Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Matemáticas

Algunos resultados de perturbación para el problema de autovalores

Autora: Eva Garijo Alcalde
Tutora: María Paz Calvo Cabrero
2023

Agradecimientos

Me gustaría agradecer a Mari Paz Calvo Cabrero su entrega y su compromiso en dirigir este TFG, por acompañarme en todo momento, y por transmitirme la belleza de las Matemáticas desde que entré en esta carrera.

También quiero dar las gracias a mis padres, por su apoyo incondicional, y a mi hermano por estar siempre a mi lado.

Índice general

Introducción

El problema del cálculo de los autovalores de una matriz surge de manera natural en distintas ramas de la Ciencia y de la Ingeniería tras la apropiada discretización de los operadores diferenciales o integrales que aparecen al modelizar matemáticamente distintos fenómenos físicos [12]. Como, en general, no es posible calcular de manera exacta dichos autovalores, se recurre al uso de métodos numéricos para su aproximación, pero es preciso disponer de resultados teóricos que determinen el grado de precisión de los autovalores calculados. Por un lado, los coeficientes de las matrices cuyos autovalores se quieren aproximar pueden conocerse solo de manera aproximada si se han obtenido de medidas experimentales y, por otro, los métodos numéricos tienen que ser implementados en un ordenador y están sujetos a los errores derivados de la aritmética finita que éste utiliza. Se hace imprescindible entonces analizar cómo pequeños cambios en los coeficientes de una matriz afectan a sus autovalores y a los autovectores asociados [13] y [6].

El análisis regresivo de los errores trata de interpretar la solución numérica de un problema dado como la solución exacta de un problema del mismo tipo, pero con los datos ligeramente *modificados o perturbados* [4]. De este modo, el estudio de los errores en la aproximación calculada con un método numérico se puede efectuar analizando la diferencia entre dos soluciones exactas de un mismo problema, pero con datos próximos. En este trabajo se pretende profundizar en esta idea cuando el problema considerado es la aproximación numérica de los autovalores de una matriz. Fue Lord Rayleigh quien popularizó y sentó las bases de la teoría de perturbaciones de valores propios en [11], donde uno de sus cálculos tenía como objetivo determinar tanto las frecuencias propias como los modos propios de las vibraciones armónicas de una cuerda con módulo de elasticidad constante y cuya densidad de masa era una pequeña perturbación de un valor constante.

En el caso de matrices y operadores hermíticos, el análisis de sus autovalores y autovectores proporciona una imagen completa de los mismos. No ocurre lo mismo cuando se trata de matrices y operadores no hermíticos para los que, a menudo, las predicciones teóricas no coinciden con las observaciones, en especial cuando los conjuntos de vectores propios asociados están mal acondicionados con respecto a la norma de interés utilizada. En el caso de la norma 2, esto significa que la matriz, o el operador, no es normal y por tanto los vectores propios no son ortogonales. El pseudoespectro ofrece una alternativa analítica y gráfica para el estudio de matrices y operadores no normales.

El objetivo de este trabajo es analizar las variaciones que sufren los autovalores y autovectores de una matriz cuando sus coeficientes se ven afectados por pequeños cambios.

En el primer capítulo de la memoria se estudia, siguiendo [13], la teoría de perturbación para el problema de autovalores. En primer lugar, imponiendo una perturbación que tiende a cero, se deducen de forma exacta, a partir de la definición de autovalor y autovector, las expresiones del autovalor y autovector de la matriz perturbada hasta primer orden en la perturbación.

A continuación se amplía el estudio al caso de autovalores múltiples utilizando el teorema de Gerschgorin, un resultado que se ha estudiado en las asignaturas de Análisis Numérico del grado, y que constituye una herramienta muy útil en este trabajo.

En la segunda parte del capítulo se considera, en particular, el caso de matrices reales simétricas. Se introduce el concepto de acondicionamiento para el problema de autovalores, que permite ampliar el estudio a perturbaciones que no son necesariamente pequeñas y a su vez, deducir resultados de gran valor práctico como son los teoremas de separación de autovalores de matrices simétricas o el principio minimax [13].

En el segundo capítulo de la memoria se consideran las distintas definiciones de pseudoespectro de una matriz que aparecen en [12], y se demuestra la equivalencia de todas ellas, así como las principales propiedades que verifica el pseudoespectro. Finalmente, se representan gráficamente los pseudoespectros de algunas familias de matrices de interés en distintas aplicaciones.

En el trabajo también se han implementado métodos numéricos apropiados para el cálculo de los autovalores y autovectores de una matriz, así como alguno de los procedimientos propuestos en [12] para el cálculo del psudoespectro de una matriz dada. De este modo se han podido ilustrar con ejemplos concretos los distintos resultados de perturbación que se estudian desde el punto de vista teórico a lo largo del trabajo.

Capítulo 1

Teoría de perturbación para el problema de autovalores

Al estudiar la teoría de perturbación para el problema de autovalores es natural que la primera idea que tengamos sea hacerlo analíticamente a través de la definición de autovalor y de autovector asociado. Como veremos, los resultados que se obtienen con esta técnica, al nivel de este trabajo, resultan poco clarificadores. A partir del Teorema de Gerschgorin, una herramienta ya conocida, veremos que estos resultados pueden demostrarse de una manera más intuitiva y clara.

Un caso al que vamos a dedicar especial atención es el de las matrices simétricas. El análisis del problema se simplifica, y a través de la definición del número de condición espectral para el problema de autovalores podremos obtener resultados más potentes que en el caso general. Ilustraremos todos ellos con ejemplos numéricos.

1.1. Primeros resultados de perturbación

Es bien sabido que dada una matriz $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ de orden n, un escalar $\lambda \in \mathbb{C}$ es un *autovalor* de A si existe un vector $\mathbf{x} \in \mathbb{C}^n$ no nulo tal que

$$A\mathbf{x} = \lambda\mathbf{x}.$$

Decimos que $\mathbf{x}$ es un *autovector* de A asociado a λ. Es inmediato comprobar que, de manera equivalente, λ es autovalor de A si es solución de la ecuación

$$|\lambda I - A| = \det(\lambda I - A) = 0, \tag{1.1}$$

donde $|\lambda I - A|$ es un polinomio en λ de grado n conocido como *polinomio característico de* A. Llamamos *multiplicidad algebraica* ($m(\lambda)$) de un autovalor λ a la multiplicidad de este como raíz del polinomio característico (1.1). Por otro lado, llamamos *multiplicidad geométrica* ($d(\lambda)$) de λ a la dimensión del subespacio vectorial formado por todos los autovectores asociados a λ, esto es, la dimensión de $V_\lambda = \{\mathbf{x} \in \mathbb{C}^n : A\mathbf{x} = \lambda\mathbf{x}\}$. Un resultado fundamental, cuya demostración puede encontrarse en [10] es que para todo autovalor λ de A,

$$1 \leq d(\lambda) \leq m(\lambda).$$

Una vez revisados los conceptos básicos relacionados con los autovalores de una matriz, pasamos a describir los primeros resultados de perturbación para el problema de autovalores. Nuestro objetivo es comparar los autovalores de una matriz dada A con los autovalores de una matriz perturbada $A + \epsilon B$, donde B es otra matriz (posiblemente aleatoria) y ϵ es un parámetro real y positivo que da cuenta del tamaño de la perturbación. Si λ_1 es un autovalor de A, denotamos por $\lambda_1(\epsilon)$ al correspondiente autovalor de la matriz perturbada $A + \epsilon B$, es decir, que satisface $\lambda_1(0) = \lambda_1$.

Ejemplo 1.1.

En la Figura 1.1 se ha representado, haciendo uso de la función `eig` de Matlab, la evolución de los autovalores de una matriz simétrica A, cuando sobre ella actúa una perturbación, también simétrica, de tamaño $\epsilon \in [0, 1]$.

Para construir la matriz A se ha utilizado como punto de partida una matiz de Clement [3]. La forma general de una matriz de Clement de orden n es como sigue

$$A_n = \begin{pmatrix} 0 & 1 & & & \\ n-1 & 0 & 2 & & \\ & \ddots & \ddots & \ddots & \\ & & 2 & 0 & n-1 \\ & & & 1 & 0 \end{pmatrix}. \tag{1.2}$$

Se trata de una matriz triadiagonal, con ceros en su diagonal principal y cuyos autovalores son bien conocidos y están dados por $\{\pm(n-1), \pm(n-3), \ldots, \pm 1\}$ si n es par, y hay que añadir a estos el autovalor 0 si n es impar.

La Figura 1.1 se ha obtenido calculando los autovalores de la matriz $\tilde{A}_{10} + \epsilon\, B$, donde $\tilde{A}_{10}$ es la matriz de Clement (1.2) de orden $n = 10$ simetrizada y normalizada ($\tilde{A}_{10} = (A_{10} + A_{10}^T)/(n-1)$). Por su parte, B es una matriz simétrica fija que se ha generado aleatoriamente y cuyos elementos están todos en el intervalo $[0, 1]$. Tomamos $\tilde{A}_{10}$ y B simétricas para que los autovalores sean reales. Por último, ϵ recorre los valores del intervalo $[0, 1]$.

