las entradas de la diagonal principal son no nulas. Basta darse cuenta de que debe existir i, $1 \leq i \leq n$, tal que $A(i,1) \neq 0$ y la submatriz A_{i1} es regular, ya que en otro caso el $\det(A)$ es nulo. Obtendríamos el resultado por indución. En adelante supondremos que A verifica dicha propiedad. Usaremos también la descomposición

$$A = D - L - U,$$

donde $D = diag(diag(A))$, $L = -tril(A, -1)$, $U = -triu(A, 1)$.

El método de Jacobi consiste en tomar $M = D$, y $N = L + U$, en el esquema general del final de la sección anterior, lo que da lugar a generar la sucesión

$$x_0 \in \mathbb{R}^n; \quad Dx_{k+1} = (L + U)x_k + b, \quad k = 0, 1, \ldots,$$

o, expresado por componentes,

$$x_{k+1}(i) = (b(i) - \sum_{j=1}^{i-1} A(i,j)x_k(j) - \sum_{j=i+1}^{n} A(i,j)x_k(j))/A(i,i), \quad 1 \leq i \leq n,$$

en cada iteración del método. Notar que cada paso del método requiere $n(2n-1)$ flops. Si tenemos una matriz sparse, como por ejemplo una tridiagnal, este número puede reducirse mucho. En el caso mencionado sólo necesitamos $5n$ flops. Además, el error cometido en un paso del método como consecuencia de la aritmética de precisión limitada, podemos decir que no influye en los siguientes, ya que es como empezar con un dato inicial nuevo.

Para calcular x_{k+1} utilizamos todas las componente de x_k, por lo que debemos matener estas en la memoria hasta que hayamos calculado x_{k+1}. Parece pues razonable obtener $x_{k+1}(i)$ utilizando las ya calculadas $x_{k+1}(j)$, $1 \leq j \leq i-1$, en lugar de las correspondientes componentes de x_k, lo que da lugar al esquema

$$x_{k+1}(i) = (b(i) - \sum_{j=1}^{i-1} A(i,j)x_{k+1}(j) - \sum_{j=i+1}^{n} A(i,j)x_k(j))/A(i,i), \quad 1 \leq i \leq n,$$

que se corresponde con tomar $M = D - L$, y $N = U$, en el esquema general. A este segundo método iterativo se le llama método de Gauss-Seidel.

Definición 6.3. *A las matrices $B_J(A) = D^{-1}(L+U)$ y $B_G(A) = (D-L)^{-1}U$ les llamaremos matriz de Jacobi y de Gauss-Seidel asociadas a A respectivamente.*

La convergencia de los métodos expuestos depende del radio espectral de las matrices introducidas. En los siguientes ejemplos se pone de manifiesto que no hay relación entre la convergencia de un método y el otro, si bien en general cuando convergen ambos Gauss-Seidel suele ser más rápido.

Ejemplo 6.1. (i) *Sea*

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

Entonces

$$\begin{aligned} B_J &= D^{-1}(L+U) = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 0 & 1 & -1 \\ -2 & 0 & -2 \\ 1 & 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1/2 & -1/2 \\ -1 & 0 & -1 \\ 1/2 & 1/2 & 0 \end{bmatrix}, \end{aligned}$$

de donde $\sigma(B_J) = \{0, \pm i\sqrt{5}/2\}$*, por lo que el método de Jacobi no es convergente, mientras que*

$$B_G = (D-L)^{-1}U = \begin{bmatrix} 1/2 & 0 & 0 \\ -1/2 & 1/2 & 0 \\ 0 & 1/4 & 1/2 \end{bmatrix} \begin{bmatrix} 0 & 1 & -1 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1/2 & -1/2 \\ 0 & -1/2 & -1/2 \\ 0 & 0 & -1/2 \end{bmatrix},$$

y por lo tanto, Gauss-Seidel si converge.

(ii) *Sea*

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

Entonces

$$B_J = \begin{bmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{bmatrix},$$

con $\sigma(B_J) = \{0\}$*, lo que implica convergencia, muy rápida, del método de Jacobi. Sin embargo,*

$$B_G = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 0 & -2 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -2 & 2 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{bmatrix},$$

con lo que $r(B_G) = 2$*, lo que implica no convergencia del método de Gauss-Seidel.*

(iii) *Mediante el ejercicio siguiente es fácil encontrar un ejemplo en el que ambos métodos son no convergentes.*

Ejercicio 6.1. *Para* $A \in \mathbb{R}^{2\times 2}$*, Jacobi es convergente si y sólo si lo es Gauss-Seidel, y convergen si y sólo si* $|A(1,2)A(2,1)| < |A(1,1)A(2,2)|$*. Decidir cuál es más rápido.*

Proposición 6.2. *Los métodos de Gauss-Seidel y de Jacobi son convergentes para matrices estrictamente diagonalmente dominantes.*

Demostración. (a) Comencemos analizando el método de Jacobi. Sea $\lambda \in \sigma(B_J(A))$ y $v \in \mathbb{C}^n$ un vector propio asociado no nulo, es decir,

$$\lambda D v = (L+U)v,$$

o equivalentemente,

$$\lambda A(i,i)v(i) = -\sum_{j=1, j\neq i}^{i-1} A(i,j)v(j), \quad 1 \leq i \leq n. \tag{6.3}$$

Perseguimos demostrar que $|\lambda| < 1$ con lo que $r(B_J(A)) < 1$. Distinguimos dos situaciones:

(i) Dominancia por filas.

Sea $|v(i_0)| = \|v\|_\infty$. Entonces tomando módulos en (6.3) para $i = i_0$ tenemos

$$|\lambda|\,|A(i_0,i_0)| \leq \sum_{j=1, j\neq i_0}^{n} |A(i_0,j)| < |A(i_0,i_0)|\,,$$

de donde $|\lambda| < 1$.

(ii) Dominancia por columnas.

Tomando módulos en (6.3), sumando para i y reordenando los sumandos para expresar las sumas por columnas se tiene

$$\begin{aligned} |\lambda| \sum_{i=1}^{n} |A(i,i)|\,|v(i)| &\leq \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} |A(i,j)|\,|v(j)| = \sum_{j=1}^{n} |v(j)| \,(\sum_{i=1, i\neq j}^{n} |A(i,j)|)\,|v(j)| \\ &< \sum_{j=1}^{n} |A(j,j)|\,|v(j)|\,, \end{aligned}$$

luego $|\lambda| < 1$.

(a) Analicemos ahora el método de Gauss-Seidel. Sea $\lambda \in \sigma(B_G(A))$ y $v \in \mathbb{C}^n$ un vector propio asociado no nulo, es decir,

$$\lambda(D - L)v = Uv,$$

o equivalentemente,

$$\lambda A(i,i)v(i) = -\lambda \sum_{j=1}^{i-1} A(i,j)v(j) - \sum_{j=i+1}^{n} A(i,j)v(j), \quad 1 \leq i \leq n. \tag{6.4}$$

Distinguimos de nuevo dos situaciones:

(i) Dominancia por filas.

Sea $|v(i_0)| = \|v\|_\infty$. Entonces tomando módulos en (6.4) para $i = i_0$ tenemos

$$|\lambda|\,|A(i_0,i_0)| \leq |\lambda| \sum_{j=1}^{i_0-1} |A(i_0,j)| + \sum_{j=i_0+1}^{n} |A(i_0,j)|\,,$$

de donde

$$|\lambda| \leq \sum_{j=i_0+1}^{n} |A(i_0,j)|\,/(|A(i_0,i_0)| - \sum_{j=1}^{i_0-1} |A(i_0,j)|) < 1.$$

(ii) Dominancia por columnas.

Tomando módulos en (6.4) y sumando para i se tiene

$$|\lambda| \sum_{i=1}^{n} (|A(i,i)|\,|v(i)| - \sum_{j=1}^{i-1} |A(i,j)|\,|v(j)|) \leq \sum_{i=1}^{n} \sum_{j=i+1}^{n} |A(i,j)|\,|v(j)|$$

y por tanto, reordenando las sumas por columnas y utilizando la dominancia por columnas se tiene

$$\begin{aligned} |\lambda| \sum_{j=1}^{n} (|A(j,j)| - \sum_{i=j+1}^{n} |A(i,j)|)\,|v(j)| &\leq \sum_{j=1}^{n} \sum_{i=1}^{j-1} |A(i,j)|\,|v(j)| \\ &\leq \sum_{j=1}^{n} (|A(j,j)| - \sum_{i=j+1}^{n} |A(i,j)|)\,|v(j)|\,, \end{aligned}$$

luego $|\lambda| < 1$. □

Lema 6.1. *Consideremos las matrices tridiagonales*

$$A(\delta) = \begin{bmatrix} d_1 & \delta^{-1}e_1 & & & \\ \delta c_2 & d_2 & \delta^{-1}e_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \delta c_{n-1} & d_{n-1} & \delta^{-1}e_{n-1} \\ & & & \delta c_n & d_n \end{bmatrix}, \quad \delta \neq 0.$$

Entonces

$$\det(A(\delta)) = \det(A(1)).$$

Demostración. Consideramos la matriz

$$D_\delta = diag([1\,\delta\,\ldots\,\delta^{n-1}]),$$

cuya inversa es $D_{\delta^{-1}}$. Entonces

$$A(\delta) = D_\delta A(1) D_{\delta^{-1}},$$

por lo que sólo resta tomar determinantes en la anterior igualdad para obtener la propiedad anunciada. □

Proposición 6.3. *Sea $A \in \mathbb{R}^{n\times n}$ regular, con $A(i,i) \neq 0$, $1 \leq i \leq n$, y tridiagonal. Entonces*

$$r(B_G(A)) = r(B_J(A))^2.$$

Por lo tanto, el método de Jacobi es convergente si, y sólo si, lo es el método de Gauss-Seidel, y este último es más rápido cuando hay convergencia.

Demostración. Analicemos los esprectros de las matrices de Jacobi y de Gauss-Seidel. Se tiene

$$\begin{aligned}\sigma(B_J) &= \{\lambda \in \mathbb{C} : \det(\lambda - D^{-1}(L+U)) = 0\} \\ &= \{\lambda \in \mathbb{C} : \det(\lambda D - L - U) = 0\} \\ &\equiv \{\lambda \in \mathbb{C} : q_J(\lambda) = 0\},\end{aligned}$$

y

$$\begin{aligned}\sigma(B_G) &= \{\lambda \in \mathbb{C} : \det(\lambda - (D-L)^{-1}U) = 0\} \\ &= \{\lambda \in \mathbb{C} : \det(\lambda D - \lambda L - U) = 0\} \\ &\equiv \{\lambda \in \mathbb{C} : q_G(\lambda) = 0\}.\end{aligned}$$

Para $\lambda \neq 0$ tenemos

$$\begin{aligned}q_G(\lambda^2) &= \det(\lambda^2 D - \lambda^2 L - U) = \det(\lambda(\lambda D - \lambda L - \lambda^{-1}U)) \\ &= \lambda^n \det(\lambda D - \lambda L - \lambda^{-1}U) = \lambda^n \det(\lambda D - L - U),\end{aligned}$$

donde la última igualdad es consecuencia del lema previo. Así,

$$q_G(\lambda^2) = \lambda^n q_J(\lambda), \quad \lambda \in \mathbb{C},$$

ya que $q_G(0) = 0$. En consecuencia, si $\lambda \in \sigma(B_J)$ entonces $\lambda^2 \in \sigma(B_G)$. Además, si $\mu \in \sigma(B_G)$ no nulo, entonces tomando $\lambda \in \sqrt{\mu}$, es decir, $\lambda^2 = \mu$, entonces $\lambda^n q_J(\lambda) = 0$, con $\lambda \neq 0$, por lo que $\lambda \in \sigma(B_J)$. Como $0 \in \sigma(B_G)$, entonces

$$\sigma(B_G) = \{\gamma^2 : \gamma \in \sigma(B_J)\} \cup \{0\},$$

y, por consiguiente,

$$r(B_G) = \max_{\mu \in \sigma(B_G)} |\mu| = \max_{\lambda \in \sigma(B_J)} |\lambda|^2 = (\max_{\lambda \in \sigma(B_J)} |\lambda|)^2 = r(B_J)^2,$$

ya que $\sigma(B_J)$ es no vacío. □

6.3. El método de relajación sucesiva

Consideramos $A \in \mathbb{R}^{n \times n}$ en las condiciones expuestas al principio de la sección anterior. Para $w \in \mathbb{R}$, $w \neq 0$, el método de relajación sucesiva es el método iterativo asociado a la descomposición

$$A = (w^{-1}D - L) - ((1-w)w^{-1}D + U),$$

es decir, generamos una sucesión $\{x_k\}_{k=0}^{\infty}$ mediante la fórmula de recurrencia

$$x_0 \in \mathbb{R}^n; \quad (D - wL)x_{k+1} = ((1-w)D + wU)x_k + wb, \quad k = 0, 1, \ldots,$$

o, escrito por componentes,

$$\begin{aligned} x_{k+1}(i) &= (1-w)x_k(i) + w\,(b(i) - \sum_{j=1}^{i-1} A(i,j)x_{k+1}(j) - \sum_{j=i+1}^{n} A(i,j)x_k(j))/A(i,i) \\ &= (1-w)x_k(i) + w\,\widetilde{x_{k+1}}(i), \quad k = 0,1,\ldots, \quad 1 \le i \le n, \end{aligned}$$

donde $\widetilde{x_{k+1}}$ representa el vector obtenido por Gauss-Seidel a partir de x_k. Así, el nuevo método se puede interpretar como una media ponderada entre el valor anterior y el obtenido por Gauss-seidel a partir de éste. Observar que Gauss-Seidel corresponde a $w = 1$. Algunos autores usan la terminología subrelajación para $w < 1$ y sobrerelajación para $w > 1$.