Es interesante observar que las trayectorias seguidas por los distintos autovalores $\lambda_i(\epsilon)$, $i = 1, 2, 3, 4$, de la matriz $\tilde{A}_{10} + \epsilon B$, no llegan a cortarse [7]. Es decir, las matrices $\tilde{A}_{10} + \epsilon B$ no llegan a tener autovalores múltiples.

1.1.1. Perturbación de autovalores simples

Consideramos en primer lugar el caso en que λ_1 es un autovalor simple de la matriz A, esto es, λ_1 es raíz simple del polinomio característico de A. Como hemos mencionado, nuestro objetivo es estudiar y comparar dicho autovalor λ_1 con el correspondiente autovalor $\lambda_1(\epsilon)$ de una matriz perturbada $A + \epsilon B$. Nos centramos en el caso de dos matrices A y B de orden n cuyos elementos satisfacen

$$|a_{ij}| < 1, \;\; |b_{ij}| < 1, \;\; 1 \leq i, j \leq n.$$

Comenzamos escribiendo la ecuación característica de A, desarrollando el determinante en potencias de λ

$$\det(\lambda I - A) \equiv \lambda^n + c_{n-1}\lambda^{n-1} + c_{n-2}\lambda^{n-2} + \cdots + c_0 = 0, \tag{1.3}$$

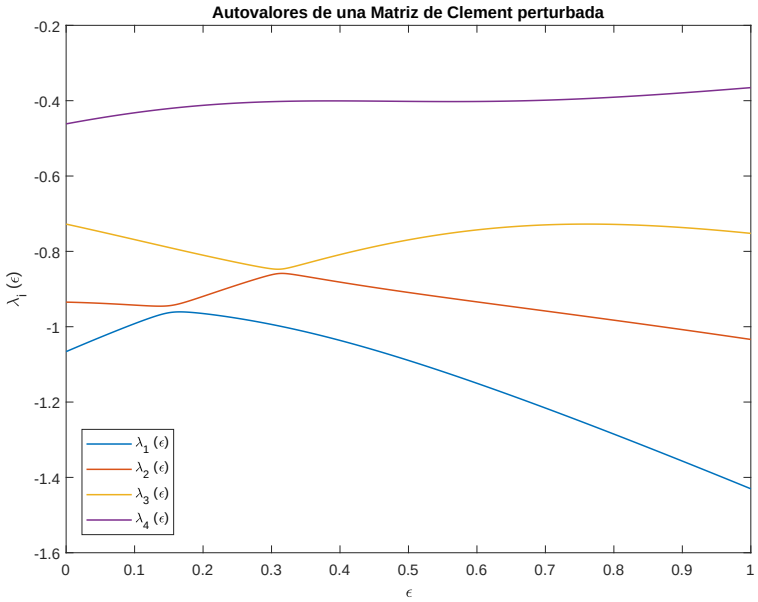

Figura 1.1: Primeros cuatro autovalores $\lambda_r(\epsilon)$ de una matriz de Clement perturbada.

y la ecuación característica de $A + \epsilon B$

$$\det(\lambda I - A - \epsilon B) \equiv \lambda^n + c_{n-1}(\epsilon)\lambda^{n-1} + c_{n-2}(\epsilon)\lambda^{n-2} + \cdots + c_0(\epsilon) = 0. \tag{1.4}$$

Desarrollando la expresión del determinante de (1.4) por su primera columna, vemos que cada coeficiente $c_r(\epsilon),\ 0 \leq r \leq n-1$, es un polinomio en ϵ de grado $n-r$, que cumple $c_r(0) = c_r$.

Por lo tanto, podemos escribir

$$c_r(\epsilon) = c_r + c_{r1}\epsilon + c_{r2}\epsilon^2 + \cdots + c_{r,n-r}\epsilon^{n-r}, \quad 0 \leq r \leq n.$$

Puesto que λ_1 es una raíz simple de la ecuación característica (1.3) estamos en condiciones de aplicar el Teorema A.1.1 sobre funciones algebraicas (ver Anexo), que garantiza que para ϵ suficientemente pequeño existe una raíz simple $\lambda_1(\epsilon)$ de (1.4) dada por la serie de potencias convergente

$$\lambda_1(\epsilon) = \lambda_1 + k_1\epsilon + k_2\epsilon^2 + \cdots . \tag{1.5}$$

Del desarrollo en serie de potencias (1.5) concluimos que $\lambda_1(\epsilon) \to \lambda_1$ cuando $\epsilon \to 0$. Además, independientemente de la multiplicidad de los restantes autovalores de la matriz A hemos obtenido

$$|\lambda_1(\epsilon) - \lambda_1| = O(\epsilon). \tag{1.6}$$

Pasamos ahora a estudiar la perturbación de un autovector $\mathbf{x}_1$ asociado al autovalor simple λ_1 de la matriz A. En primer lugar, escribimos la expresión explícita de $\mathbf{x}_1$.

Partimos de un resultado conocido de la teoría general de autovalores y autovectores que afirma que

$$\dim(V_{\lambda_1}) = n - rg(A - \lambda_1 I),$$

donde V_{λ_1} es el subespacio generado por el autovector $\mathbf{x}_1$. Por ser λ_1 un autovalor simple $\dim(V_{\lambda_1}) = 1$ y por tanto $rg(A - \lambda_1 I) = n - 1$. Entonces existe al menos un menor no nulo de orden $n-1$ de $(A - \lambda_1 I)$. Podemos suponer sin pérdida de generalidad que este menor está formado por las primeras $n-1$ filas y $n-1$ columnas de $(A - \lambda_1 I)$.

De la resolución del sistema lineal de ecuaciones que define el autovector $\mathbf{x}_1$, resulta que

$$\mathbf{x}_1 = [A_{n1}, A_{n2}, \ldots, A_{nn}]^T,$$

donde A_{ni} denota el cofactor del elemento (n, i) de la matriz A, que sabemos será un polinomio en λ_1 de grado menor o igual que $n-1$, como puede demostrarse fácilmente utilizando la regla de Cramer.

Sea $\mathbf{x}_1(\epsilon)$ el autovector asociado al autovalor $\lambda_1(\epsilon)$ de la matriz perturbada $A + \epsilon B$. Podemos aplicar los mismos resultados obteniendo que las componentes de $\mathbf{x}_1(\epsilon)$ serán ahora polinomios en $\lambda_1(\epsilon)$ y ϵ. Además, puesto que la serie de potencias que define $\lambda_1(\epsilon)$ es convergente para ϵ suficientemente pequeño, las componentes de $\mathbf{x}_1(\epsilon)$ estarán dadas por una serie de potencias en ϵ convergente, cuyo término independiente ha de ser la correspondiente componente de $\mathbf{x}_1$. Podemos escribir entonces

$$\mathbf{x}_1(\epsilon) = \mathbf{x}_1 + \epsilon \mathbf{z}_1 + \epsilon^2 \mathbf{z}_2 + \cdots, \tag{1.7}$$

donde cada $\mathbf{z}_i$ es a su vez una serie de potencias en ϵ convergente. De nuevo en correspondencia con (1.6), obtenemos que

$$|\mathbf{x}_1(\epsilon) - \mathbf{x}_1| = O(\epsilon).$$

Si nos centramos ahora en el caso particular en que A es una matriz diagonalizable, entonces existen sendos conjuntos $\{\mathbf{x}_i\}_{i=1}^n$ e $\{\mathbf{y}_i\}_{i=1}^n$ de vectores propios de A por la derecha y por la izquierda, respectivamente, que cumplen

$$\mathbf{y}_i^T \mathbf{x}_j = 0 \ \ (i \neq j).$$

Además, estos vectores propios son únicos (salvo producto por un escalar) solo si todos los autovalores son simples. De esta forma, por ser $\{\mathbf{x}_i, \ \ i = 1, \ldots, n\}$ una base de $\mathbb{C}^n$ podemos expresar los vectores $\mathbf{z}_i$ de (1.7) como combinación lineal de los autovectores de A

$$\mathbf{z}_i = \sum_{j=1}^{n} s_{ji} \mathbf{x}_j, \quad i = 1, 2, \ldots, n.$$

Sustituyendo en (1.7) tenemos

$$\mathbf{x}_1(\epsilon) = \mathbf{x}_1 + \epsilon \sum_{j=1}^{n} s_{j1} \mathbf{x}_j + \epsilon^2 \sum_{j=1}^{n} s_{j2} \mathbf{x}_j + \cdots,$$

y agrupando los términos que multiplican a cada autovector $\mathbf{x}_i$ queda

$$\begin{aligned} \mathbf{x}_1(\epsilon) =& (1 + \epsilon s_{11} + \epsilon^2 s_{12} + \cdots)\mathbf{x}_1 + (\epsilon s_{21} + \epsilon^2 s_{22} + \cdots)\mathbf{x}_2 \\ &+ \cdots + (\epsilon s_{n1} + \epsilon^2 s_{n2} + \cdots)\mathbf{x}_n. \end{aligned}$$

La convergencia de cada una de las series de portencias que aparecen entre paréntesis viene dada por la convergencia absoluta de la serie en (1.7). Como estamos interesados en comparar $\mathbf{x}_1$ con un autovector de la matriz perturbada asociado al autovalor $\lambda_1(\epsilon)$, en la expresión obtenida para $\mathbf{x}_1(\epsilon)$ dividimos entre $(1 + \epsilon s_{11} + \epsilon^2 s_{12} + \cdots)$, que es no nulo para ϵ suficientemente pequeño. Redefinimos así el autovector de $A + \epsilon B$ con el que vamos a comparar, que volvemos a denotar como $\mathbf{x}_1(\epsilon)$, de la siguiente manera

$$\mathbf{x}_1(\epsilon) = \mathbf{x}_1 + (\epsilon t_{21} + \epsilon^2 t_{22} + \cdots)\mathbf{x}_2 + \cdots + (\epsilon t_{n1} + \epsilon^2 t_{n2} + \cdots)\mathbf{x}_n, \tag{1.8}$$

donde de nuevo las series entre paréntesis son convergentes para ϵ suficientemente pequeño.

En el análisis que hemos hecho hasta ahora hemos utilizado los vectores propios $\mathbf{x}_i$ obtenidos directamente de las ecuaciones $(A - \lambda_i I)\mathbf{x}_i = \mathbf{0}$ que los definen. Podemos normalizar y renombrar dichos vectores para tener

$$\|\mathbf{x}_i\|_2 = 1, \quad 1 \leq i \leq n.$$

Esta acción introduce un factor multiplicativo adicional en cada uno de los términos entre paréntesis de (1.8) que puede ser absorbido por los coeficientes t_{ij}. Notamos, sin embargo, que haciendo esto, los autovectores perturbados $\mathbf{x}_i(\epsilon)$ no están normalizados cuando $\epsilon \neq 0$.

Expresión explícita de la perturbación

Nuestro objetivo a continuación es obtener una relación explícita que nos permita obtener la expresión exacta en términos de $\mathbf{x}_i$ e $\mathbf{y}_j$ del autovector perturbado hasta primer orden en ϵ. Para ello, en primer lugar, definimos las cantidades

$$s_i = \mathbf{y}_i^T \mathbf{x}_i, \ \ (i = 1, \ldots, n),$$

donde los vectores propios por la derecha $\{\mathbf{x}_i\}_{i=1}^n$ y por la izquierda $\{\mathbf{y}_i\}_{i=1}^n$ de A, tienen norma euclídea unidad.

Notamos que si A tiene autovalores múltiples entonces los correspondientes $\mathbf{x}_i$ e $\mathbf{y}_i$ no serán únicos. En ese caso vamos a suponer que el valor del s_i correspondiente que estamos usando proviene de alguna elección concreta de $\mathbf{x}_i$ e $\mathbf{y}_i$. Por otra parte, los vectores propios asociados a autovalores simples están definidos salvo producto por un escalar de módulo unidad (para conservar la normalización de los vectores) y en ese caso $|s_i|$ sí está completamente definido, como podemos comprobar fácilmente

$$|\tilde{s}_i| = |\tilde{\mathbf{y}}_i^T \tilde{\mathbf{x}}_i| = |\rho_1 \mathbf{y}_i^T \rho_2 \mathbf{x}_i| = |\rho_1||\rho_2||\mathbf{y}_i^T \mathbf{x}_i| = |\mathbf{y}_i^T \mathbf{x}_i| = |s_i|.$$

En cualquier caso, se cumple que

$$|s_i| = |\mathbf{y}_i^T \mathbf{x}_i| \leq \|\mathbf{y}_i\|_2 \|\mathbf{x}_i\|_2 = 1.$$

Por otro lado, dada la matriz B que determina la perturbación, definimos también los escalares

$$\beta_{ij} = \mathbf{y}_i^T B \mathbf{x}_j, \ \ (i, j = 1, \ldots, n), \tag{1.9}$$

y utilizando que $\|B\|_2 \leq n$ y propiedades básicas de normas matriciales tenemos

$$|\beta_{ij}| = |\mathbf{y}_i^T B \mathbf{x}_j| \leq \|\mathbf{y}_i\|_2 \|B\mathbf{x}_j\|_2 \leq \|\mathbf{y}_i\|_2 \|B\|_2 \|\mathbf{x}_j\|_2 \leq n. \tag{1.10}$$

Con todo ello, podemos enunciar el siguiente resultado.

Teorema 1.1.1 (Autovalor simple a primer orden en ϵ) *Sea λ_1 un autovalor simple de una matriz $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ diagonalizable y sea $\lambda_1(\epsilon)$ el correspondiente autovalor de la matriz perturbada $A + \epsilon B$, donde $\epsilon > 0$ y cada elemento de B verifica $|b_{ij}| < 1$. Entonces*

$$\lambda_1(\epsilon) = \lambda_1 + k_1\epsilon + O(\epsilon^2), \tag{1.11}$$

donde

$$k_1 = \frac{\beta_{11}}{s_1} = \frac{\mathbf{y}_1^T B \mathbf{x}_1}{\mathbf{y}_1^T \mathbf{x}_1}. \tag{1.12}$$

Demostración.

La ecuación que cumplen el autovalor perturbado $\lambda_1(\epsilon)$ y el correspondiente autovector $\mathbf{x}_1(\epsilon)$

$$(A + \epsilon B)\mathbf{x}_1(\epsilon) = \lambda_1(\epsilon)\mathbf{x}_1(\epsilon) \tag{1.13}$$

nos proporciona una igualdad entre series de potencias en ϵ, pues $\lambda_1(\epsilon)$ y cada componente de $\mathbf{x_1}(\epsilon)$ lo son. Por tanto, lo que debemos hacer es igualar los términos que acompañan a ϵ en cada lado de la igualdad (1.13). Utilizando (1.5) y (1.8) queda

$$A\left(\sum_{i=2}^{n} t_{i1}\mathbf{x}_i\right) + B\mathbf{x}_1 = \lambda_1\left(\sum_{i=2}^{n} t_{i1}\mathbf{x}_i\right) + k_1\mathbf{x}_1. \tag{1.14}$$

Por ser $\{\mathbf{x}_i\}_{i=1}^n$ autovectores de A asociados a los autovalores $\{\lambda_i\}_{i=1}^n$, la igualdad (1.14) se convierte en

$$\sum_{i=2}^{n}(\lambda_i - \lambda_1)t_{i1}\mathbf{x}_i + B\mathbf{x}_1 = k_1\mathbf{x}_1. \tag{1.15}$$

Multiplicando ambos términos de la igualdad anterior por $\mathbf{y}_1^T$, y recordando que $\mathbf{y}_1^T\mathbf{x}_i = 0$ para todo $i \neq 1$, queda

$$\mathbf{y}_1^T B\mathbf{x}_1 = k_1\mathbf{y}_1^T\mathbf{x}_1 \Longrightarrow k_1 = \frac{\beta_{11}}{s_1}.$$

De (1.10) se sigue que

$$|k_1| \leq \frac{n}{|s_1|}.$$

□

Hemos obtenido, por tanto, que para ϵ suficientemente pequeño, para el que los términos de orden mayor o igual que 2 en el desarrollo en potencias de ϵ de $\lambda_1(\epsilon)$ puedan despreciarse, se tiene

$$\lambda_1(\epsilon) \approx \lambda_1 + k_1\epsilon,$$

esto es, la diferencia $\lambda_1(\epsilon) - \lambda_1$ se comporta como $k_1\epsilon$. No obstante, en la relación (1.12) debemos tener en cuenta que la cantidad $|s_1|$ puede ser arbitrariamente pequeña.

Para completar el estudio del problema de perturbación a primer orden, vamos a obtener ahora una ecuación explícita que nos dé la expresión exacta a primer orden en ϵ del autovector perturbado $\mathbf{x}_1(\epsilon)$.

Teorema 1.1.2 (Autovector asociado a primer orden en ϵ) *Con las mismas hipótesis que en el Teorema 1.1.1, si $\mathbf{x}_1$ es el autovector de A asociado al autovalor simple λ_1 y $\mathbf{x}_1(\epsilon)$ el correspondiente autovector perturbado asociado a $\lambda_1(\epsilon)$ entonces*

$$\mathbf{x}_1(\epsilon) = \mathbf{x}_1 + \epsilon \left[\frac{\beta_{21}\mathbf{x}_2}{(\lambda_1 - \lambda_2)s_2} + \frac{\beta_{31}\mathbf{x}_3}{(\lambda_1 - \lambda_3)s_3} + \cdots + \frac{\beta_{n1}\mathbf{x}_n}{(\lambda_1 - \lambda_n)s_n} \right] + O(\epsilon^2). \qquad (1.16)$$

Demostración.

Partimos de la expresión (1.15) y multiplicamos ahora los dos términos de la igualdad por $\mathbf{y}_i^T$ con $i \neq 1$ obteniendo

$$(\lambda_i - \lambda_1)t_{i1}\mathbf{y}_i^T\mathbf{x}_i + \mathbf{y}_i^T B\mathbf{x}_1 = 0, \quad (i = 2, 3, \ldots, n),$$

esto es,

$$(\lambda_i - \lambda_1)t_{i1}s_i + \beta_{i1} = 0, \quad (i = 2, 3, \ldots, n), \qquad (1.17)$$

de donde despejamos $t_{i1} = \dfrac{\beta_{i1}}{(\lambda_1 - \lambda_i) \cdot s_i}$ y sustituimos en el primer paréntesis de (1.14) para obtener (1.16).