Definición 6.4. *Para $w \in \mathbb{R}$, $w \neq 0$, llamaremos matriz de relajación asociada a A a la matriz $B_w(A) = (D - wL)^{-1}((1-w)D + wU)$.*

Nos interesa determinar para que valores del parámetro w se tiene convergencia y dentro de estos parámetros determinar aquellos que proporcionen una convergencia más rápida. De entrada tenemos que restringirnos a $0 < w < 2$, ya que fuera de dicho intervalo no puede haber convergencia como expresa la siguiente proposición.

Proposición 6.4. *Se cumple $r(B_w(A)) \ge |w-1|$. Por consiguiente, si el método es convergente, entonces $0 < w < 2$.*

Demostración. Por definición $|\lambda| \le r(B_w(A))$, $\forall \lambda \in \sigma(B_w(A))$, luego

$$|\det(B_w(A))|^{1/n} = \left(\prod_{\lambda \in \sigma(B_w(A))} |\lambda| \right)^{1/n} \le r(B_w(A)),$$

si en el productorio entendemos que contamos cada valor propio tantas veces como indique su orden de multiplicidad como raíz del polinomio característico asociado a A, y donde hemos utilizado que el determinante de una matriz coincide con el producto de todos sus valores propios. El resultado buscado es ya consecuencia de que

$$\det(B_w(A)) = \frac{\det((1-w)D + wU)}{\det(D - wL)} = (1-w)^n.$$

□

De forma completamente análoga a como se razono en la sección previa se tiene el siguiente resultado de convergencia.

Proposición 6.5. *Para $0 < w \le 1$, el método de relajación es convergente para matrices estrictamente diagonalmente dominantes.*

Ejemplo 6.2. *En este ejemplo ponemos de manifiesto que para otros valores de w puede no ser cierto el resultado anterior. Sea*

$$A = \begin{bmatrix} 7 & -2 & 3 \\ -1 & 5 & -2 \\ 1 & 1 & -3 \end{bmatrix}.$$

Cosideramos $w = 3/2$. Tenemos

$$B_{3/2} = \begin{bmatrix} -1/2 & 3/7 & -9/14 \\ -3/20 & -13/35 & 57/140 \\ -13/40 & 1/35 & -173/280 \end{bmatrix},$$

cuyo polinomio característico es

$$P(x) = x^3 + \frac{417}{280}x^2 + \frac{159}{280}x + \frac{1}{8}.$$

Es fácil comprobar que $P(-1)P(-2) < 0$, por lo que $r(B_{3/2}(A)) > 1$.

Proposición 6.6. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica definida positiva. Consideremos el método iterativo asociado a la descomposición $A = M - N$, con M regular y $M' + N$ definida positiva. Entonces el método iterativo es convergente.*

Demostración. Por las características de A, la aplicación $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n \to x'Ay$ define un producto escalar sobre $\mathbb{R}^n$ con norma asociada $\|x\|_A = \sqrt{x'Ax}$, $x \in \mathbb{R}^n$. Denotamos de la misma forma la norma matricial subordinada. La prueba del resultado pasará por mostrar que

$$\left\|M^{-1}N\right\|_A = \left\|I_n - M^{-1}A\right\|_A = \max_{\|x\|_A=1} \left\|x - M^{-1}Ax\right\|_A < 1.$$

Sea $x \in \mathbb{R}^n$ con $\|x\|_A = 1$. Sea $y = M^{-1}Ax \neq 0$. Entonces

$$\begin{aligned} \|x - y\|_A^2 &= \|x\|_A^2 + \|y\|_A^2 - y'Ax - x'Ay \\ &= 1 + \|y\|_A^2 - y'My - y'M'y \\ &= 1 - y'(N + M')y < 1, \end{aligned}$$

por ser $M' + N$ definida positiva. La arbitrariedad de x permite afirmar que $\left\|M^{-1}N\right\|_A < 1$. □

Nota 6.5. *El resultado previo es cierto si pedimos alternativamente que $M + N$ es definida positiva, ya que $\|y\|_A^2 - y'Ax - x'Ay = \|y\|_A^2 - 2y'Ax = -y'(M + N)y < 0$.*

Corolario 6.1. *Sea $A \in \mathbb{R}^{n\times n}$ simétrica definida positiva. El método de relajación es convergente para todo $0 < w < 2$.*

Demostración. Recordar que $M = w^{-1}D - L$, luego M es regular, ya que al ser A definida positiva, $A(i, i) > 0$, $1 \leq i \leq n$. Bastará con comprobar que $M' + N$ es definida positiva. Al ser A simétrica, $L' = U$, luego

$$M' + N = w^{-1}D - L' + w^{-1}(1 - w)D + U = w^{-1}(2 - w)D,$$

que es definida positiva por la elección de w y por tener la matriz diagonal D todas sus entradas en la diagonal principal positivas. □

Ejemplo 6.3. *Es natural preguntarse si el método deJacobi es convergente para matrices simétricas definidas positivas. La respuesta es negativa y para buscar un ejemplo de ello habrá que trabajar, al menos, con matrices de orden* 3*, por supuesto no tridiagonales ni estrictamente diagonalmente dominantes.*

Sea

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

Entonces

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

cuyo polinomio característico es $x^3 - x + \frac{8}{27} = 0$*, y por tanto* $\sigma(B_J) = \{\frac{1}{3}, -\frac{1}{6} \pm \frac{1}{6}\sqrt{33}\}$*, de donde* $r(B_J) = \frac{1}{6} + \frac{1}{6}\sqrt{33} > 1$*.*

En el siguiente resultado se establece un valor óptimo del parámetro w para cierta clase de matrices importante desde el punto de vista de las aplicaciones.

Teorema 6.3. *Sea* $A \in \mathbb{R}^{n\times n}$ *simétrica definida positiva y tridiagonal (válido también para tridiagonales por bloques). La función* $w \in]0, 2[\to r(B_w)$ *tiene un único punto de mínimo en*

$$w_0 = \frac{2}{1 + \sqrt{1 - r(B_J)^2}},$$

siendo el valor mínimo de la función $w_0 - 1$*.*

Capítulo 7

Sistemas sobredeterminados

7.1. Introducción: el problema de mínimos cuadrados, SVD y ecuaciones normales

Consideremos el sistema $Ax = b$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. El problema consiste en expresar b como combinación lineal de las columnas de A. Por tanto, el problema tiene solución si, y sólo si, $b \in \left\langle \{A(:,j)\}_{j=1}^n \right\rangle = ran(A)$. Como la $\dim(ran(A)) = rank(A) \leq p = \min(m,n)$ (recuérdese que $\dim(\ker(A)) + rank(A) = n$), claramente no tiene porqué existir solución en el sentido habitual. Fijamos una norma $\|\|$ sobre $\mathbb{R}^n$ y nos planteamos el problema alternativo de minimizar el residuo repecto de dicha norma; es decir, encontrar $x_0 \in \mathbb{R}^n$ tal que

$$\|Ax_0 - b\| = \inf_{x\in\mathbb{R}^n} \|Ax - b\| .$$

Ejemplo 7.1. *Sean $b_1 \leq b_2 \leq b_3$. Consideramos el sistema sobredeterminado*

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} x = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix},$$

que tiene solución si, y sólo si, $b_1 = b_2 = b_3$ y ésta es $x = b_1$. Nos planteamos tres problemas alternativos:

(i) *Sea $\phi(x) = \|Ax - b\|_2^2 = \sum_{i=1}^3 (x - b_i)^2$, $x \in \mathbb{R}$. Vamos a estudiar la existencia de mínimo de ϕ. Como $\phi'(x) = 6x - 2\sum_{i=1}^3 b_i$, entonces $x_0 = \sum_{i=1}^3 b_i/3$ es un candidato a mínimo de ϕ. Como $\phi''(x_0) = 6 > 0$, entonces estamos ante un mínimo con $\phi(x_0) = 2(\sum_{i=1}^3 {b_i}^2 - b_1b_2 - b_1b_3 - b_2b_3)/3$.*

(ii) *Sea $\gamma(x) = \|Ax - b\|_1 = \sum_{i=1}^3 |x - b_i|$, $x \in \mathbb{R}$. Vamos a estudiar la existencia de mínimo de γ. Como*

$$\gamma'(x) = \begin{cases} -3, & x < b_1, \\ -1, & b_1 < x < b_2, \\ 1, & b_2 < x < b_3, \\ 3, & b_3 < x, \end{cases}$$

entonces γ pasa de decreciente a creciente en $x_0 = b_2$, por lo que x_0 es un mínimo de γ con $\gamma(x_0) = |b_2 - b_1| + |b_2 - b_3|$.

(iii) *Sea $\eta(x) = \|Ax - b\|_\infty = \max_{1\leq i\leq 3} |x - b_i|$, $x \in \mathbb{R}$. En $x_0 = (b_1 + b_3)/2$, punto medio de los valores extremos de los b_i, se alcanza un mínimo con $\eta(x_0) = (b_3 - b_1)/2$. (Es fácil convencerse de la certeza de las afirmaciones observando que b_2 no interviene en la discusión)*

En el ejemplo anterior se observa la dificultad de trabajar con normas distintas de la euclídea; en contraposición, el cuadrado de ésta es derivable lo que habilita el uso de técnicas del análisis. Además, la norma euclídea tiene la ventaja adicional de ser invariante por transformaciones ortogonales, es decir, para $Q \in \mathbb{R}^{n\times n}$, ortogonal, se tiene

$$\|Ax - b\|_2 = \|QAx - Qb\|_2 ,$$

lo que sugiere buscar transformaciones ortogonales que simplifiquen nuestro sistema en algún sentido.

Definición 7.1. *El problema de mínimos cuadrados asociado al sistema $Ax = b$ es el de obtener $x_0 \in \mathbb{R}^n$ tal que*

$$\|Ax_0 - b\|_2 = \min_{x\in\mathbb{R}^n} \|Ax - b\|_2 . \tag{7.1}$$

Nota 7.1. *Notar que $\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2 = (\min_{x\in\mathbb{R}^n} \|Ax - b\|_2)^2$, por lo que a todos los efectos podemos considerar la norma al cuadrado.*

Sea $A = U\Sigma V'$ una SVD de A. Entonces

$$Ax - b = U\Sigma V'x - b \underset{c=U'b}{\overset{y=V'x}{=}} U(\Sigma y - c)$$

luego

$$\|Ax - b\|_2 = \|\Sigma y - c\|_2$$

Por tanto, la SVD nos ha permitido realizar un cambio de variables de forma que el (7.1) se ha reducido a forma diagonal,

$$\min_{y\in\mathbb{R}^n} \|\Sigma y - c\|_2 ,$$

fácilmente resoluble. Si $rank(A) = k \leq p$, entonces

$$\begin{aligned} \|\Sigma y - c\|_2^2 &= \|[\mu_1 y(1) - c(1); \ldots ; \mu_k y(k) - c(k); -c(k+1:m)]\|_2^2 \\ &= \sum_{i=1}^{k} (\mu_i y(i) - c(i))^2 + \sum_{i=k+1}^{m} c(i)^2, \end{aligned}$$

que se minimiza cuando

$$y(i) = c(i)/\mu_i,\ 1 \leq i \leq k,$$

cometiéndose un error

$$\left(\sum_{i=k+1}^{m} c(i)^2 \right)^{1/2}.$$

Nótese que hemos probado **existencia** de solución de (7.1). Observar que las componentes $y(k+1:n)$ están indeterminadas, y pueden asignarse arbitrariamente sin efecto sobre la longitud de $\Sigma y - c$. Así, si $k = n$, tenemos unicidad (nótese que en este caso $m \geq n$), y si $k = m$, solución exacta del sistema de ecuaciones. Sumarizamos estos hechos en el siguiente teorema.

Teorema 7.1 (Existencia y unicidad de solución del LSP)**.** *Sea $A = U\Sigma V'$ una SVD de A. Sea $k = rank(A) \leq p$. Si $y \in \mathbb{R}^n$ es cualquier vector cumpliendo*

$$y(i) = (U'b)(i)/\mu_i \quad (1 \leq i \leq k),$$

entonces $x = Vy$ es una solución del problema de mínimos cuadrados (7.1), con valor del mínimo

$$\left\| (U'b)(k+1:m) \right\|_2.$$

En particular, si $k = n$, entonces se tiene unicidad de solución del (7.1), y si $k = m$, las soluciones del (7.1) son soluciones exactas del sistema de ecuaciones lineales.

El proceso a seguir para obtener una solución consistiría en la siguiente secuencia de cálculos:

$$A = U\Sigma V' \rightarrow c = U'b;\ k = rank(A) \rightarrow \left\{ \begin{array}{c} y(i) = c(i)/\mu_i,\ 1 \leq i \leq k \\ y(i) = 0,\ k+1 \leq i \leq n \end{array} \right. \rightarrow x = Vy$$

Esta secuencia se escribe de forma compacta como

$$x = V\Sigma^+ U'b$$

donde Σ^+ se obtiene transponiendo Σ y luego invirtiendo los elementos no nulos de la diagonal principal, es decir,

$$\Sigma^+ = diag([1/\mu_1; \ldots; 1/\mu_k; 0; \ldots; 0]) \in \mathbb{R}^{n \times m}.$$

A la matriz

$$A^+ = V\Sigma^+ U'$$

se le llama **inversa generalizada de Moore-Penrose** de A. Si A es regular, $A^+ = A^{-1}$.