□

Del resultado obtenido notamos que si A es una matriz tal que el autovalor λ_1 que estamos estudiando está lo suficientemente separado del resto de autovalores λ_i, $i \neq 1$, y tal que las cantidades s_i, $i \neq 1$, no son demasiado pequeñas, entonces la diferencia $\mathbf{x}_1(\epsilon) - \mathbf{x}_1$ es poco sensible a las perturbaciones de A. No obstante, si alguna de las situaciones anteriores no ocurre, entonces el autovector $\mathbf{x}_1$ es muy sensible a los posibles cambios en la matriz A.

Procediendo de la misma forma podemos hallar el coeficiente de ϵ^2 en el desarrollo en potencias de $\lambda_1(\epsilon)$ y de $\mathbf{x}_1(\epsilon)$ así como el resto de términos de orden superior en ambos.

Podemos ilustrar esta idea hallando la expresión de k_2 en (1.5). Igualando los coeficientes de ϵ^2 en ambos lados de la igualdad (1.13) obtenemos

$$A\left(\sum_{i=2}^{n} t_{i2}\mathbf{x}_i\right) + B\left(\sum_{i=2}^{n} t_{i1}\mathbf{x}_i\right) = k_2\mathbf{x}_1 + k_1\left(\sum_{i=2}^{n} t_{i1}\mathbf{x}_i\right) + \lambda_1\left(\sum_{i=2}^{n} t_{i2}\mathbf{x}_i\right),$$

es decir

$$\sum_{i=2}^{n} t_{i2}(\lambda_i - \lambda_1)\mathbf{x}_i + \sum_{i=2}^{n} t_{i1}B\mathbf{x}_i = k_2\mathbf{x}_1 + k_1\left(\sum_{i=2}^{n} t_{i1}\mathbf{x}_i\right).$$

Multiplicando ambos miembros de la igualdad por $\mathbf{y}_1^T$ por la izquierda queda

$$\sum_{i=2}^{n} t_{i1}\beta_{1i} = k_2 s_1,$$

y despejando k_2 y sustituyendo el valor de t_{i1} que ya conocemos de (1.17) obtenemos

$$k_2 = \frac{1}{s_1}\sum_{i=2}^{n} t_{i1}\beta_{1i} = \frac{1}{s_1}\sum_{i=2}^{n} \frac{\beta_{i1}\beta_{1i}}{(\lambda_1 - \lambda_i)s_i}.$$

Cuando la matriz A tiene autovalores múltiples, el estudio de la perturbación de estos autovalores utilizando el polinomio característico de la matriz es poco intuitivo, por lo que a continuación abordamos el problema utilizando un resultado bien conocido y de gran valor práctico.

1.2. Teoría de perturbación basada en el teorema de Gerschgorin

Comenzamos enunciando y probando el teorema de Gerschgorin. Aun siendo un resultado que se ha estudiado en las asignaturas de Análisis Numérico del grado, lo incluimos aquí ya que, como veremos, permite obtener resultados más fuertes e intuitivos para el estudio de la perturbación de autovalores y autovectores.

Teorema 1.2.1 (Teorema de Gerschgorin) *Sean $\lambda_1, \lambda_2, \ldots, \lambda_n$ los autovalores de una matriz cuadrada $A = (a_{ij})$, $1 \leq i, j \leq n$, repetidos con su multiplicidad. Para cada $1 \leq i \leq n$, sea $D_i = D(a_{ii}, r_i)$ el disco cerrado del plano complejo centrado en a_{ii} de radio $r_i = \sum_{i \neq j} |a_{ij}|$. Entonces*

1. *Cada autovalor λ_i de A está en la unión $\mathcal{R} = \bigcup_{i=1}^{n} D_i$.*
2. *Si una componente conexa de dicha región $\mathcal{R}$ del plano complejo está formada por s discos, entonces contiene exactamente s autovalores de A.*

Demostración.

Comenzamos con la prueba de la primera afirmación del teorema. Sea λ un autovalor de la matriz A. Entonces existe $\mathbf{x} \neq \mathbf{0}$ autovector asociado a λ tal que

$$A\mathbf{x} = \lambda\mathbf{x},$$

esto es, para cada $1 \leq i \leq n$ se tiene

$$\sum_{j=1}^{d} a_{ij}x_j = \lambda x_i$$

y

$$\sum_{j \neq i} a_{ij}x_j = \lambda x_i - a_{ii}x_i = (\lambda - a_{ii})x_i.$$

Tomando módulos en la igualdad anterior y aplicando la desigualdad triangular obtenemos

$$|\lambda - a_{ii}||x_i| = |\sum_{j\neq i} a_{ij}x_j| \leq \sum_{j\neq i} |a_{ij}||x_j| \leq \|\mathbf{x}\|_\infty \sum_{j\neq i} |a_{ij}|, \quad 1 \leq i \leq n.$$

Sea ahora i_0 tal que $\|\mathbf{x}\|_\infty = |x_{i_0}| \neq 0$. Entonces

$$|\lambda - a_{i_0 i_0}||x_{i_0}| \leq |x_{i_0}| \sum_{j\neq i_0} |a_{i_0 j}|,$$

y, tras dividir por $|xi_0|$, concluimos que

$$d(\lambda, a_{i_0 i_0}) = |\lambda - a_{i_0 i_0}| \leq \sum_{j\neq i_0} |a_{i_0 j}| = r_{i_0},$$

es decir,

$$\lambda \in D(a_{i_0 i_0}, r_{i_0}) \subseteq \mathcal{R}.$$

Vamos ahora con la segunda parte de la demostración. Sea $t \in [0,1]$, contruimos la matriz $B(t) = \Delta + t(A - \Delta)$ donde $\Delta = \text{diag}(a_{11}, a_{22}, \ldots, a_{nn})$. Es decir, $B(t)$ es una matriz tal que sus elementos diagonales son los elementos de la diagonal de A, y sus elementos no diagonales son t veces los correspondientes elementos no diagonales de A. Consideramos la aplicación

$$\begin{array}{ccc} [0,1] & \longrightarrow & \mathcal{M}_{n\times n} \\ t & \longmapsto & B(t) \end{array}$$

que es continua, pues para cada coeficiente es una aplicación afín. Además,

$$B(0) \;=\; \Delta, \qquad B(1) \;=\; A.$$

A continuación, consideramos los autovalores $\lambda_1(t), \lambda_2(t), \ldots, \lambda_n(t)$ de la matriz $B(t)$ para cada $t \in [0,1]$, en particular $\lambda_i(0) = a_{ii}$ y $\lambda_i(1) = \lambda_i$. Si aplicamos a $B(t)$ la primera parte del teorema, obtenemos que el $i-$ésimo disco para $B(t)$ tiene por centro a_{ii} y radio tr_i. Entonces los autovalores $\lambda_1(t), \lambda_2(t), \ldots, \lambda_n(t)$ de la matriz $B(t)$ están contenidos en la unión de discos para $B(t)$ y por tanto, también en la unión de discos para A, que tienen el mismo centro y radio mayor.

Entonces para cada $1 \leq i \leq n$ tomamos la aplicación ϕ definida por

$$\begin{array}{rccc} \phi: & [0,1] & \longrightarrow & \mathbb{C} \\ & t & \longmapsto & \lambda_i(t). \end{array}$$

La aplicación ϕ es continua en virtud del teorema de la función implícita pues cada imagen $\lambda_i(t)$ es raíz de la ecuación polinómica $\det(B(t) - \lambda I) = 0$. Además, la imagen está contenida en la unión de discos $\mathcal{R}$ de A.

Por ser ϕ continua, la imagen de cualquier conjunto conexo es también conexa. En particular, por ser $[0,1]$ conexo, el camino seguido por $\lambda_i(t)$ es conexo, y por tanto, comienza y termina en la misma componente conexa de $\mathcal{R}$. Así, vamos a tener tantos $\lambda_i(1)$ contenidos en cada componente conexa como $\lambda_i(0) = a_{ii}$ estén en dicha componente. Y puesto que los a_{ii} son los centros de los discos cuya unión forman la región $\mathcal{R}$, tantos como discos formen dicha componente conexa.

□

Vamos a estudiar a continuación el problema de perturbación para autovalores y autovectores aplicando el teorema que acabamos de demostrar.

1.2.1. Perturbación de autovalores simples de una matriz diagonalizable.

Como ya sabemos, si A es una matriz de orden n diagonalizable entonces tiene n vectores propios linealmente independientes, y por tanto existe una matriz H invertible tal que

$$H^{-1}AH = \text{diag}(\lambda_1, \ldots, \lambda_n), \tag{1.18}$$

donde las columnas de H son paralelas a los autovectores $\{\mathbf{x}_i\}_{i=1}^n$ por la derecha de A, y las filas de H^{-1} son paralelas a los traspuestos de los autovectores $\{\mathbf{y}_i\}_{i=1}^n$ por la izquierda de A.

Podemos tomar los autovectores normalizados, de forma que

$$\|\mathbf{x}_i\|_2 = \|\mathbf{y}_i\|_2 = 1, \quad i = 1, \ldots, n.$$

Sea entonces $\mathbf{x}_i$ la i-ésima columna de H e $\dfrac{\mathbf{y}_i^T}{s_i}$ con $s_i = \mathbf{y}_i^T\mathbf{x}_i \neq 0$ la i-ésima fila de H^{-1}. Notamos de nuevo que por ser λ_1 un autovalor simple, los vectores propios $\mathbf{x}_1$ e $\mathbf{y}_1$ son únicos salvo producto por un escalar, que debe ser de módulo unidad para que sigan estando normalizados.