Ejemplo 7.2 (LSP)**.** *Tenemos la siguiente descomposición SVD:*

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \\ 1/\sqrt{6} & -1/\sqrt{2} & 1/\sqrt{3} \\ 2/\sqrt{6} & 0 & -1/\sqrt{3} \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}'.$$

Como ya sabíamos, A tiene rango máximo. La pseudoinversa de A es

$$\begin{aligned}
A^+ &= \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \\ 1/\sqrt{6} & -1/\sqrt{2} & 1/\sqrt{3} \\ 2/\sqrt{6} & 0 & -1/\sqrt{3} \end{bmatrix}' \\
&= \begin{bmatrix} -\frac{1}{6}\sqrt{2}\sqrt{3} & \frac{1}{2}\sqrt{2} & 0 \\ \frac{1}{6}\sqrt{2}\sqrt{3} & \frac{1}{2}\sqrt{2} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{6}\sqrt{6} & \frac{1}{6}\sqrt{6} & \frac{1}{3}\sqrt{6} \\ \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} & 0 \\ \frac{1}{3}\sqrt{3} & \frac{1}{3}\sqrt{3} & -\frac{1}{3}\sqrt{3} \end{bmatrix} \\
&= \begin{bmatrix} -\frac{1}{36}\sqrt{6}\sqrt{3}\sqrt{2}+\frac{1}{2} & -\frac{1}{36}\sqrt{6}\sqrt{3}\sqrt{2}-\frac{1}{2} & -\frac{1}{18}\sqrt{6}\sqrt{3}\sqrt{2} \\ \frac{1}{36}\sqrt{6}\sqrt{3}\sqrt{2}+\frac{1}{2} & \frac{1}{36}\sqrt{6}\sqrt{3}\sqrt{2}-\frac{1}{2} & \frac{1}{18}\sqrt{6}\sqrt{3}\sqrt{2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{1}{3} \end{bmatrix}.
\end{aligned}$$

Nos planteamos los siguientes sistemas sobre determinados.

- $Ax = [1; -1; 0]$

  Nótese que $A^+[1; -1; 0] = [1; 1]$.

  $$c = \begin{bmatrix} \frac{1}{6}\sqrt{6} & \frac{1}{6}\sqrt{6} & \frac{1}{3}\sqrt{6} \\ \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} & 0 \\ \frac{1}{3}\sqrt{3} & \frac{1}{3}\sqrt{3} & -\frac{1}{3}\sqrt{3} \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \sqrt{2} \\ 0 \end{bmatrix}$$

  luego la única solución de LSP es

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

  y verifica exactamente el sistema planteado.

- $Ax = [1; 1; 1]$.

  $$c = \begin{bmatrix} \frac{1}{6}\sqrt{6} & \frac{1}{6}\sqrt{6} & \frac{1}{3}\sqrt{6} \\ \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} & 0 \\ \frac{1}{3}\sqrt{3} & \frac{1}{3}\sqrt{3} & -\frac{1}{3}\sqrt{3} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{2}{3}\sqrt{6} \\ 0 \\ \frac{1}{3}\sqrt{3} \end{bmatrix}$$

  luego la única solución de LSP es

  $$x = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \frac{2}{3}\sqrt{2} \\ 0 \end{bmatrix} = \begin{bmatrix} -\frac{2}{3} \\ \frac{2}{3} \end{bmatrix}$$

  con

  $$\left\| A[-\frac{2}{3}; \frac{2}{3}] - [1; 1; 1] \right\|_2 = \frac{\sqrt{3}}{3}.$$

Nota 7.2. *Sea* $f(x) = \|Ax - b\|_2^2 = \sum_{i=1}^{m} (\sum_{j=1}^{n} A(i,j)x(j) - b(i))^2$. *Un punto crítico de* f, *en particular un mínimo de* f, *debe anular todas las*

parciales, es decir,

$$\begin{aligned} 0 &= \sum_{i=1}^{m} A(i,k)(\sum_{j=1}^{n} A(i,j)x(j) - b(i)) \\ &= \sum_{i=1}^{m} A'(k,i)(Ax-b)(i),\ 1 \le k \le n, \end{aligned}$$

o lo que es equivalente

$$A'(Ax-b) = 0.$$

Definición 7.2. *El sistema de ecuaciones normales asociado al problema de mínimos cuadrados es*

$$A'Ax = A'b.$$

Recordar que $A'A \in \mathbb{R}^{n\times n}$ es simétrica semidefinida positiva, y que es definida positiva si, y sólo si, A es inyectiva. Si $m \ge n$, entonces $rank(A) = n$ equivale a A inyectiva, pero si $m < n$, entonces $rank(A) = m$, no implica A inyectiva; en efecto,

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

Nota 7.3. *Si* $r_x = \|Ax-b\|_2$, *entonces*

$$r_x^2 = r_y^2 + \|A(x-y)\|_2^2 + 2(x-y)'A'(Ay-b).$$

Teorema 7.2. *Toda solución del sistema de ecuaciones normales es una solución del (7.1) y viceversa.*

Demostración. ea x_0 una solución de las ecuaciones normales. Por la nota anterior

$$r_x^2 = r_{x_0}^2 + \|A(x-y)\|_2^2 \ge r_{x_0}^2,\ \forall x \in \mathbb{R}^n.$$

Recíprocamente, si x_0 una solución de (7.1), entonces es un punto crítico de la función $f(x) = \|Ax-b\|_2^2$, $x \in \mathbb{R}^n$, luego verifica las ecuaciones normales. □

Ejemplo 7.3 (Inestabilidad al formar las ecuaciones normales)**.** *En MATLAB se tiene*

$$A = \begin{bmatrix} 1 & 1 \\ eps & 0 \\ 0 & eps \end{bmatrix} \rightarrow A'A \doteq \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

por lo que no podemos obtener la única solución de LSP de los sistemas sobredeterminados asociados a A resolviendo las ecuaciones normales asociadas.

7.2. La factorización QR

Consideremos $Ax = b$ con $A \in \mathbb{R}^{m\times n}$, $m \geq n$, y $b \in \mathbb{R}^m$. Supongamos que disponemos de una matriz $Q \in \mathbb{R}^{m\times m}$ ortogonal tal que $R = Q'A$ es triangular superior, es decir, $R = [R_1; zeros(m-n, n)]$ con $R_1 \in \mathbb{R}^{n\times n}$ triangular superior. Observar que lo que estamos solicitando es poder actorizar A en la forma $A = QR$. Bajo esta suposición, podemos pasar al sistema sobredeterminado equivalente $Rx = Q'b = [c; d]$, con $c \in \mathbb{R}^n$.

Por la invarianza de la norma euclídea bajo transformaciones ortogonales tenemos

$$\|Ax - b\|_2^2 = \left\|Rx - Q'b\right\|_2^2 = \|R_1x - c\|_2^2 + \|d\|_2^2\,,$$

de donde

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2 = \min_{x\in\mathbb{R}^n} \|R_1x - c\|_2^2 + \|d\|_2^2\,,$$

es decir, tienen el mismo punto de mínimo, la única diferencia reside en el valor del mínimo que varía de un caso al otro en la cantidad $\|d\|_2^2$.

Situémonos en el caso de $rg(A) = n$ y comprobemos que entonces R_1 es regular, con lo que el correspondiente problema de mínimos cuadrados se traduce en resolver un sistema de ecuaciones triangular. Si $R_1y = 0$, entonces $Q'Ay = 0$, de donde $Ay = 0$, lo que implica $y = 0$ ya que A es inyectiva.

Veamos como alcanzar la situación descrita mediante una clase de matrices, aunque existen otras (por ejemplo las matrices de Givens) que permiten llegar al mismo punto.

Definición 7.3. *A las matrices $H_v \in \mathbb{R}^{m\times m}$ de la forma*

$$H_v = I_m - 2\frac{vv'}{v'v},$$

con $v \in \mathbb{R}^m$, no nulo, les llamaremos matrices de Householder.

Observar que son matrices simétricas y además ortogonales ya que

$$\begin{aligned} H_v'H_v &= H_v^2 = I_m - 4\frac{vv'}{v'v} + 4\frac{vv'vv'}{(v'v)^2} \\ &= I_m - 4\frac{vv'}{v'v} + 4\frac{vv'}{v'v} = I_m. \end{aligned}$$

Notar también que $H_{\alpha v} = H_v$, $\alpha \in \mathbb{R}$.

Estudiemos el número de flops que requiere H_vx con $x \in \mathbb{R}^m$. Calcularíamos $\alpha = -2/(v'v)$ que requiere $2m$ flops, luego $\beta = \alpha v'x$ que requiere otros $2m$ flops para pasar a calcular $w = x - \beta v$ que requiere de otros $2m$ flops. En total, $6m$ flops. Si queremos realizar H_vA, $A \in \mathbb{R}^{mxn}$, necesitaremos realizar un total de $(4n + 2)m$ flops. Observar que para realizar estas operaciones no necesitamos conocer explícitamente la matriz H_v, si no tan sólo v.

Sea $x \in \mathbb{R}^m$, tal que $x(2 : m)$ es no nulo. Veamos que es posible elegir v de forma que el vector H_vx tiene sus entradas nulas excepto, a lo sumo, la primera. Si se diera la condición buscada, existiría $c \in \mathbb{R}$, no nulo, tal que

$$x - 2v'xv/v'v = H_vx = ce_1,$$

de donde despejando, ya que $v'x \neq 0$,

$$v = \frac{2v'v}{v'x}(x - ce_1).$$

Por la propiedad mencionada de invarianza de las matrices de Householder bajo multiplicación del vector generador por un escalar, podríamos tomar $v = x - ce_1$, con $c \in \mathbb{R}$ a determinar. Además,

$$\|x\|_2 = \|H_v x\|_2 = |c|\,,$$

por lo que podríamos tomar $c = \pm\|x\|_2$. Partamos de $v = x \pm \|x\|_2 e_1$. Se tiene

$$v'v = 2(\|x\|_2^2 \pm \|x\|_2 x(1)) \text{ y } v'x = \|x\|_2^2 \pm \|x\|_2 x(1),$$

por lo que en ambos casos

$$H_v x = x - (x \pm \|x\|_2 e_1) = \mp\|x\|_2 e_1,$$

como deseabamos. Para evitar obtener v nulo, ya que $v'v = 2\|x\|_2(\|x\|_2 \pm x(1))$, elegiremos el signo de forma que $\pm x(1) \geq 0$. En resumen, tomamos

$$v_x = x + sign(x(1))\|x\|_2 e_1 = [x(1) + sign(x(1))\|x\|_2; x(2); \ldots; x(m)],$$

con lo que

$$H_v x = -sign(x(1))\|x\|_2 e_1 = [-sign(x(1))\|x\|_2; 0; \ldots; 0].$$

Notar que lo que acabamos de decir es válido para cualquier $x \in \mathbb{R}^m$ no nulo.

Ejemplo 7.4. *Consideramos $x = [-2; 2; -1]$ con $\|x\|_2 = 3$. Entonces $v_x = x - 3e_1 = [-5; 2; -1]$ y calculamos*

$$H_v x = [3; 0; 0]\,.$$

Realicemos el mismo cálculo a partir de la matriz H_v. Tenemos

$$\begin{aligned} H_v &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{2}{30}\begin{bmatrix} -5 \\ 2 \\ -1 \end{bmatrix}\begin{bmatrix} -5 & 2 & -1 \end{bmatrix} \\ &= \begin{bmatrix} -\frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \\ \frac{2}{3} & \frac{11}{15} & \frac{2}{15} \\ -\frac{1}{3} & \frac{2}{15} & \frac{14}{15} \end{bmatrix}, \end{aligned}$$

luego $H_v x = [3; 0; 0]$.

Analicemos ahora como es la matriz H_v para $v = [zeros(k, 1), w]$ con $w \in \mathbb{R}^{m-k}$, no nulo, y $1 \leq k < m$. Entonces para $0 \leq j \leq k$ tenemos

$$H_v(:, j) = e_j - \frac{2}{v'v}vv(j) = e_j,$$

y para $k+1 \leq i, j \leq m$,

$$\begin{aligned} H_v(i,j) &= e_j(i) - \frac{2}{v'v} v(i)v(j) = \delta_{ij} - \frac{2}{w'w} w(i-k)w(j-k) \\ &= \delta_{(i-k)(j-k)} - \frac{2}{w'w} w(i-k)w(j-k) = H_w(i-k, j-k), \end{aligned}$$

de donde por la simetría de se puede concluir que

$$H_v = \begin{bmatrix} I_k & 0 \\ 0 & H_w \end{bmatrix}.$$

En consecuencia, $H_v x = [x(1:k); H_w x(k+1:m)]$. Por lo tanto, para cualquier $x \in \mathbb{R}^m$, no nulo, podemos obtener un vector $v \in \mathbb{R}^m$, no nulo, de forma que $H_v x$ conserve inalteradas las primeras $0 \leq k \leq m-2$ entradas de x, y que anule las entradas desde la $k+2$ hasta m. En efecto, si tomamos

$$\begin{aligned} v &= [zeros(k,1); v_{x(k+1:n)}] \\ &= [zeros(k,1); [x(k+1) + sign(x(k+1)) \left\|x(k+1:n)\right\|_2 ; x(k+2); \ldots ; x(m)]] \end{aligned}$$

tendríamos

$$H_v x = [x(1:k); -sign(x(k+1)) \left\|x(k+1:n)\right\|_2 ; zeros(m-k-2,1)].$$

Es interesante notar que para $y \in \mathbb{R}^m$, con $y(k+1:m)$ nulo, la anterior matriz dejaría y inalterado ya que

$$H_v y = \begin{bmatrix} I_k & 0 \\ 0 & H_w \end{bmatrix} y = y.$$

Con lo anteriormente expuesto, podemos triangularizar una matriz $A \in \mathbb{R}^{m \times m}$,de rango máximo, mediante matrices de Householder. En el primer paso se considera la matriz H_v que haga ceros en la primera columna desde la posición segunda en adelante. Se transforman cada una de las columnas de A (insistir en que para este proceso nunca calculamos las matrices ni hacemos productos de matrices). A continuación hacemos nulas las entradas de la segunda columna desde la tercera posición en adelante. La tranformación elegida no modifica la primera columna de A, por lo que se conserva la estructura triangular antes conseguida. Se transforman las columnas de la tercera en adelante. Se sigue con el proceso, n etapas, hasta hacer la matriz triangular superior. En realidad en la etapa k-ésima bastaría con trabajar con la matriz correspondiente a las posiciones $(k:m, k:m)$.

Veamos el proceso en un ejemplo concreto para pasar luego a exponerlo de forma genérica.

Ejemplo 7.5. *Sean* $A = \begin{bmatrix} 0 & 2 \\ 0 & 0 \\ -1 & -2 \end{bmatrix}$ *y* $b = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$, *y* $W = [A, b]$. *Tendremos que realizar dos etapas de las anteriormente descritas:*

(i) $v_1 = v_{[0;0;-1]} = [1; 0; -1]$.

$H_{v_1}W(:,1) = [-1;0;0]$.

$H_{v_1}W(:,2) = [2;0;-2]-[1;0;-1]'[2;0;-2][1;0;-1] = [2;0;-2]-[4;0;-4] = [-2;0;2]$.

$H_{v_1}W(:,3) = [1;1;0] - [1;0;-1]'[1;1;0][1;0;-1] = [1;1;0] - [1;0;-1] = [0;1;1]$.

$$W^{(1)} = H_{v_1}W = \begin{bmatrix} -1 & -2 & 0 \\ 0 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix}.$$

$$H_{v_1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

(ii) $v_2 = [0; v_{[0;2]}] = [0;2;2]$. *(En realidad podemos trabajar con* $\widetilde{W^{(1)}} = W^{(1)}(2:3, 2:3)$*)*

$H_{v_2}W^{(1)}(:,2) = [0;-2;0]$.

$H_{v_2}W^{(1)}(:,3) = [0;1;1] - 4^{-1}[0;2;2]'[0;1;1][0;2;2] = [0;-1;-1]$.

$$W^{(2)} = H_{v_2}W^{(1)} = \begin{bmatrix} -1 & -2 & 0 \\ 0 & -2 & -1 \\ 0 & 0 & -1 \end{bmatrix}.$$

$$H_{v_2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \tfrac{1}{4}\begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix} \begin{bmatrix} 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}.$$

Sea $R = W^{(2)}(:, 1:2)$. *Notar que* $A = QR$ *con*

$$Q = H_{v_1}H_{v_2} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Para obtener la solución del problema de mínimos cuadrados asociado al sistema $Ax = b$ *resolvemos*

$$\begin{bmatrix} -1 & -2 \\ 0 & -2 \end{bmatrix} x = \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

cuya solución es $[-1; 1/2]$. *Entonces,*

$$\|A[-1;1/2] - b\|_2^2 = \min_{x\in\mathbb{R}^3} \|Ax - b\|_2^2 = 1.$$

Por otra parte el sistema de ecuaciones normales es

$$\begin{bmatrix} 1 & 2 \\ 2 & 8 \end{bmatrix} x = \begin{bmatrix} 0 \\ 2 \end{bmatrix},$$

cuya solución es, como sabemos, la misma del sistema anterior.

Nota 7.4. *La factorización QR no es única. El propio proceso descrito lo pone de manifiesto.*

Nota 7.5. *Sii añadimos la condición de que* $R(i,i) > 0$, $1 \le i \le n$, *¿es cierto que la factorización QR es única?*

Ejemplo 7.6. *Sea $0 < \varepsilon < 1$. Trabajaremos en un sistema en punto flotante para el que $1+\varepsilon^2 \doteq 1$. Consideremos el sistema sobredeterminado con matriz ampliada*

$$W = \begin{bmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \end{bmatrix}.$$

Este sistema no tiene solución. La solución del sistema de ecuaciones normales

$$\begin{bmatrix} 1+\varepsilon^2 & 1 \\ 1 & 1+\varepsilon^2 \end{bmatrix} x = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

es

$$x = \begin{bmatrix} \frac{1}{2+\varepsilon^2} \\ \frac{1}{2+\varepsilon^2} \end{bmatrix} \doteq \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

Utilicemos ahora la factorización QR.

(i) $v_1 = v_{[1;\varepsilon;0]} = [2;\varepsilon;0]$, pues $(1+\varepsilon^2)^{1/2} \doteq 1$. Además $\|v_1\|_2^2 = 4+\varepsilon^2 \doteq 4$.

$H_{v_1}W(:,1) = [-1;0;0]$.

$H_{v_1}W(:,2) = [1;0;\varepsilon] - 2^{-1}[2;\varepsilon;0]'[1;0;\varepsilon][2;\varepsilon;0]' = [1;0;\varepsilon] - [2;\varepsilon;0] = [-1;-\varepsilon;\varepsilon]$.

$H_{v_1}W(:,3) = [1;0;0] - 2^{-1}[2;\varepsilon;0]'[1;0;0][2;\varepsilon;0]' = [1;0;0] - [2;\varepsilon;0] = [-1;-\varepsilon;0]$.

$$W^{(1)} = H_{v_1}W = \begin{bmatrix} -1 & -1 & -1 \\ 0 & -\varepsilon & -\varepsilon \\ 0 & \varepsilon & 0 \end{bmatrix}.$$

(ii) $v_2 = [0; v_{[-\varepsilon;\varepsilon]}] = [0; -\varepsilon(\sqrt{2}+1); \varepsilon] \equiv [0; -(\sqrt{2}+1); 1]$. $\|v_2\|_2^2 = 4+2\sqrt{2}$

$H_{v_2}W^{(1)}(:,2) = [-1;\sqrt{2}\varepsilon;0]$.

$H_{v_2}W^{(1)}(:,3) = [-1;\varepsilon\frac{\sqrt{2}+1}{\sqrt{2}+2};-\varepsilon\frac{\sqrt{2}+1}{\sqrt{2}+2}]$.

$$W^{(2)} = H_{v_2}W^{(1)} = \begin{bmatrix} -1 & -1 & -1 \\ 0 & \sqrt{2}\varepsilon & \varepsilon\frac{\sqrt{2}+1}{\sqrt{2}+2} \\ 0 & 0 & -\varepsilon\frac{\sqrt{2}+1}{\sqrt{2}+2} \end{bmatrix}.$$

Para obtener la solución del problema de mínimos cuadrados asociado al sistema $Ax = b$ resolvemos

$$\begin{bmatrix} -1 & -1 \\ 0 & \sqrt{2}\varepsilon \end{bmatrix} x = \begin{bmatrix} -1 \\ \varepsilon\frac{\sqrt{2}+1}{\sqrt{2}+2} \end{bmatrix},$$

cuya solución es

$$x(2) = \frac{\sqrt{2}+1}{\sqrt{2}(\sqrt{2}+2)} = \frac{1}{2},\ x(1) = 1 - \frac{1}{2} = \frac{1}{2}.$$

Nota 7.6 ([4, pág. 123-124])**.** *La estabilidad del método utilizando matrices ortogonales se debe a que cuando hacemos XA para una matriz X cualquiera*

con aritmética de precisión limitada, se puede probar que lo que obtenemos es

$$fl(\widetilde{X}A) = XA + E = X(A + X^{-1}E) = X(A + \Delta A),$$

donde $\widetilde{X}$ *representa la representación en punto flotante de la matriz, y* $\|E\|_2 \leq O(u)\|X\|_2$. *Entonces,*

$$\|\Delta A\|_2 \leq \left\|X^{-1}\right\|_2 \|E\|_2 \leq O(u)k_2(X),$$

es decir, $k_2(X)$ *es el factor que amplifica los errores. Por lo tanto, conseguimos no amplificar el error si* $k_2(X) = 1$, *lo que sucede con las matrices ortogonales.*

Ejercicio 7.1. *La cuestión que surje de forma natural es: ¿*$k_2(X) = 1$ *implica que* X *es un matriz ortogonal?*

Capítulo 8

Cálculo de Valores y de Vectores Propios

La notación y resultados básicos relativos a estos conceptos fueron introducidos en el apartado 2.3. Vamos a introducir y analizar una serie de aspectos relacionados con el tema que nos ocupa. Sea $A \in \mathbb{R}^{n\times n}$.

(i) Obtención de los coeficientes del polinomio característico de A.

Se puede probar que los coeficientes dependen continuamente de A. Por Cayley-Hamilton sabemos que A verifica la ecuación característica

$$A^n + a_{n-1}A^{n-1} + \ldots + a_1 A + a_0 = 0.$$

En particular, si tomamos $x_0 \in \mathbb{R}^n$ llegamos a que los coeficientes del polinomio característico satisfacen el sistema de acuaciones lineales

$$\left[x_0,\ Ax_0,\ \ldots,\ A^{n-1}x_0\right] z = -A^n x_0.$$

Esta técnica se conoce como método de Krylov, y es claro que presenta problemas dependiendo de la elección de x_0. Existen otros métodos como el método de Leverrier_Faddeev, que en particular también proporciona la matriz inversa, si es que estamos interesada en ella.

(ii) Valores propios como ceros de funciones.

Obtener los valores propios de una matriz se puede interpretar como la resolución de la ecuación característica $P_A(\lambda) = 0$. Podemos aplicar el método de la secante: elegimos λ_0, λ_1 dos aproximaciones iniciales de una valor propio, y generamos la sucesión

$$\lambda_{m+1} = \lambda_m - \frac{\lambda_m - \lambda_{m-1}}{P_A(\lambda_m) - P_A(\lambda_{m-1})} P_A(\lambda_m) \ \ (m = 1, 2, \ldots),$$

donde $P_A(\lambda_m)$ es el determinante de una matriz, que recordemos puede evaluarse mediante el método de eliminación de Gauss. No necesitamos pues conocer los coeficientes del polinomio característico.

(iii) Localización de los valores propios.

En primer lugar tenemos que $r(A) \leq \|A\|_p$, $1 \leq p \leq \infty$. En la misma dirección tenemos:

Teorema 8.1 (Gerschgorin)**.** *Sea $A \in \mathbb{R}^{n\times n}$. Sea D_i el disco cerrado de centro $A(i,i)$ y de radio $\sum_{j=1,j\neq i}^{n} |A(i,j)|$, $1 \leq i \leq n$. Entonces*

$$\sigma(A) \subset \bigcup_{i=1}^{n} D_i.$$

Además, si la reunión, Ω, de r discos de los anteriores resulta ser un conjunto conexo disjunto de la reunión, Δ, del resto de discos, entonces en Ω hay r valores propios de A y en Δ hay $n-r$ valores propios de A.

(iv) Transformaciones de semejanza.

Estas transformaciones conservan los valores propios, aunque no sucede lo mismo con los vectores propios asociados, si bien existe una relación sencilla que los relaciona si conocemos la transformación de semejanza.

Se puede probar que toda matriz $A \in \mathbb{R}^{n\times n}$ es semejante por transformaciones ortogonales a una matriz Hessenberg superior (tridiagonal si A es simétrica), de ahí que sea suficiente con tratar esta clases de matrices.

(v) Comportamiento de valores propios bajo perturbaciones de los datos.

Si bien los valores propios dependen continuamente de las entradas de la matriz, en la práctica se pueden presentar situaciones anómalas como la del siguiente ejemplo.

Ejemplo 8.1. *Sea $\delta \geq 0$. Sea $A_\delta \in \mathbb{R}^{16\times 16}$ cuya diagonal principal tiene todas sus entradas igual a 2, la diagonal de posición $k = 1$ tiene todas sus entradas igual a 1, y en la posición $(16,1)$ tenemos el valor δ. Obviamente, el único valor propio de A_0, con multiplicidad algebraica 16, es 2. Sin embargo, los valores propios de A_δ distan todos $\delta^{1/16}$ de 2, de manera que para $\delta = 10^{-16}$ la distancia es de 0.1. Un error absoluto de un dato de entrada de 10^{-16} ha provocado un error absoluto de 0.1 en la salida.*

El siguiente resultado pone de manifiesto que esto no sucede en el caso de matrices simétricas; es decir, el problema que nos ocupa es, en este caso, bien condicionado.