Nuestro objetivo ahora es estudiar los autovalores de la matriz perturbada $A + \epsilon B$ utilizando el Teorema de Gerschgorin 1.2.1. Dicha matriz tiene exactamente los mismos autovalores que la matriz semejante que resulta de hacer $H^{-1}(A+\epsilon B)H$ cuya expresión es

$$H^{-1}(A+\epsilon B)H = \text{diag}(\lambda_1, \ldots, \lambda_n) + \epsilon \begin{pmatrix} \frac{\beta_{11}}{s_1} & \frac{\beta_{12}}{s_1} & \ldots & \frac{\beta_{1n}}{s_1} \\ \frac{\beta_{21}}{s_2} & \frac{\beta_{22}}{s_2} & \ldots & \frac{\beta_{2n}}{s_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\beta_{n1}}{s_n} & \frac{\beta_{n2}}{s_n} & \ldots & \frac{\beta_{nn}}{s_n} \end{pmatrix}, \tag{1.19}$$

donde hemos utilizado la definición de $\beta_{ij} = \mathbf{y}_i^T B \mathbf{x}_j$ dada por (1.9).

Aplicando ahora el Teorema de Gerschgorin obtenemos que los autovalores perturbados (los autovalores de $A+\epsilon B$) están en la unión de los discos de centro $\lambda_i + \epsilon\frac{\beta_{ii}}{s_i}$ y radio $r_i = \sum_{j\neq i} \epsilon|\frac{\beta_{ij}}{s_i}|$.

Ya hemos visto que bajo la hipótesis $|b_{ij}| < 1$ para todo $1 \leq i, j \leq n$ se tiene que $|\beta_{ij}| \leq n$, entonces el radio del i-ésimo disco será

$$r_i = \sum_{j\neq i} \epsilon \left|\frac{\beta_{ij}}{s_i}\right| = \frac{\epsilon}{|s_i|}\sum_{j\neq i}|\beta_{ij}| \leq \frac{\epsilon}{|s_i|}\sum_{j\neq i} n = \frac{\epsilon n(n-1)}{|s_i|}.$$

Finalmente, como λ_1 es un autovalor simple de A, esto es $|\lambda_1 - \lambda_i| > 0$, para $i = 2, \ldots, n$, entonces existe ϵ suficientemente pequeño para el que $\lambda_1(\epsilon)$ es el único autovalor en el disco correspondiente.

El resultado obtenido es algo decepcionante, pues esperaríamos de nuestro análisis previo (1.11) que el autovalor $\lambda_1(\epsilon)$ correspondiente a λ_1 estuviera en un disco centrado en $\lambda_1 + \epsilon\dfrac{\beta_{11}}{s_1}$ pero de radio $r_1 = O(\epsilon^2)$ cuando $\epsilon \to 0$.

A continuación vamos a ver que, efectivamente, es posible reducir el radio de los discos de Gerschgorin para obtener el resultado esperado.

En primer lugar notamos que dada una matriz A tal que λ es un autovalor de A y $\mathbf{x}$ un autovector asociado, $A\mathbf{x} = \lambda\mathbf{x}$, entonces λ también es autovalor de la matriz obtenida de A multiplicando su $i-$ésima fila por $\frac{1}{m}$ y su $i-$ésima columna por m. En efecto, si tomamos $i = 1$, podemos ver que las ecuaciones características de ambas matrices coinciden sin más que desarrollar los determinantes por la primera columna.

Aplicando este razonamiento, y utilizando $i = 1$ y $m = k/\epsilon$, con k un parámetro por determinar, llegamos a que la matriz (1.19) tiene los mismos autovalores que la matriz

$$\operatorname{diag}(\lambda_1, \ldots, \lambda_n) + \begin{pmatrix} \epsilon\frac{\beta_{11}}{s_1} & \epsilon^2\frac{\beta_{12}}{ks_1} & \cdots & \epsilon^2\frac{\beta_{1n}}{ks_1} \\ \frac{k\beta_{21}}{s_2} & \epsilon\frac{\beta_{22}}{s_2} & \cdots & \epsilon\frac{\beta_{2n}}{s_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{k\beta_{n1}}{s_n} & \epsilon\frac{\beta_{n2}}{s_n} & \cdots & \epsilon\frac{\beta_{nn}}{s_n} \end{pmatrix}. \tag{1.20}$$

Nuestro objetivo es tomar el valor del parámetro k de manera que el primer disco de Gerschgorin sea tan pequeño como sea posible, manteniendo el resto de discos sin intersecar con el primero.

Podemos observar que los elementos de la primera fila salvo el $(1,1)$ contienen el factor ϵ^2, mientras que los elementos de la primera columna salvo el $(1,1)$ son independientes de ϵ.

Lema 1.2.1 *Dado ϵ suficientemente pequeño, para que la componente conexa de D_1 solo contenga este disco, k puede ser el mayor valor consistente con*

$$\left| k\frac{\beta_{i1}}{s_i} \right| \leq |\lambda_1 - \lambda_i|, \quad i = 2, \ldots, n. \tag{1.21}$$

Demostración.

Aplicando el Teorema de Gerschgorin a la matriz (1.20) sabemos que

- $\lambda_1(\epsilon)$ está en el disco de centro $c_1 = \lambda_1 + \epsilon\dfrac{\beta_{11}}{s_1}$ y radio $r_1 = \dfrac{\epsilon^2}{|ks_1|}\displaystyle\sum_{j\neq 1} |\beta_{1j}|$.

- Para cada $i \neq 1$, $\lambda_i(\epsilon)$ está en el disco de centro $c_i = \lambda_i + \epsilon\dfrac{\beta_{ii}}{s_i}$ y radio $r_i = |k\dfrac{\beta_{i1}}{s_i}| + \dfrac{\epsilon}{|s_i|}\displaystyle\sum_{j\neq 1,i} |\beta_{ij}|$.

Es claro que la condición para que los discos D_1 y D_i no solapen es

$$|c_i - c_1| > r_1 + r_i,$$

como puede verse en la Figura 1.2.

Vamos a demostrar que la condición (1.21) es suficiente para que esto suceda en el caso de autovalores reales. Para ello, supongamos sin pérdida de generalidad que $\lambda_1 > \lambda_i$.

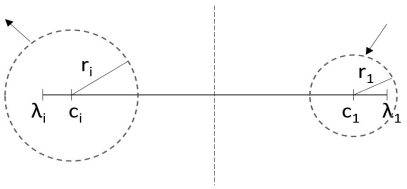

Figura 1.2: Discos D_1 y D_i dados por el Teorema de Gerschgorin, para ilustrar la demostración del Lema 1.2.1

En el peor de los casos, la situación es como la de la Figura 1.2 donde $c_1 < \lambda_1$ y $c_i > \lambda_i$. Comenzamos notando que

$$c_i + r_i = \lambda_i + \epsilon\frac{\beta_{ii}}{s_i} + |k\frac{\beta_{i1}}{s_i}| + \frac{\epsilon}{|s_i|}\sum_{j\neq 1,i}|\beta_{ij}| \leq \lambda_i + \frac{1}{2}|\lambda_1 - \lambda_i| + \frac{\epsilon}{|s_i|}\sum_{j\neq 1}|\beta_{ij}|.$$

Si ϵ es suficientemente pequeño para que $\dfrac{\epsilon}{|s_i|}\displaystyle\sum_{j\neq 1}|\beta_{ij}| \leq \frac{1}{4}|\lambda_1 - \lambda_i|$, obtenemos que

$$c_i + r_i \leq \lambda_i + \left(\frac{1}{2} + \frac{1}{4}\right)|\lambda_1 - \lambda_i| = \lambda_i + \frac{3}{4}|\lambda_1 - \lambda_i|. \tag{1.22}$$

Por otro lado,

$$c_1 - r_1 = \lambda_1 - \epsilon\frac{|\beta_{11}|}{|s_1|} - \frac{\epsilon^2}{k|s_1|}\sum_{j\neq 1}|\beta_{1j}| = \lambda_1 - \left[\epsilon\frac{|\beta_{11}|}{|s_1|} + \frac{\epsilon^2}{k|s_1|}\sum_{j\neq 1}|\beta_{1j}|\right],$$

donde, bajo la condición impuesta a ϵ en el paso anterior, el término entre corchetes es menor o igual $|\lambda_1 - \lambda_i|/4$. Por lo tanto

$$c_1 - r_1 \geq \lambda_1 - \frac{1}{4}|\lambda_1 - \lambda_i| = \lambda_i + \frac{3}{4}|\lambda_1 - \lambda_i|. \tag{1.23}$$

Juntando (1.22) y (1.23) llegamos al resultado que queremos demostrar

$$c_i + r_i \leq \lambda_i + \frac{3}{4}|\lambda_1 - \lambda_i| \leq c_1 - r_1.$$

Por lo tanto, los discos no solapan.