Teorema 8.2. *Sea $A \in \mathbb{R}^{n\times n}$ tal que existe $R \in \mathbb{C}^{n\times n}$ regular con $D := R^{-1}AR$ diagonal. Sea $\Delta A \in \mathbb{R}^{n\times n}$. Si $\mu \in \sigma(A + \Delta A)$, entonces existe $\lambda \in \sigma(A)$ tal que*

$$|\lambda - \mu| \leq k_p(R)\,\|A\|_p \quad (1 \leq p \leq \infty).$$

En particular, si A es simétrica se tiene

$$|\lambda - \mu| \leq \|A\|_2 .$$

Demostración. Sea $\mu \in \sigma(A + \Delta A)$, que podemos suponer no está en $\sigma(A)$. Entonces

$$1 \leq k_p(R) \|A\|_p \max_{1 \leq i \leq n} |\lambda - D(i,i)|^{-1} ,$$

de donde se deduce que existe i_0, $1 \leq i_0 \leq n$, tal que

$$|\lambda - D(i_0, i_0)| \leq k_p(R) \|A\|_p ,$$

y sabemos que $D(i_0, i_0) \in \sigma(A)$. □

8.1. El método de la potencia y de la potencia inversa

Sea $A \in \mathbb{R}^{n \times n}$. Obsérvese que si $\lambda \in \sigma(A)$ y $x \in S_\lambda(A)$ es no nulo, entonces

$$\lambda = \frac{\overline{x}' A x}{\overline{x}' x} \quad \text{(cociente de Rayleigh)},$$

por lo que es suficiente con preocuparse por determinar vectores propios.

Supongamos que tenemos $\sigma(A) = \{\lambda_i\}_{i=1}^n$, donde $\lambda_1 = \lambda_2 == \lambda_r$ es un valor propio dominate en el sentido de que se verifica

$$|\lambda_1| = \ldots = |\lambda_r| > |\lambda_{r+1}| \geq |\lambda_{r+2}| \geq \ldots \geq |\lambda_n| ,$$

y que existe una base $\{v_i\}_{i=1}^n$ de $\mathbb{C}^n$ de vectores propios con $Av_i = \lambda_i v_i$, $1 \leq i \leq n$. Esta última condición será eliminada en la siguiente sección.

Tomamos $y_0 \in \mathbb{R}^n$ de partida. Entonces

$$y_0 = \sum_{i=1}^n \alpha_i v_i,$$

por lo que

$$A^k y_0 = \lambda_1^k \left(\sum_{i=1}^r \alpha_i v_i + \sum_{i=r+1}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k v_i \right) \quad (k \in \mathbb{N}).$$

Es claro que si hacemos tender k a infinito, el término entre paréntesis tiende a $\sum_{i=1}^r \alpha_i v_i \in S_{\lambda_1}(A)$, que es no nulo si el vector inicial tiene una componente en la dirección del subespacio propio asociado al valor propio dominante. Para evitar el término conflictivo λ_1^k normalizamos los vectores. En efecto, la sucesión

$$z_k := \frac{A^k y_0}{\|A^k y_0\|_2} \quad (k \in \mathbb{N}),$$

se aproxima a un vector propio asociado a λ_1 tanto como queramos, y

$$\lim_k z_k' A z_k = \lambda_1.$$

La técnica expuesta se denomina *método de la potencia*.

Evidentemente tendremos problemas con esta técnica si tenemos un valor propio dominante complejo o valores propios reales distintos dominantes, pero existen técnicas adaptadas para solucionar el problema. El método funcionará tanto mejor cuanto mayor sea el cociente $|\lambda_{r+1}/\lambda_1|$.

Si suponemos que A es regular, los inversos de los valores propios de A son valores propios de A^{-1} y viceversa (los vectores propios se conservan), de ahí que esta técnica sea válida también para aproximar los valores propios minorantes, sin más que aplicar la técnica a la matriz inversa A^{-1} (método de la *potencia inversa*).

El hecho de que los valores propios de la matriz trasladada sean los valores propios trasladados, permite localizar cualquier valor propio de tamaño intermedio, siempre que éste este separado de los restantes en módulo. Aplicaríamos el método de la potencia inversa a la matriz trasladada, por supuesto que con ciertos cambios para deshacer la traslación realizada.

Destaquemos por último que cada valor propio obtenido puede ser eliminado por la técnica de *deflación*. En efecto, si conocemos $\lambda \in \sigma(A)$ y $x \in S_\lambda(A)$, unitario para la norma euclidea, entonces determinamos una matriz de Householder P tal que $Px = \pm e_1$ y entonces

$$P^{-1}APx = \lambda e_1,$$

de donde se deduce que la matriz $P^{-1}AP$, semejante a A, tiene la forma

$$\begin{bmatrix} \lambda & w \\ 0 & B \end{bmatrix},$$

donde B es una matriz cuadrada de tamaño $n-1$. Continuaremos la busqueda de los valores propios con la matriz B.

8.2. El método de la potencia para matrices no diagonalizables

Sea $A \in R^{n\times n}$. Es conocido que

$$C^n = \bigoplus_{i=1}^{k} \ker(A - \lambda_i)^{k_i}$$

donde los valores propios de A, distintos dos a dos, son λ_i, $1 \le i \le k$, con multiplicidades algebraicas k_i, respectivamente. En particular $\sum_{i=1}^{k} k_i = n$. Determinamos una base $\{v_{ij}\}_{j=1}^{k_i}$ de $\ker(A-\lambda_i)^{k_i}$ para cada $1 \le i \le k$. Así, $\cup_{i=1}^{k}\{v_{ij}\}_{j=1}^{k_i}$ es una base de C^n.

Supondremos que hay un valor propio dominante, es decir,

$$|\lambda_1| > |\lambda_2| \ge \ldots \ge |\lambda_{k-1}| \ge |\lambda_k|$$

Teorema 8.3. *Sea $y_0 \in \mathbb{R}^n$ tal que tiene una componente en la dirección del subespacio* $\ker(A-\lambda_1)^{k_1}$ *($y_0 \notin (\ker(A-\lambda_1)^{k_1})^{\perp}$). Consideramos la sucesión*

$$y_s \equiv A^s y_0 / \left\|A^s y_0\right\|_2, \quad s = 1, 2, \ldots$$

Se puede elegir un término de dicha sucesión tan cercano a un vector propio asociado a λ_1 como se desee.

Demostración. En primer lugar expresamos y_0 en términos de la base introducida

$$y_0 = \sum_{i=1}^{k} \sum_{j=1}^{k_i} \alpha_{ij} v_{ij} \quad (\alpha_{ij} \in \mathbb{C})$$

donde, por hipótesis, alguna de las coordenadas α_{1j}, $1 \leq j \leq k_1$, es no nula. En otras palabras, el vector $w_0 := \sum_{j=1}^{k_1} \alpha_{1j} v_{1j}$ es no nulo.

Para s suficientemente grande tenemos

$$\begin{aligned} A^s v_{ij} &= (A - \lambda_i + \lambda_i)^s v_{ij} = \sum_{r=0}^{s} \binom{s}{r} \lambda_i^{s-r} (A-\lambda_i)^r v_{ij} \\ &= \sum_{r=0}^{k_i-1} \binom{s}{r} \lambda_i^{s-r} (A-\lambda_i)^r v_{ij}, \end{aligned}$$

luego

$$A^s y_0 = \sum_{i=1}^{k} \sum_{j=1}^{k_i} \alpha_{ij} \sum_{r=0}^{k_i-1} \binom{s}{r} \lambda_i^{s-r} (A-\lambda_i)^r v_{ij}.$$

Analicemos la parte correspodiente al subespacio $\ker(A-\lambda_1)^{k_1}$. Tenemos

$$\sum_{j=1}^{k_1} \sum_{r=0}^{k_1-1} \binom{s}{r} \lambda_1^{s-r} (A-\lambda_1)^r v_{1j} = \sum_{r=0}^{k_1-1} \binom{s}{r} \lambda_1^{s-r} (A-\lambda_1)^r w_0.$$

Usémos la notación

$$w_r := (A-\lambda_1)^r w_0 \ \ (1 \leq r \leq k_1 - 1).$$

Sea $r_1 = \text{máx}\{0 \leq r \leq k_1 - 1 : w_r \text{ es no nulo}\}$. Por la elección de r_1 se tiene

$$0 \neq w_{r_1} \in \ker(A - \lambda_1).$$

Entonces

$$\begin{aligned} A^s y_0 &= \binom{s}{r_1} \lambda_1^{s-r_1} \left[w_{r_1} + \sum_{r=0}^{r_1-1} \frac{\binom{s}{r}}{\binom{s}{r_1}} \lambda_1^{r_1-r} w_r \right. \\ &\quad \left. + \sum_{i=2}^{k} \sum_{j=1}^{k_i} \alpha_{ij} \sum_{r=0}^{k_i-1} \frac{\binom{s}{r} \lambda_i^{s-r}}{\binom{s}{r_1} \lambda_1^{s-r_1}} (A-\lambda_i)^r v_{ij} \right], \end{aligned}$$

siendo ya evidente que la parte entre corchetes tiende al vector propio w_{r_1} cuando $s \to \infty$, ya que es sabido, aplíquese por ejemplo el criterio del cociente a la correspondiente serie numérica, que

$$\lim_{m\to\infty} m^l \beta^m = 0 \ (|\beta| < 1,\, l \in \mathbb{N}).$$

En consecuencia, la sucesión y_s se aproxima a un vector propio asociado a λ_1, o a su opuesto, tanto como queramos. □

Nota 8.1. *La presentación de la prueba podría hacerse más sencilla partiendo de que* $y_0 = \sum_{i=1}^{k} v_j$ *con* $v_j \in \ker(A-\lambda_i)^{k_i}$, $1 \le i \le k$.

Nota 8.2. *La base de* $\ker(A-\lambda)^{k(\lambda)}$ *puede elegirse completamente fuera del núcleo: Sea* $A = [1\ 1; 0\ 1]$ *cuyo único valor propio es* $\lambda = 1$ *y el subespacio propio está generado por* $[1; 0]$. *Una base de* $\ker(A-1)^2 = \mathbb{R}^2$ *es* $\{[0; 1],[1; 1]\}$ *y ninguno es vector propio.*

Nota 8.3. *¿El método de la potencia en las condiciones del teorema anterior aplicado a vectores propios iniciales independientes proporciona vectores propios independientes? Obviamente no es cierto en general. Volvamos al ejemplo de la nota previa. Si* $y_0 = [1; 0]$ *entonces*

$$A^s y_0 = 1^s y_0 + s(A-1)y_0 = y_0 + s[0; 1] = [1; s]$$

de donde

$$y_s = [1; s]/\sqrt{1+s^2} \to [1; 0].$$

Lo mismo sucede si partimos de $y_0 = [1; 1]$ *y son independientes. La pregunta natural es como conseguir vectores propios independientes adicionales (si es que hay más de uno; puede parecer que el ejemplo propuesto depende de esta circunstancia pero no es así). ¿Partir de vectores iniciales ortogonales entre sí, por ejemplo?*

Nota 8.4. *En esta nota analizaremos situaciones bajo las que el método de la potencia no funciona.*

- *Valores propios dominantes reales distintos de igual módulo*

  Esta situación se detecta en la práctica al observarse que se aproximan dos vectores diferentes según el orden del término de la sucesión que consideremos correspondiendo a la suma y la diferencia de vectores propios asociados a cada valor propio. Así que únicamente habría que cambiar el algoritmo y dar como vector propio asociado la suma de las dos últimos términos de la suceción. Luego el valor propio se aproxima mediante el correspondiente cociente de Rayleigh.

  Otra opción es tener en cuenta que

  $$\sigma(A^2) = \{\lambda^2 : \lambda \in \sigma(A)\}$$

  y aplicar el método de la potencia a A^2. *Con la aproximación del valor propio, tras extraer raíz cuadrada, pasaríamos a aplicar el método de la potencia inversa con desplazamiento.*

- *Valores propios complejos conjugados*

8.3. El método QR

El método QR persigue alcanzar el resultado del teorema de Schur: toda matriz es semejante por transformaciones unitarias a una matriz triangular superior en cuya diagonal principal estarán los valores propios de la matriz de partida ($T \equiv \overline{Q}'AQ$ triangular superior con $\overline{Q}'Q = I_n$). Sabemos que no es posible alcanzar, en general, dicha situación mediante un método directo, por lo que evidentemente estaremos ante un método iterativo en el que por sucesivas tranformaciones de semejanza mediante matrices unitarias (ortogonales pues nos centramos en artimética real) desearíamos alcanzar la situación descrita.

Consiste en realizar el siguiente esquema

Algoritmo 8.1. *Datos:A,nits*
for $i = 1 : nits$
$A = Q * R$ *(Obtener una factorización QR de la matriz, p.e. con matrices Householder)*
$A = R * Q$ *(Revertir el orden del producto)*
end
Salida:A

Más explícitamente, se trata de hacer lo siguiente:
$A \equiv A^{(0)} = Q_1R_1 \to A^{(1)} \equiv R_1Q_1 = Q_1'AQ_1$ (Semejante a A) $\to A^{(1)} = Q_2R_2 \to A^{(2)} \equiv R_2Q_2 = Q_2'A^{(1)}Q_2 = Q_2'Q_1'AQ_1Q_2 \to \ldots \to A^{(k)} \equiv Q_k' \ldots Q_2'Q_1'AQ_1Q_2 \ldots Q_k$
de forma que la sucesión $A^{(k)}$ converja a una matriz triangular (probablemente por bloques si pensamos en la presencia de posibles valores propios complejos conjugados y en valores propios múltiples).

Si llamamos $Q^{(k)} := Q_1Q_2 \ldots Q_k$ y $R^{(k)} := R_kR_{k-1} \ldots R_1$, $k \in \mathbb{N}$, entonces

$$A^{(k)} = Q^{(k)\prime}AQ^{(k)} \ \ (k \in \mathbb{N}).$$

Además,

$$A^k = Q^{(k)}R^{(k)} \ \ (k \in \mathbb{N}).$$

En efecto, $A^1 = Q^{(1)}R^{(1)}$ y por inducción

$$\begin{aligned} A^k &= AA^{k-1} = Q^{(1)}R^{(1)}Q^{(k-1)}R^{(k-1)} = Q_1(R_1Q_1)Q_2 \ldots Q_{k-1}R^{(k-1)} \\ &= Q_1Q_2(R_2Q_2) \ldots Q_{k-1}R^{(k-1)} = Q_1Q_2(Q_3R_3) \ldots Q_{k-1}R^{(k-1)} \\ &= \ldots = Q^{(k)}R^{(k)}. \end{aligned}$$

Intuitivamente esperamos que la sucesión $Q^{(k)}$ converja a una matriz ortogonal Q cuyas columnas son vectores propios de A ya que $A^k(R^{(k)})^{-1} = Q^{(k)}$ ($R^{(k)}$ regular por serlo A) y al hacer k tender a infinito es como si aplicáramos el método de la potencia a las columnas de $(R^{(k)})^{-1}$ (no hace falta normalizar como en el método de la potencia pues $Q^{(k)}$ ortogonal; obsérvese que no puede converger, para columnas diferentes, a vectores que no sean ortogonales y unitarios, lo que da idea de porque no convergen varias columnas a un vector propio 'dominante').

Nota 8.5. *Es simple probar que el límite de matrices ortogonales (triangular) es ortogonal (triangular).*

Nota 8.6. *Si tenemos una sucesión de números reales acotada,* $\{x_n\}_{n\in\mathbb{N}}$*, de forma que cualquier subsucesión convergente lo hace a un mismo número* l*, entonces existe* $\lim x_n = l$*. (por reducción al absurdo tendríamos una subsucesión convergente, por ser acotada la de partida, que no puede converger a* l*)*

Teorema 8.4. *Sea* $A \in \mathbb{R}^{n\times n}$*. Supongamos que* A *tiene valores propios* $\{\lambda_i\}_{i=1}^n$ *cumpliendo*

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_{n-1}| > |\lambda_n| > 0.$$

En particular, existe $P \in \mathbb{R}^{n\times n}$ *regular de forma que*

$$D := PAP^{-1} = diag([\lambda_1, \lambda_2, \ldots, \lambda_n]).$$

Supondremos además que P^{-1} *es factorizable* LU*. Entonces la sucesión matricial* $\{A^{(k)}\}_{k\in\mathbb{N}}$ *generada mediante el método QR (con unicidad de la factorización) cumple:*

$$\lim_{k\to\infty} A^{(k)}(i,j) = \begin{cases} \lambda_i,\ 1 \le i \le n,\ i = j, \\ 0,\ 1 \le j < i \le n. \end{cases}$$

Demostración. En primer lugar nótese que asumimos que A es regular. Además, tiene n valores propios distintos dos a dos, por lo que la matriz es diagonalizable.

En la nota previa se dió una condición suficiente para la unicidad de la factorización QR de una matriz regular que va a ser clave en la prueba, al permitir analizar el comportamiento de la sucesión $\{Q^{(k)}\}$.

De entrada nótese que $A^k = Q^{(k)}R^{(k)}$ es la única factorización de ese tipo de la matriz A^k, y de las hipótesis vamos a encontrar otra. Sea $P = QR$ una factorización QR de P y $P^{-1} = LU$ la factorización cuya existencia se asume en el enunciado (hay una diferencia de un cambio de filas previo). Tenemos

$$A^k = PD^kP^{-1} = QRD^kLU = QRD^kLD^{-k}D^kU.$$

Es fácil comprobar que

$$(D^kLD^{-k})(i,j) = \left(\frac{\lambda_i}{\lambda_j}\right)^k L(i,j)\ \ (1 \le i,j \le n),$$

y como L es triangular inferior con unos en la diagonal inferior se deduce, por la relación existente entre los valores propios, que

$$\lim_{k\to\infty} D^kLD^{-k} = I_n.$$

Nótese que la convergencia anterior depende de

$$\max_{1\leq i<j\leq n}\left|\frac{\lambda_i}{\lambda_j}\right|.$$

Usamos la notación $E_k := D^k L D^{-k} - I_n$. Tenemos

$$RD^k L D^{-k} = (I + RE_k R^{-1})R.$$

Como $\lim_{k\to\infty} E_k = 0$, entonces $I + RE_kR^{-1}$ es invertible para k suficientemente grande. Admite pues una factorización única QR que denotamos

$$I + RE_kR^{-1} = \widetilde{Q}_k\widetilde{R}_k.$$

La sucesión $\{\widetilde{Q}_k\}$ está acotada por ser de matrices ortogonales, luego existe una subsucesión convergente a una matriz $\widetilde{Q}$ que debe ser ortogonal. Por lo tanto, también la correspondiente subsucesión de $\{\widetilde{R}_k\}$ converge a una matriz $\widetilde{R}$ triangular superior con $\widetilde{R}(i,i) \geq 0$, $1 \leq i \leq n$. Tomando límites en la subsucesión se obtiene $I_n = \widetilde{Q}\widetilde{R}$, lo que asegura que $\widetilde{R}(i,i) > 0$, $1 \leq i \leq n$, y la unicidad de la factorización QR de I_n nos proporciona finalmente que $\widetilde{Q} = \widetilde{R} = I_n$. Lo mismo se concluye para cualesquiera subsucesiones convergentes de $\{\widetilde{Q}_k\}$ y $\{\widetilde{R}_k\}$, luego por una nota previa se tiene que ambas sucesiones convergen a I_n.

En definitiva, tenemos

$$Q^{(k)}R^{(k)} = A^k = Q\widetilde{Q}_k\widetilde{R}_k RD^k U.$$

La matriz $\widetilde{R}_k RD^k U$ es triangular superior con entradas no nulas en la diagonal principal. Podemos determinar una matriz diagonal D_k en cuya diagonal principal elegimos ± 1 convenientemente de forma que $D_k\widetilde{R}_k RD^k U$ ya tiene entradas positivas en diagonal principal. Por la unicidad de la QR

$$Q'Q^{(k)} = \widetilde{Q}_k D_k.$$

Con esta observación tenemos

$$A^{(k)} = Q^{(k)\prime}AQ^{(k)} = D_k\widetilde{Q}'_k RDR^{-1}\widetilde{Q}_k D_k.$$

Sea $S_k := \widetilde{Q}'_k RDR^{-1}\widetilde{Q}_k$. Obsérvese que $\lim_{k\to\infty} S_k = RDR^{-1}$, que es una matriz triangular superior cuya diagonal principal coincide con la de D. En definitiva,

$$A^{(k)}(i,j) = D_k(i,i)S_k(i,j)D_k(j,j) \;\; (1 \leq i,j \leq n),$$

lo que proporciona el resultado deseado. □

Nota 8.7. *El resto de componentes de $A^{(k)}$ puede no converger, pero queda claro tras la prueba que*

$$\lim_{k\to\infty}\left|A^{(k)}(i,j)\right| = \left|RDR^{-1}(i,j)\right| \;\; (1 \leq i < j \leq n).$$

Nota 8.8. *La condición sobre P^{-1} no es esencial, pero complica la prueba si no se asume. A este respecto, nótese que si A es diagonal, entonces el método QR va a proporcionar en cada iteración la propia A en cada iteración, sin reordenar en forma decreciente en módulo su diagonal principal. Eso esperamos también si no asumimos la condición mencionada; es decir,*

$$\lim_{k\to\infty} S_k = RDR^{-1}$$

donde D es diagonal con los valores propios de A reorganizados en un orden, en principio, distinto al exigido.

Los cambios sin la citada condición son los siguientes: existe una matriz permutación, P_σ, tal que $P_\sigma P^{-1} = LU$. Tenemos

$$\begin{aligned} A^k &= PD^kP^{-1} = QRP_\sigma D^k P_\sigma' LU \\ &= QR\widetilde{D}^k L\widetilde{D}^{-k}\widetilde{D}^k U \end{aligned}$$

donde ahora $PP_\sigma' = QR$ (factorizamos QR esta nueva matriz y no P) y $\widetilde{D} := P_\sigma D P_\sigma'$. Entonces,

$$(\widetilde{D}^k L\widetilde{D}^{-k})(i,j) = \left(\frac{\lambda_{\sigma(i)}}{\lambda_{\sigma(j)}}\right)^k L(i,j) \;\; (1 \le i,j \le n).$$

Un análisis más detallado de la factorización $P_\sigma P^{-1} = LU$ nos segura que

$$L(i,j) = 0 \text{ si } \sigma(i) > \sigma(j),$$

lo que soluciona el problema de que en dicho caso $\lambda_{\sigma(i)}/\lambda_{\sigma(j)} > 1$. Así, $\widetilde{D}^k L\widetilde{D}^{-k} := I_n + E_k$ con $E_k \to 0$. A partir de este punto razonamos de igual modo que en el teorema para llegar por unicidad de la QR a que

$$Q^{(k)} = Q\widetilde{Q}_k D_k \text{ con } \widetilde{Q}_k.$$

Por consiguiente, como

$$A = PP_\sigma'\widetilde{D}P_\sigma P^{-1} = QR\widetilde{D}Q'R^{-1},$$

tenemos

$$A^{(k)} = Q^{(k)\prime}AQ^{(k)} = D_k\widetilde{Q}_k' R\widetilde{D}R^{-1}\widetilde{Q}_k D_k$$

y

$$\widetilde{Q}_k' R\widetilde{D}R^{-1}\widetilde{Q}_k \to R\widetilde{D}R^{-1}.$$

Nota 8.9. *Nótese que el método presentará anomalías en el caso de tener valores propios complejos conjugados, como sucede con el método de la potencia. Si se lo aplicamos a $[1,1,0;1,1,0;0,0,-7]$, obtenemos la misma matriz siempre. En general, el método convergerá a una matriz 'triangular superior por bloques', es decir, en la 'diagonal principal' tendremos matrices cuadradas de tamaño el número de valores propios con igual módulo.*

Por supuesto, si la matriz de partida es triangular superior, obtendremos una sucesión constante de matrices con todos los términos igual a ella.

Capítulo 9

Cuestiones

1. Calcúlese $\|xy'\|_p$, $p = 1, 2, \infty$, donde $x \in \mathbb{R}^m$ e $y \in \mathbb{R}^n$.

2. Sea $A \in \mathbb{R}^{m\times n}$. Pruébense las siguientes afirmaciones:

   a) $\|A\|_F = \sqrt{\text{traza}(A'A)} = \sqrt{\text{traza}(AA')}$.
   b) $\|PAQ\|_F = \|A\|_F$, para toda $P \in \mathbb{R}^{m\times m}$ y $Q \in \mathbb{R}^{n\times n}$ ortogonales.

3. Demuéstrense las siguientes desigualdades para $A \in \mathbb{R}^{m\times n}$ y proporciónese una matriz no nula donde se alcanze cada desigualdad:

   a) $\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\,\|A\|_2$.
   b) $\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty$.

4. Sea $A \in \mathbb{R}^{n\times n}$ diagonalizable tal que sus valores propios son positivos. Pruébese que existe una matriz B tal que $A = B^2$ (B es una raíz cuadrada de A). ¿Toda matriz cuadrada posee una raíz cuadrada?

5. Sea $A \in \mathbb{R}^{n\times n}$. Demuéstrese que

$$\sigma(A^2) = \{\lambda^2 : \lambda \in \sigma(A)\}.$$

6. Sea $A \in R^{m\times n}$ y sea $A^+ \in R^{n\times m}$ la inversa generalizada de A. Entonces

$$\min_{X\in R^{n\times m}} \|AX - I\|_F = \left\|AA^+ - I\right\|_F = m - rank(A).$$

7. Sea $A \in R^{m\times n}$ y sea $A^+ \in R^{n\times m}$ la inversa generalizada de A. Entonces

$$\left\|AA^+ - I\right\|_2 = \begin{cases} 0, & rank(A) = m \\ 1, & rank(A) < m \end{cases}.$$

8. La factorización QR de una matriz regular $A \in R^{n\times n}$, donde $Q \in R^{n\times n}$ es ortogonal, $R \in R^{n\times n}$ es triangular superior con $R(i,i) > 0$, $1 \leq i \leq n$, es única.

9. Sea $A \in R^{n\times n}$ ortogonal, triangular superior y con $A(i,i) > 0$, $1 \leq i \leq n$. Entonces $A = I$.

10. Sea $A \in R^{n\times n}$ regular factorizable LU. Sea $A = LU$ una tal factorización y sea $A = QR$ una factorización QR de A. Entonces

    $$L^{-1}(i,:)Q(:,j) = 0 \text{ (fila ortogonal a ciertas columnas)}, \quad 1 \leq j \leq i-1,\ 1 \leq i \leq n.$$

11. Sea $A \in R^{n\times n}$ simétrica de forma que existe $B \in R^{n\times n}$ con $A = B^2$ (B es una raíz cuadrada de A). Entonces A es definida positiva.