□

Entonces, la condición impuesta sobre el valor de k queda

$$k \leq \frac{|s_i|}{2|\beta_{i1}|}|\lambda_1 - \lambda_i|, \quad i = 2, \ldots, n,$$

donde el factor $\frac{1}{2}$ se puede reemplazar por cualquier factor menor que 1. Es decir, basta tomar

$$|k| = \min_{2\leq i\leq n}\left(\left|\frac{(\lambda_1-\lambda_i)s_i}{2\beta_{i1}}\right|\right),$$

para que el primer disco de Gerschgorin sea tan pequeño como sea posible manteniendo el resto de discos sin solapar con el primero.

Perturbación del correspondiente autovector

Pasamos a estudiar ahora el cambio en el autovector $\mathbf{x}_1$ de A, asociado al autovalor simple λ_1. Denotamos por $\{\mathbf{x}_i\}_{i=1}^n$ a las columnas de la matriz H dada por (1.18), es decir, a una base de autovectores de A. Si $\{\mathbf{x}_i(\epsilon)\}_{i=1}^n$ son los correspondientes autovectores de $A+\epsilon B$ y $\{\mathbf{z}_i(\epsilon)\}_{i=1}^n$ son los autovectores de $H^{-1}(A+\epsilon B)H$, es claro que

$$\mathbf{x}_i(\epsilon) = H\mathbf{z}_i(\epsilon), \quad 1\leq i\leq n.$$

Si λ_1 es un autovalor simple de A, podemos probar el siguiente resultado.

Teorema 1.2.2 *Sea $\mathbf{z}_1(\epsilon)$ autovector de la matriz $H^{-1}(A+\epsilon B)H$ asociado al autovalor $\lambda_1(\epsilon)$. Entonces, si $\mathbf{z}_1(\epsilon)$ está normalizado en norma $\|\cdot\|_\infty$, esto es, si $\|\mathbf{z}_1(\epsilon)\|_\infty = 1$, se tiene que*

$$z_{11}(\epsilon) = 1. \tag{1.24}$$

Demostración.

Vamos a razonar por reducción al absurdo. Supongamos que existe una componente del vector normalizado $\mathbf{z}_1(\epsilon)$, distinta de la primera, para la que

$$z_{1j}(\epsilon) = 1, \quad \text{con } j\neq 1. \tag{1.25}$$

La ecuación que satisface $\mathbf{z}_1(\epsilon)$ es

$$H^{-1}(A+\epsilon B)H\mathbf{z}_1(\epsilon) = \lambda_1(\epsilon)\mathbf{z}_1(\epsilon). \tag{1.26}$$

Sustituyendo (1.25) en la ecuación resultante de (1.26) para la componente j-ésima, obtenemos que

$$\lambda_1(\epsilon) = \lambda_j + \frac{\epsilon}{s_j}\sum_{i=1}^n \beta_{ji}z_{1i}(\epsilon),$$

de donde se deduce que

$$\lambda_1(\epsilon) - \lambda_j \to 0 \text{ cuando } \epsilon\to 0,$$

lo que es absurdo, pues ya hemos probado que $\lambda_1(\epsilon)\to\lambda_1\neq\lambda_j$ para $j\neq 1$, por ser λ_1 autovalor simple de A. □

Teorema 1.2.3 *Bajo las condiciones del Teorema 1.2.2, existe una constante $K>0$ tal que las componentes del vector $\mathbf{z}_1(\epsilon)$ verifican*

$$z_{1j}(\epsilon) < K\cdot\epsilon \quad (\epsilon\to 0), j\neq 1.$$

Demostración.

La ecuación resultante de (1.26) para la componente j-ésima es ahora

$$\lambda_1(\epsilon)z_{1j}(\epsilon) = \lambda_j z_{1j}(\epsilon) + \frac{\epsilon}{s_j}\sum_{i=1}^{n}\beta_{ji}z_{1i}(\epsilon). \tag{1.27}$$

Por tanto, utilizando que ahora sabemos que $|z_{1i}(\epsilon)| < 1$ para $i \neq 1$, llegamos a que

$$|\lambda_1(\epsilon) - \lambda_j||z_{1j}(\epsilon)| \leq \frac{\epsilon}{|s_j|}\sum_{i=1}^{n}|\beta_{ji}||z_{1i}(\epsilon)| \leq \frac{\epsilon}{|s_j|}\sum_{i=1}^{n}|\beta_{ji}|.$$

Entonces, dado ϵ suficientemente pequeño para que $|\lambda_1(\epsilon) - \lambda_j| \geq \frac{1}{2}|\lambda_1 - \lambda_j|$, se tiene

$$|z_{1j}(\epsilon)| \leq \frac{2\sum_{i=1}^{n}|\beta_{ji}|}{|s_j||\lambda_1 - \lambda_j|}\,\epsilon, \quad j \neq 1,$$

y basta tomar

$$K = \frac{2\sum_{i=1}^{n}|\beta_{ji}|}{|s_j||\lambda_1 - \lambda_j|}.$$

□

Ahora, si llevamos el resultado que acabamos de probar a la relación (1.27), llegamos a que

$$(\lambda_1(\epsilon) - \lambda_j)z_{1j}(\epsilon) = \frac{\epsilon\beta_{j1}}{s_j} + O(\epsilon^2), \quad j \neq 1.$$

Sustituyendo $\lambda_1(\epsilon) = \lambda_1 + O(\epsilon)$, obtenemos

$$\left|z_{1j}(\epsilon) - \frac{\epsilon\beta_{j1}}{s_j(\lambda_1 - \lambda_j)}\right| = O(\epsilon^2), \quad j \neq 1.$$

De su definición, $\mathbf{x}_1(\epsilon) = H\mathbf{z}_1(\epsilon)$, y de (1.24) sabemos que $z_{11}(\epsilon) = 1$. Entonces,

$$\mathbf{x}_1(\epsilon) = \sum_{j=1}^{n} z_{1j}(\epsilon)\mathbf{x}_j = \mathbf{x}_1 + \epsilon\sum_{j=2}^{n}\frac{\beta_{j1}}{s_j(\lambda_1 - \lambda_j)}\mathbf{x}_j + O(\epsilon^2),$$

un resultado análogo a (1.16).

1.2.2. Perturbación de un autovalor múltiple de una matriz diagonalizable.

Manteniendo la matriz A diagonalizable, pasamos a estudiar el problema de perturbación para un autovalor múltiple de A. Suponemos que λ_1 es un autovalor de A de multiplicidad m. Sin pérdida de generalidad podemos suponer que los m autovalores λ_1 son los primeros, de forma que la ecuación (1.19) se transforma en

$$H^{-1}(A + \epsilon B)H = \mathrm{diag}(\lambda_1, \ldots, \lambda_1, \lambda_{m+1}, \ldots, \lambda_n) + \epsilon\begin{pmatrix} \frac{\beta_{11}}{s_1} & \frac{\beta_{12}}{s_1} & \cdots & \frac{\beta_{1n}}{s_1} \\ \frac{\beta_{21}}{s_2} & \frac{\beta_{22}}{s_2} & \cdots & \frac{\beta_{2n}}{s_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\beta_{n1}}{s_n} & \frac{\beta_{n2}}{s_n} & \cdots & \frac{\beta_{nn}}{s_n} \end{pmatrix}, \tag{1.28}$$

donde no imponemos ninguna condición sobre la multiplicidad de los autovalores restantes.

En primer lugar, el caso de los autovalores simples se puede estudiar aplicando la teoría desarrollada en la Sección 1.2.1. Sea por ejemplo λ_i, $m+1 \leq i \leq n$, un autovalor simple de la matriz A. Multiplicando la i-ésima fila y columna por $\frac{\epsilon}{k}$ y por $\frac{k}{\epsilon}$ respectivamente, para un valor apropiado del parámetro k, podemos localizar un autovalor simple en un disco de centro $\lambda_i + \epsilon\dfrac{\beta_{ii}}{s_i}$ y radio $O(\epsilon^2)$ cuando $\epsilon \to 0$.

Consideramos ahora los m autovalores de la matriz perturbada $A + \epsilon B$ asociados al autovalor múltiple λ_1 de A, que están en discos con centro $\lambda_1 + \epsilon\dfrac{\beta_{ii}}{s_i}$ para $i = 1, \ldots, m$. Los correspondientes radios son de orden $O(\epsilon)$.

Para ϵ suficientemente pequeño, la unión de discos $\mathcal{R}_1$ formada por todos los discos de Gerschgorin asociados a las primeras m filas de la matriz (1.28) estará separada del resto de discos, pero no podemos asegurar, en general, que los discos individuales de la unión estén aislados unos de otros.

Para entender mejor la última afirmación, partimos, por ejemplo de λ_1 que es un autovalor de multiplicidad m de la matriz original A. Para la matriz perturbada $A + \epsilon B$ obtenemos los m autovalores perturbados correspondientes $\lambda_{11}(\epsilon), \ldots, \lambda_{1m}(\epsilon)$ de manera que

$$|\lambda_{1i}(\epsilon) - \lambda_1| = O(\epsilon), \quad \epsilon \to 0,\ i = 1, \ldots, m,$$

pero no es cierto, en general, que haya un autovalor en cada uno de los m discos con centro $\lambda_1 + \epsilon\frac{\beta_{ii}}{s_i}$ para $i = 1, \ldots, m$ y radio de orden $O(\epsilon^2)$.

No obstante, sí podemos reducir el radio de dichos discos hasta cierto punto. Nos centramos en el autovalor múltiple λ_1 y reescribimos (1.28) como

$$H^{-1}(A + \epsilon B)H = \operatorname{diag}(\lambda_1, \ldots, \lambda_1, \lambda_{m+1}, \ldots \lambda_n) + \epsilon \begin{pmatrix} P & Q \\ R & S \end{pmatrix},$$

donde la submatriz denotada por P es cuadrada de orden m. Multiplicando las m primeras filas por $\dfrac{\epsilon}{k}$ y las m primeras columnas por $\dfrac{k}{\epsilon,}$ obtenemos

$$H^{-1}(A + \epsilon B)H = \operatorname{diag}(\lambda_1, \ldots, \lambda_1, \lambda_{m+1}, \ldots \lambda_n) + \begin{pmatrix} \epsilon P & \frac{\epsilon^2}{k} Q \\ kR & \epsilon S \end{pmatrix}.$$

De esta forma, podemos tomar un valor de k independiente de ϵ, de forma que los m primeros discos estén aislados del resto. Así, para tal valor de k, hay m autovalores en la unión de los discos

$$\mathcal{R} = \bigcup_{j=1}^{m} D(c_{1j}, r_j),$$

donde para cada $j = 1, \ldots, m$ el centro y el radio de cada disco vienen dados por

$$c_{1j} = \lambda_1 + \epsilon\frac{\beta_{jj}}{s_j};\ r_j = \frac{\epsilon}{|s_j|} \sum_{k=1;k\neq j}^{m} |\beta_{jk}| + \frac{\epsilon^2}{k|s_j|} \sum_{k=m+1;k\neq j}^{n} |\beta_{jk}|.$$

Perturbación del correspondiente autovector

Sea λ_1 un autovalor de multiplicidad m de la matriz A. Procediendo como en la Sección 1.2.1 podemos probar que si $\mathbf{z}_1(\epsilon)$ es el autovector, normalizado en norma $\|\cdot\|_\infty$, de la matriz $H^1(A+\epsilon B)H$ asociado al autovalor $\lambda_{11}(\epsilon)$, entonces

$$z_{1j}(\epsilon) < 1, \quad \text{para} \quad j > m.$$

Por lo tanto, el vector $\mathbf{z}_1(\epsilon)$ debe tener alguna de las siguientes formas:

$$\begin{gathered}[1, K(\epsilon), \ldots, K(\epsilon), z_{1m+1}(\epsilon), \ldots, z_{1n}(\epsilon)], \\ \vdots \\ [K(\epsilon), \ldots, K(\epsilon), 1, z_{1m+1}(\epsilon), \ldots, z_{1n}(\epsilon)],\end{gathered}$$

donde $|K(\epsilon)| \leq 1$. El correspondiente autovector $\mathbf{x}_1(\epsilon)$ de $A+\epsilon B$ tiene componentes de orden 1 en ϵ en las direcciones $\{\mathbf{x}_{m+1}, \ldots, \mathbf{x}_n\}$. Pero las componentes en el subespacio generado por $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ pueden ser

$$\begin{gathered}\mathbf{x}_1 + K(\epsilon)\mathbf{x}_2 + \cdots + K(\epsilon)\mathbf{x}_m, \\ \vdots \\ K(\epsilon)\mathbf{x}_1 + K(\epsilon)\mathbf{x}_2 + \cdots + \mathbf{x}_m.\end{gathered}$$

No podemos mejorar este resultado, ya que ningún vector de este subespacio es realmente autovector de A.

1.2.3. Matrices no diagonalizables

El siguiente paso lógico es estudiar la perturbación de matrices no diagonalizables. Si bien no lo haremos de forma general en este trabajo, vamos a estudiar un par de ejemplos sencillos pero ilustrativos que, además, dan cuenta de las dificultades de la teoría desarrollada hasta aquí.

Sea A la matriz 2×2 dada por

$$A = \begin{pmatrix} a & 1 \\ 0 & b \end{pmatrix}, \quad a,\, b \in \mathbb{R}.$$

A es una matriz triangular superior que, en el caso $a \neq b$ es diagonalizable y cuyos autovalores son $\{\lambda_1, \lambda_2\} = \{a, b\}$. Los autovectores asociados por la derecha y por la izquierda normalizados son

$$\begin{aligned}\mathbf{x}_1^T &= (1, 0), & \alpha\mathbf{x}_2^T &= (1, b-a), \\ \alpha\mathbf{y}_1^T &= (a-b, 1), & \mathbf{y}_2^T &= (0, 1),\end{aligned}$$

donde $\alpha = \sqrt{1+(b-a)^2}$ es el factor de normalización. Tenemos entonces que

$$s_1 = \mathbf{y}_1^T\mathbf{x}_1 = \frac{a-b}{\alpha}, \qquad s_2 = \mathbf{y}_2^T\mathbf{x}_2 = \frac{b-a}{\alpha}. \tag{1.29}$$

Para $a \neq b$, los autovalores a y b son simples, y de la Sección 1.2.1 sabemos que la sensibilidad de los autovalores es proporcional a los valores $|s_i^{-1}|$, $i = 1, 2$. Por lo tanto, cuando hacemos tender b hacia a, los autovalores son cada vez más sensibles a perturbaciones de la matriz A.

No obstante, de (1.29) vemos que $s_1^{-1} + s_2^{-1} = 0$. Por lo tanto, aunque ambos inversos tienden a infinito cuando b tiende a a, no son independientes.

Consideramos ahora la matriz A de dimensión 2×2 dada por

$$A = \begin{pmatrix} a & 10^{-10} \\ 0 & a \end{pmatrix}, \quad a \in \mathbb{R}. \tag{1.30}$$

La matriz A posee un autovalor doble dado por $\lambda = a$ y no es diagonalizable. Si pensamos en una perturbación ϵ del elemento $(2, 1)$, entonces los autovalores de la matriz perturbada son $a \pm \sqrt{10^{-10}\,\epsilon}$.

Es importante notar con ello que, aunque los autovalores de una matriz A son funciones continuas de sus coeficientes, estos no son necesariamente funciones diferenciables de los elementos de A. En este caso, tenemos que

$$\frac{d\lambda}{d\epsilon} = \pm\sqrt{\frac{10^{-10}}{\epsilon}},$$

que no está definida para $\epsilon = 0$. Vemos que una perturbación de tamaño $O(\epsilon)$ en el coeficiente $(2, 1)$ de la matriz genera un cambio de tamaño $O(\epsilon^{\frac{1}{2}})$ en los autovalores. En general, una perturbación de orden $O(\epsilon)$ produce un cambio de orden $O(\epsilon^{\frac{1}{p}})$ en los autovalores asociados con un bloque de Jordan de dimensiones $p \times p$ [9].

No obstante, en el ejemplo (1.30), para una perturbación arbitraria de la forma

$$B = \begin{pmatrix} \epsilon_{11} & \epsilon_{12} \\ \epsilon_{21} & \epsilon_{22} \end{pmatrix},$$

resulta más natural pensar en una perturbación

$$\tilde{B} = \begin{pmatrix} \epsilon_{11} & \epsilon_{12} + 10^{-10} \\ \epsilon_{21} & \epsilon_{22} \end{pmatrix},$$

de la matriz

$$\tilde{A} = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix},$$

que es diagonalizable y a la que podemos aplicar la teoría de la Sección 1.2.2.

1.3. Acondicionamiento del problema de autovalores

Si estamos interesados en un estudio completo del problema de perturbación para autovalores surge la necesidad de cuantificar de alguna manera la sensibilidad de una matriz a dichas perturbaciones mediante una o varias constantes que dependan de dicha matriz y que proporcionen cotas de error para el valor que recuperamos de λ cuando hay una perturbación en la matriz.

Tener una idea de la sensibilidad de la matriz para el problema de autovalores es importante, ya que en la práctica si tenemos una estimación del tamaño de la perturbación introducida, ϵ, esperaríamos que si $|\epsilon|$ es pequeño, $|\lambda(\epsilon) - \lambda|$ también lo sea. Sin embargo, si la sensibilidad de la matriz A es alta esto puede no ser cierto.

1.3.1. Condición espectral de una matriz respecto del problema de autovalores.

Antes de estudiar el número de condición para el problema de autovalores, recordamos la definición del número de condición espectral de una matriz.

Definición 1.3.1 *Dada una matriz H, se define su número de condición espectral como*

$$\kappa_2(H) = \|H\|_2 \|H^{-1}\|_2,$$

donde $\|\cdot\|_2$ denota la norma matricial derivada de la norma vectorial euclídea.

Sea A una matriz diagonalizable y H una matriz de autovectores de A, tal que

$$H^{-1}AH = \text{diag}(\lambda_1, \ldots, \lambda_n).$$

Vamos a demostrar que el número de condición euclídeo de la matriz de autovectores, $\kappa_2(H) = \|H^{-1}\|_2 \|H\|_2$, es una buena elección como número de condición para el problema de autovalores. Es importante mencionar que para que los resultados que vamos a obtener en esta sección sean ciertos, no estamos requiriendo que ϵ sea pequeño.