12. Sea $A \in R^{m\times n}$. Consideramos la matriz

    $$\widetilde{A} = \begin{bmatrix} 0 & A' \\ A & 0 \end{bmatrix} \in R^{(m+n)\times(m+n)}$$

    Si μ es un valor singular de A, entonces $\pm\mu$ son valores propios de $\widetilde{A}$.

13. Sea (B, c) un método iterativo convergente para sistemas asociados a una matriz regular $A \in R^{n\times n}$. Supongamos además que $B \in R^{n\times n}$ regular. Entonces el método iterativo $(B^2, (B+I)c)$ es convergente y converge más rápidamente que el original.

14. Sea $A \in R^{n\times n}$. Existe $\lambda \in R$ tal que para todo $x \in R^n$ no nulo los vectores $(A-\lambda)^k x$ son no nulos para $k = 0, 1, 2, \ldots$

15. Sea $A \in R^{n\times n}$ con $A(i,j) = i + j - 1$, $1 \leq i, j \leq n$. Entonces

    $$|\lambda| \leq \frac{3}{2}\,(n^2 + n) \quad (\lambda \in \sigma(A)).$$

16. A es factorizable LU sii A^2 (A^n $(n \in N)$) es factorizable LU.

17. Si A es simétrica definida positiva, entonces A^n es factorizable LU.

18. Si $B \in R^{n\times n}$ es ortogonal, entonces

    $$\mu_i(AB) = \mu_i(A) \quad (A \in R^{m\times n}, 1 \leq i \leq \text{mín}(m,n)).$$

19. Considerar el sistema $Ax = b$ con A regular. El método iterativo $((A^2+1), Ab)$ para el sistema anterior es convergente.

20. Toda solución del problema LSP asociado a $Ax = b$ tiene el mismo residuo.

21. $k(A^2) \leq k(A)^2$, y como $1 \leq k(A) \leq k(A)^2$, ¿es de esperar que A^2 esté peor condicionada que A?

22. Sea $\{A_r\}_{n\in N} \subset R^{n\times n}$ una sucesión de matrices regulares convergente a la matriz A, de modo que existe y es finito $\lim_{r\to\infty} k(A_r)$. Entonces A es regular.

23. Si $\lim A_r = A$, entonces $\lim k(A_r) = k(A)$.

24. Sea $A \in R^{n\times n}$ verificando

$$A(i,i) = 1 \text{ y } \sum_{j=1, j\neq i}^{n} |A(i,j)| < 1 \quad (1 \leq i \leq n).$$

El método de Jacobi aplicado a A es convergente.

25. Para $\lambda \in R$ definimos

$$A(\lambda) \equiv \begin{bmatrix} 1 & -\lambda \\ -\lambda & 1 \end{bmatrix}.$$

Los métodos de Jacobi y relajación, $0 < w < 2$, aplicados a $A(\lambda)$ son convergentes para $-1 < \lambda < 1$.

26. Sea $A \in R^{n\times n}$ simétrica y sean $\lambda \in C$ y $x \in C^n$ con $\|x\|_2 = 1$. Sea $r \equiv \|Ax - \lambda x\|_2$. Entonces existe $\mu \in \sigma(A)$ tal que

$$|\lambda - \mu| \leq r.$$

27. Sea $0 < |\varepsilon| < 1$. Sea

$$A(\varepsilon) = \begin{bmatrix} 1 & -1 & 1 \\ -1 & \varepsilon & \varepsilon \\ 1 & \varepsilon & \varepsilon \end{bmatrix}$$

Entonces, $k_\infty(A) \geq 1.5\,|\varepsilon|^{-1}$.

28. Sea

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

Entonces la factorización LU existe y $L = U'$.

29. Sea A regular y B tal que $\delta \equiv \left\|A^{-1}\right\| \|B - A\| < 1$. Entonces$B$ es regular y

$$\left\|A^{-1} - B^{-1}\right\| \leq \frac{\delta}{1-\delta} \left\|A^{-1}\right\|.$$

30. Sea x solución del LSP asociado al sistema $Ax = b$. Sea $r \equiv Ax - b$. Entonces $[r; x]$ es solución del sistema de ecuaciones

$$\begin{bmatrix} I & A \\ A' & 0 \end{bmatrix} y = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

Si A es inyectiva, el sistema tiene solución única.

31. Sea A regular. Entonces es factorizable en la forma QR de forma única si le pedimos que $R(i,i) = 1$ $(1 \leq i \leq n)$.

32. Sea A real. Sea x un vector real unitario para la norma euclídea. Sea $\rho := x'Ax$ y sea $z := (A - \rho)x$. Entonces el conjunto

$$[\rho - \|z\|_2 , \rho + \|z\|_2] \cap \sigma(A)$$

es no vacío.

33. Sea (B, c) un método iterativo convergente con B no nula. Sea $\alpha \in \mathbb{R}^n$ el único vector cumpliendo $B\alpha + c = \alpha$. Existe $x_0 \in \mathbb{R}^n$, $x_0 \neq \alpha$, tal que $x_k = \alpha$ para algún $k \geq 1$, donde $\{x_k\}_{k\geq 0}$ representa la sucesión generada mediante el método iterativo a partir del dato inicial x_0.

34. Sea $A \in \mathbb{R}^{m\times n}$ bidiagonal superior de rango máximo. Entonces $B = A'A$ es tridiagonal y definida positiva.

35. Sea $A \in \mathbb{R}^{m\times n}$ con ancho de banda $\{0, 1\}$; es decir, bidiagonal superior, y $rank(A) = n \leq m$. Entonces $B = A'A$ es tridiagonal y definida positiva.

36. $x = A^+b$ es la solución de LSP asociado al sistema $Ax = b$ de menor norma.

37. Sea $A \in \mathbb{R}^{n\times n}$ simétrica. Sea $\varepsilon > 0$ y sean $1 \leq i, j \leq n, i \neq j$. Definimos la matriz $A_{i,j,\varepsilon}$ mediante

$$A_{i,j,\varepsilon}(r,s) = \begin{cases} A_{i,j,\varepsilon}(r,s) + \varepsilon, & r, s \in \{i, j\},\ r \neq s, \\ A(r,s), & \text{en otro caso} \end{cases}$$

Sea $\mu \in \sigma(A)$. Existe $\lambda \in \sigma(A_{i,j,\varepsilon})$ tal que

$$|\lambda - \mu| \leq \sqrt{2}\varepsilon.$$

38. Sea $A \in R^{n\times n}$ simétrica con $A^2 = A$ y no idénticamente nula. Entonces $\|A\|_2 = 1$.

39. $A \in R^{n\times n}$ es simétrica sii $r(A) = \|A\|_2$.

40. Sea $A \in R^{m\times n}$ con $rank(A) = 1$. Existen $x \in R^m$ e $y \in R^n$ tales que

$$A = xy'.$$

41. Sea $A \in R^{n\times n}$ regular. Podemos modificar un elemento cualquiera de la matriz A, sumándole un número, de forma que la matriz obtenida sigue siendo regular.

42. Sea $A \in R^{n\times n}$ regular y sean $x, y \in R^n$. Si $x'A^{-1}y + 1 = 0$, entonces $A + yx'$ es singular.

43. Sea $A \in \mathbb{R}^{n\times n}$ tal que $AA' = A'A$. Entonces A admite una factorización SVD con $U = V$; es decir, de la forma

$$A = U\Sigma U',$$

con $U \in \mathbb{R}^{n\times n}$ ortogonal y $\Sigma \in \mathbb{R}^{n\times n}$ diagonal con entradas en la diagonal principal no negativas y ordenadas en forma decreciente. En particular, A es simétrica.

Capítulo 10

Ejercicios prácticos con MATLAB

1. Prográmese en MATLAB la regla de Cramer para determinar la solución de un sistema de ecuaciones lineales $Ax = b$ usando la función *det* proporcionada por MATLAB para el cálculo de determinantes.

   Sean $A = \text{rand}(100)$ y $b = A(:, 1)$. Resuélvase mediante la función generada el sistema de ecuaciones lineales $Ax = b$. Compárese con la opción $A\backslash b$ proporcionada por MATLAB usando las funciones *tic* y *toc*.

2. Proporciónese una función con la estructura

   $$[ds, vsx] = splinc(vu, du, a, b, m, x)$$

   para hallar el splin cúbico asociado a una función u y a una partición del intervalo $[a, b]$ en m subintervalos. El argumento vu representará un vector con los valores de la función u en los nodos de la partición, du un vector con los valores de u' en los extremos del intervalo, y x la abcisa en la que se quiere aproximar u. La función proporcionará un vector con las derivadas del splin en los nodos de la partición (ds) y el valor del splin en x (vsx).

   Considérese la función $u(x) = 1 - \cos(\pi x)$. Sabiendo que $u'(0) = u'(1) = 0$ y utilizando los valores de u en los nodos de la partición uniforme del intervalo $[0, 1]$ asociada a $m = 200$, aproxímese el valor de u en $x = 0.998$ mediante el correspondiente splin cúbico.

3. Constrúyase una función con la estructura

   $$vu = dfinitas(f, g, a, b, alfa, beta, m)$$

   para el método en diferencias finitas propuesto para aproximar la solución u del problema de contorno

   $$\begin{cases} -y'' + f(x)y = g(x), \\ y(a) = \alpha,\ y(b) = \beta, \end{cases}$$

siendo vu un vector con las aproximaciones de los valores de u en los nodos de la partición uniforme del intervalo $[a, b]$ determinada por m.

Utilícese dicha función para aproximar la solución del problema de contorno

$$-y'' + (x+1)y = (x^2-2)e^{x-2};\ y(0) = 0,\ y(2) = 2,$$

en los nodos de la partición uniforme del intervalo $[0, 2]$ correspondiente a $m = 100$. Compárense los resultados con la solución exacta $u(x) = xe^{x-2}$, $x \in [0, 2]$, creando una tabla o matriz con los valores exactos, los aproximados y el error relativo en cada nodo de la partición.

4. Proporciónense diferentes matrices que tengan como valores propios los primeros 25 números naturales (utilícense las funciones *poly*, *diag*, *compan*, etc.). Obténganse mediante la función *eig* sus valores propios y vectores propios asociados. Hállense los polinomios característicos de las matrices construidas y luego sus raíces (*roots*).

5. Dado un sistema ortonormal de n vectores de $\mathbb{R}^m$ $(1 \leq n < m)$ sabemos que podemos encontrar $m - n$ vectores de $\mathbb{R}^m$ que junto con los primeros formen una base ortonormal de $\mathbb{R}^m$.

   a) Sea $A \in \mathbb{R}^{m\times n}$ de forma que sus vectores columna formen un sistema ortonormal. Constrúyase una función con la estructura

   $$B = basgramsm(A)$$

   que obtenga una base ortonormal de $\mathbb{R}^m$ de la que formen parte los vectores $\{A(:, j)\}_{j=1}^n$. Los vectors de la base obtenida seran los vectores columna de la matriz B.

   b) Aplíquese la función construida en el apartado anterior para obtener una base de $\mathbb{R}^4$ de la que formen parte los vectores columna de la matriz

   $$A = \begin{pmatrix} 0 & 3/5 \\ -1 & 0 \\ 0 & 4/5 \\ 0 & 0 \end{pmatrix}$$

6. Proporciónese una función que halle la norma matricial subordinada $\|\|\|_1$ de una matriz, y un vector unitario donde se alcance dicha norma. Hágase lo propio con $\|\|\|_\infty$.

7. Obténgase una función que dibuje el conjunto

   $$\{Ax : \|x\|_2 = 1\}$$

   a partir de una matriz $A \in \mathbb{R}^{2\times 2}$. ¿Qué determina $\|A\|_2$ en la figura obtenida?

8. Constrúyase una función que genere de forma aleatoria una matriz permutación de tamaño arbitrario. (Utilícese la función *rand* para crear de forma aleatoria una permutación de $\{1, \ldots, n\}$. Aunque en MATLAB existe la función *randperm* que resuelve el problema, se pide una función alternativa).

9. Utilícese la función SVD proporcionada por MATLAB para generar una función con la siguiente estructura:

$$[B, err, rc] = ap(A, k)$$

donde A es la matriz a aproximar, k es el $rank$ requerido de la aproximación, B una aproximación con $rank(B) = k$ ($B := A^{(k)}$), err es el error relativo cometido en la norma espectral y rc el ratio de compresión logrado.

10. Una de las imágenes test más ulitizada en el tratamiento de imágenes es la imagen de Lena que podemos encontrar en

$$http://www.lenna.org$$

Cópiese en el directorio activo de MATLAB con el nombre *lena.jpg*. Para cargar el archivo gráfico anterior en MATLAB utilícese

$$A = imread('lena.jpg')$$

Compruébese que A es un array de tamaño $m \times n \times 3$ con datos de la clase $uint8$ (*help class*). Hállese su tamaño. Obténgase una imagen en tonos de grises haciendo

$$A = A(:, :, 1)$$

Véase la imagen obtenida haciendo

$$imwrite(A,' lenabn.jpg')$$

Ejecútese

$$A = double(A)$$

para poder usar el comando SVD. Mediante la función construida en el apartado previo, obténganse la aproximacíón correspondiente a $k = 10$ de A. Para convertir la matriz obtenida en una imagen ($lenabn10.jpg$) úsense los comandos $uint8$ e $imwrite$.

Repítase el proceso expuesto para $k = 20, 50, 100$.

11. Análicese el caso de imágenes a color siguiendo las ideas del ejercicio anterior.

12. Usando el comando \ de MATLAB comprúebese numéricamente la desigualdad

$$\frac{\|x-y\|_\infty}{\|x\|_\infty} \leq \frac{k_\infty(A)}{1-\|\Delta A\|_\infty \|A^{-1}\|_\infty}\left(\frac{\|\Delta A\|_\infty}{\|A\|_\infty}+\frac{\|\Delta b\|_\infty}{\|b\|_\infty}\right),$$

en los siguientes casos:

a) $A=\begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}, \Delta A=\begin{bmatrix} 0 & 0 & 0.1 & 0.2 \\ 0.08 & 0.04 & 0 & 0 \\ 0 & -0.02 & -0.01 & 0 \\ -0.01 & -0.01 & 0 & -0.02 \end{bmatrix},$

$b=\begin{bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{bmatrix}$ e $\Delta b=\begin{bmatrix} 0.1 \\ -0.1 \\ 0.1 \\ 7 \end{bmatrix}$.

b) $A, b, \Delta b$ como en el apartado previo e

$$\Delta A=\begin{bmatrix} 0 & 0 & 0.001 & 0.002 \\ 0 & 0.004 & 0 & 0 \\ 0 & -0.002 & -0.001 & 0 \\ -0.001 & -0.001 & 0 & -0.002 \end{bmatrix}.$$

13. Genérese una función que acote inferiormente el número de condición, $k_\infty(A)$, de una matriz A triangular superior, resolviendo por remonte un adecuado sistema de ecuaciones asociado a dicha matriz tal y como se ha explicado en las clases teóricas. Aplíquese a la matriz

$$A=\begin{bmatrix} 1 & 0 & \alpha & -\alpha \\ 0 & 1 & -\alpha & \alpha \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

con $\alpha \in \mathbb{R}$. Compárese la cota obtenida con el valor exacto.

Recordando que

$$\frac{\|A\|_\infty}{\|A-B\|_\infty} \leq k_\infty(A) \quad (B \in \mathbb{R}^{n\times n} \text{ singular}),$$

propóngase alguna alternativa para la situación general de acotar inferiormente el número de condición, $k_\infty(A)$, de una matriz A regular cualquiera, y utilícese la función obtenida con la matriz

$$A=\begin{bmatrix} 7 & 8 & 9 \\ 8 & 9 & 10 \\ 9 & 10 & 8 \end{bmatrix}.$$

14. La *matriz de Hilbert de orden* n es una matriz cuadrada cuya entrada (i, j) es $1/(i + j - 1)$, $1 \leq i,\ j \leq n$. Es un ejemplo famoso de matriz mal condicionada.

    Demuéstrese que es una matriz simétrica definida positiva.

    Utilícese la función generada en el ejercicio anterior para confirmar el mal condicionamiento mencionado.

15. Sea $A \in \mathbb{R}^{2\times 2}$ regular. Demuéstrese que

$$k_2(A) = \alpha + \sqrt{\alpha^2 - 1},$$

    donde

$$\alpha = \frac{\|A\|_F^2}{2\,|\det(A)|}.$$

16. Sea $A \in \mathbb{R}^{n\times n}$. Pruébense las siguientes afirmaciones:

    a) Si A es estrictamente diagonalmente dominante por filas, entonces

$$k_\infty(A) \leq \frac{\|A\|_\infty}{\min\{|A(i,i)| - \sum_{1\leq j\leq n, j\neq i} |A(i,j)| : 1 \leq i \leq n\}}.$$

    b) Si $D = diag(d) \in \mathbb{R}^{n\times n}$ verifica que AD es estrictamente diagonalmente dominante por filas, entonces

$$k_\infty(A) \leq \frac{\|D\|_\infty \|A\|_\infty}{\min\{|A(i,i)d(i)| - \sum_{1\leq j\leq n, j\neq i} |A(i,j)d(j)| : 1 \leq i \leq n\}}.$$

    Escríbase un programa que decida si una matriz es o no estrictamente diagonalmente dominante por filas, y proporcione una cota superior de $k_\infty(A)$ cuando lo sea.

17. Prográmense los métodos de descenso y de remonte para matrices triangulares inferiores y triangulares superiores respectivamente, considerando la posibilidad de que las matrices puedan tener además estructura de matriz banda.

18. Constrúyase una función que resuelva sistemas de ecuaciones lineales por el método de eliminación de Gauss. Contémplese la posibilidad de que estemos interesados en resolver varios sistemas con la misma matriz de coeficientes, y en obtener el determinante de dicha matriz.

19. Adáptese el ejercicio anterior a matrices banda para las que sepamos que no es necesaria la elección del pivote. Utilícese esta función para obtener la inversa de la matriz de coeficientes del ejercicio 3 de la práctica 1 (splines cúbicos).

20. Modifíquese convenientemente la función definida en el apartado 2 para obtener una función que al introducir una matriz A nos indique si ésta es factorizable LU o no, y en caso afirmativo, nos devuelva una factorizaciónLU de A.

21. Proporciónese una función que devuelva la factorización $PA = LU$ de una matriz regular A utilizando pivote parcial, y otra función para obtener la factorización $PAQ = LU$ usando pivote total.

22. Prográmese el proceso de mejora iterativa para intentar mejorar una solución de un sistema $Ax = b$. La función a construir debe tener la estructura

$$miter(A, b, x0, k, tol)$$

donde $x0$ representa la aproximación de la solución de la que partimos, y tol la tolerancia, de tal forma que el proceso debe finalizar cuando se alcance la condición

$$\frac{\|x_{m+1} - x_m\|_\infty}{\|x_{m+1}\|_\infty} < tol,$$

o el número máximo de iteraciones k, siendo x_m el vector m-ésimo construido mediante el método. Nuestra función debe trabajar con la factorización LU de la matriz A en cada iteración, obtenida una única vez, para lo que nos serviremos de las funciones de los ejercicios anteriores.

23. Obténgase una función que permita determinar si una matriz es simétrica definida positiva, desde el punto de vista numérico, sin calcular determinantes ni valores propios de la matriz. Aplíquese dicha función a la matriz de Hilbert de diferentes órdenes.

24. Prográmense los métodos de Jacobi y de relajación. La estructura de las funciones a crear será de la forma

$$nombre\_función(A, b, x_0, n, tol).$$

La aproximación de la solución del sistema $Ax = b$ debe finalizar cuando se alcanza el número máximo de iteraciones n ó cuando se consigue que

$$\|x_k - x_{k-1}\|_\infty \leq tol \ \ \|x_{k-1}\|_\infty .$$

En el caso de relajación añádase el argumento w a la función.

25. Considérese el sistema de ecuaciones lineales

$$\begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix} x = \begin{pmatrix} 6 \\ 9 \\ 3 \\ 4 \end{pmatrix},$$

cuya solución exacta es $\alpha = [1; 2; 0; 1]$.

(i) Aproxímese la solución mediante el método de Jacobi partiendo del vector nulo con $tol = 10^{-5}$. Estímese el factor asintótico de convergencia generando la sucesión

$$\|x_k - \alpha\|_\infty^{1/k}\,,\ k = 0, 1, 2, \ldots \tag{10.1}$$

y compárese con el valor real $r(B_J)$.

(ii) Realícese la misma tarea con el método de relajación asociado a los valores $w = m/10$, $m \in \mathbb{N}$, $0 \leq m \leq 20$.

Represéntese el número de iteraciones realizado en cada caso para alcanzar la tolerancia frente al parámetro w y determínese de esta forma un valor óptimo del mismo.

Mediante los datos obtenidos represéntese de forma aproximada la función $w \to r(B_w)$, y aproxímese su mínimo y dónde se alcanza. Compárese con el valor real $w_0 = 2/(1 + \sqrt{1 - r(B_J)^2})$.

(Indicación: Se recomienda construir una función que genere la sucesión (10.1) a partir de una sucesión dada $\{x_k\}$ y de α).

26. Mediante el método de relajación con $w = 1.5$ aproxímese la solución del sistema de ecuaciones $Ax = b$, donde $A \in \mathbb{R}^{100\times 100}$ es una matriz bidiagonal inferior con $A(i, i-1) = 1$, $A(i, i) = 1.5$ y $b(i) = 2.5$. La solución exacta del sistema viene dada por $\alpha(i) = 1 - (-2/3)^i$. Úsese como aproximación inicial la solución exacta α.

    Represéntese gráficamente el error relativo cometido en cada iteración para $\|\|_\infty$. Analícese si el método con aritmética exacta es convergente.

27. Créese una función que nos proporcione la matriz inversa generalizada de Moore-Penrose de una matriz $A \in \mathbb{R}^{m\times n}$. Utilícese para ello la función SVD de MATLAB.

    Analícese el problema planteado en clase sobre la utilización numérica del sistema de ecuaciones normales asociado a un sistema de ecuaciones rectangular para resolver el correspondiente problema de mínimos cuadrados (LSP). A tal fin, considérese el sistema

$$\begin{bmatrix} 1 & 1 \\ 10^{-20} & 0 \\ 0 & 10^{-20} \end{bmatrix} x = \begin{bmatrix} 2 \\ 10^{-20} \\ 10^{-20} \end{bmatrix}. \tag{10.2}$$

28. Constrúyase una función que proporcione una factorización $A = QR$ para una matriz $A \in \mathbb{R}^{m\times n}$. Resuélvase mediante la factorización QR el problema (LSP) asociado al sistema (10.2) y compárese con la técnica basada en el sistema de ecuaciones normales.

29. Obténgase una función que proporcione una matriz semejante, del tipo Hessenberg superior, a una matriz $A \in \mathbb{R}^{n\times n}$ dada mediante matrices ortogonales.

30. Sea $A \in \mathbb{R}^{m\times n}$ y sean unos ciertos naturales i, j, k tales que $1 \leq i < j \leq m$, $1 \leq k \leq n$ y $A(j,k) \neq 0$. Constrúyase una función con la estructura

$$G = givensm(A, i, j, k)$$

que proporcione una matriz de Givens $G \in \mathbb{R}^{m\times m}$ tal que $(GA)(j,k) = 0$ y se modifique el elemento $(GA)(i,k)$.

Aplíquese esta función para obtener una matriz ortogonal $M \in \mathbb{R}^{3\times 3}$ tal que $(MA)(3,2) = 0$, $(MA)(2,3) = 0$ y $(MA)(3,3) = 0$, donde

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}.$$

31. Programar el método de la potencia y de la potencia inversa para aproximar valores y vectores propios de una matriz. La estructura de las funciones a crear será de la forma

$$[vp, vecp, nit, tola1, tola2] = nombre\_función(A, x_0, n, tol),$$

donde

- vp es un vector conteniendo las aproximaciones obtenidas del valor propio λ que se desea obtener.
- $vecp$ es una matriz conteniendo las aproximaciones obtenidas de un vector propio v asociado a λ.
- nit es el número de iteraciones realizadas.
- $tola1$ es la distancia relativa entre las dos últimas aproximaciones de λ.
- $tola2$ es la distancia relativa en norma infinito entre las dos últimas aproximaciones de v.
- x_0 es la aproximación inicial a v.
- n es el número máximo de iteraciones que se admite se pueden realizar.
- tol representa la distancia relativa mínima que se admite, tanto entre las aproximaciones consecutivas de λ como de v, para continuar si aún no se han realizado n iteraciones.

32. El objetivo de este ejercicio es testear el comportamiento de las funciones creadas en el ejercicio previo. Para ello es conveniente trabajar en primer término con matrices diagonales como por ejemplo las siguientes:

   a) $diag([1; -3; 2; -0.2; 7])$.
   b) $diag([1; -3; 2; -0.2; -3])$.

c) $diag([1; -3; 2; -0.2; 3])$.

d) $diag([5 + i; -3; 2; -0.2; -3])$.

Considerar la matriz

$$A = \begin{bmatrix} 1 & 1 & 0.5 \\ 1 & 1 & 0.25 \\ 0.5 & 0.25 & 2 \end{bmatrix}.$$

Encontrar el valor propio dominante partiendo de $x_0 = [1; 1; 1]$. Repetir la operación partiendo de $x_0 = [-0.64966116; 0.7482216; 0]$ y comentar los resultados obtenidos. Usar el método de la potencia inversa para aproximar el valor propio más próximo a 1.5.

33. Prográmese el método QR utilizando el comando qr de MATLAB.

Bibliografía

[1] F. Arandiga, R. Donat, y P. Mulet, *Mètodes Numèrics per a L'àlgebra Lineal*, Universitat de València, 2000.

[2] P. G. Ciarlet, Introduction a l'Analyse Numérique Matricielle et a l'Optimisation, Dunod, 1998.

[3] C. Conde y G. Winter, *Métodos y Algoritmos Básicos del Álgebra Numérica*, Editorial Reverté, 1990.

[4] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, 1997.

[5] G. Golub and C. Van Loan, *Matrix Computations*, Jonhs Hopkins University Press, third edition, 1996.

[6] N. J. Higham, *Accuracy ans Stability of Numerical Algorithms*, SIAM, ,Second Edition, 2002.

[7] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, 1980.

[8] Ll. N. Threfethen and D. Bau, *Numerical Linear Algebra*, SIAM, 1997.