Teorema 1.3.1 *Sean $\{\lambda_1, \ldots, \lambda_n\}$ $i = 1, \ldots, n$ los autovalores de una matriz A diagonalizable, y sea λ un autovalor de la matriz perturbada $A + \epsilon B$. Entonces existe al menos un índice i para el que*

$$|\lambda_i - \lambda| \leq \epsilon \kappa_2(H) \|B\|_2. \tag{1.31}$$

Demostración.

Si λ es un autovalor de la matriz perturbada $A + \epsilon B$ entonces, por definición, la matriz $A + \epsilon B - \lambda I$ es singular y su determinante es nulo. Tomando determinantes en la siguiente igualdad

$$H^{-1}(A + \epsilon B - \lambda I)H = \text{diag}(\lambda_1 - \lambda, \ldots, \lambda_n - \lambda) + \epsilon H^{-1}BH,$$

llegamos a que

$$\begin{aligned} \det\left(\text{diag}(\lambda_1 - \lambda, \ldots, \lambda_n - \lambda) + \epsilon H^{-1}BH\right) &= \det\left(H^{-1}(A + \epsilon B - \lambda I)H\right) \\ &= \det\left(A + \epsilon B - \lambda I\right) = 0. \end{aligned}$$

Podemos distinguir dos casos:

1. Existe algún $i \in \{1, 2, \ldots, n\}$ tal que $\lambda = \lambda_i$. Entonces el resultado que queremos probar es cierto, pues el lado izquierdo de la igualdad (1.31) es nulo para ese índice i.

2. Se tiene que $\lambda \neq \lambda_i$ para todo $1 \leq i \leq n$, de manera que, en este caso, podemos escribir

$$\begin{aligned} &\operatorname{diag}(\lambda_1 - \lambda, \ldots, \lambda_n - \lambda) + \epsilon H^{-1} B H \\ &= \operatorname{diag}(\lambda_1 - \lambda, \ldots, \lambda_n - \lambda)[I + \epsilon \operatorname{diag}\left((\lambda_1 - \lambda)^{-1}, \ldots, (\lambda_n - \lambda)^{-1}\right) H^{-1} B H]. \end{aligned}$$

De nuevo tomamos determinantes a ambos lados de la última igualdad y obtenemos

$$\det\left(I + \epsilon \operatorname{diag}\left((\lambda_1 - \lambda)^{-1}, \ldots, (\lambda_n - \lambda)^{-1}\right) H^{-1} B H\right) = 0.$$

Ahora, vamos a utilizar que para una matriz general X sabemos que si $(I + X)$ es regular y $\|X\| < 1$, entonces

$$(I + X)^{-1} = \sum_{n=0}^{\infty} (-1)^n X^n. \tag{1.32}$$

Tomando normas en la expresión (1.32) se obtiene una serie geométrica de razón $||X||$. Por tanto, si $(I + X)$ es singular, la serie es divergente y debe ser $||X|| \geq 1$ para cualquier norma matricial derivada de una vectorial.

En particular, tomando la norma $|| \cdot ||_2$, llegamos a que

$$\|\epsilon \operatorname{diag}\left((\lambda_1 - \lambda)^{-1}, \ldots, (\lambda_n - \lambda)^{-1}\right) H^{-1} B H\|_2 \geq 1.$$

Utilizando la propiedad $\|A \cdot B\| \leq \|A\| \cdot \|B\|$ que cumple toda norma matricial derivada de una norma vectorial, obtenemos

$$\begin{aligned} 1 \leq \epsilon \, \|\operatorname{diag}\left((\lambda_1 - \lambda)^{-1}, \ldots, (\lambda_n - \lambda)^{-1}\right\|_2 \|H^{-1}\|_2 \|B\|_2 \|H\|_2 \\ = \epsilon \max \left| \frac{1}{\lambda_i - \lambda} \right| \kappa_2(H) \|B\|_2 \\ \iff \min |\lambda_i - \lambda| \leq \epsilon \, \kappa_2(H) \|B\|_2 \end{aligned} \tag{1.33}$$

donde $\kappa_2(H) = \|H^{-1}\|_2 \|H\|_2$.

Así, en cualquiera de los dos casos descritos podemos afirmar que existe al menos un valor de i para el que la desigualdad (1.31) es cierta, es decir, λ está en al menos un disco centrado en λ_i y de radio $\epsilon \, \kappa_2(H) \, \|B\|_2$.

□

Hemos demostrado hasta aquí que la sensibilidad global de los autovalores de la matriz A con respecto a la perturbación ϵB depende del tamaño de $\kappa_2(H)$, de modo que $\kappa_2(H)$ será considerado como un número de condición de A con respecto al problema de autovalores.

Cabe destacar que el resultado obtenido en la ecuación (1.33) es válido para cualquier norma matricial derivada de una norma vectorial que verifique

$$\|\operatorname{diag}\left((\lambda_1 - \lambda)^{-1}, \ldots, (\lambda_n - \lambda)^{-1}\right)\| = \text{máx} \left| \frac{1}{\lambda_i - \lambda} \right|.$$

En particular se cumple para las normas $\|\cdot\|_1$ y $\|\cdot\|_\infty$.

Utilizando el concepto de continuidad podemos localizar las raíces de forma más precisa con el mismo método usado para probar el segundo resultado del Teorema de Gerschgorin. Esto nos lleva a formular el siguiente resultado, que completa al Teorema 1.3.1.

Teorema 1.3.2 *Sean $\{\lambda_1, \ldots, \lambda_n\}$ los autovalores de una matriz A diagonalizable. Para cada $1 \leq i \leq n$, sea $D_i = D(\lambda_i, r)$, el disco cerrado del plano complejo centrado en λ_i de radio $r = \epsilon\, \kappa_2(H)\|B\|_2$, donde B es una matriz arbitraria. Entonces*

1. *Cada autovalor λ de $A + \epsilon B$ está en la unión $\mathcal{R} = \bigcup_{i=1}^{n} D_i$.*
2. *Si una componente conexa de dicha región $\mathcal{R}$ del plano complejo está formada por s discos, entonces contiene exactamente s autovalores de $A + \epsilon B$.*

1.3.2. Propiedades del número de condición $\kappa_2(H)$

A la hora de definir el número de condición de una matriz A con respecto al problema de autovalores como $\kappa_2(H) = \|H^{-1}\|_2\|H\|_2$, donde H es una matriz regular tal que $H^{-1}AH = \mathrm{diag}(\lambda_i)$, nos encontramos con un problema debido a la falta de unicidad en la elección de la matriz H. Incluso si los autovalores de A son distintos, cada columna de H puede multiplicarse por un factor arbitrario.

Para resolver este inconveniente vamos a dar la siguiente definición unívoca para el número de condición.

Definición 1.3.2 *Se define el número de condición espectral con respecto al problema de autovalores de una matriz A, como el valor más pequeño de $\kappa_2(H)$ para cualquier matriz H cuyas columnas sean una base de autovectores de la matriz A. Lo denotamos por κ^A.*

Lo primero que podemos observar directamente de esta definición es que, en cualquier caso,

$$\kappa_2(H) = \|H^{-1}\|_2\|H\|_2 \geq \|H^{-1}H\|_2 = 1.$$

Por definición, una matriz A es normal si conmuta con su traspuesta conjugada, es decir, si $A^* \cdot A = A \cdot A^*$. De manera equivalente, una matriz A es normal si, y solo si existen una matriz unitaria U y una matriz diagonal D tales que $A = U^* \cdot D \cdot U$. Este resultado lleva a que para matrices normales y, en particular, para matrices hermíticas y unitarias, tomando H unitaria se tiene

$$\kappa^A = 1.$$

Esta observación permite afirmar que el problema de autovalores está siempre bien acondicionado para matrices normales. No obstante, esto no es necesariamente cierto para el correspondiente problema de autovectores.

Recordamos que dado un autovalor λ_i de la matriz A, los autovectores normalizados por la derecha y por la izquierda asociados están dados respectivamente por

$$\mathbf{x}_i = \frac{H\mathbf{e}_i}{\|H\mathbf{e}_i\|_2} \quad , \quad \mathbf{y}_i = \frac{(H^{-1})^T\mathbf{e}_i}{\|(H^{-1})^T\mathbf{e}_i\|_2}, \quad 1 \leq i \leq n,$$

donde $\mathbf{e}_1, \ldots, \mathbf{e}_n$ denota la base ordenada canónica de $\mathbb{R}^n$.

Resulta intuitivo el hecho de que el número $\kappa_2(H)$ vaya a estar relacionado con las cantidades $|s_i^{-1}|$, $i = 1, \ldots, n$, definidas anteriormente como $|s_i^{-1}| = |\mathbf{y}_i^T \mathbf{x}_i|^{-1}$. A través de las siguientes propiedades vamos a concretar dichas relaciones y a deducir cómo la cantidad $|s_i^{-1}|$ gobierna la sensibilidad individual del autovalor λ_i.

Propiedades

1. Sea H una matriz de autovectores de A. Se verifica $|s_i^{-1}| \leq \kappa_2(H)$, para todo $1 \leq i \leq n$.

   Demostración.

   Comenzamos razonando a partir de la definición de $|s_i|$, en concreto

   $$|s_i| = |\mathbf{y}_i^T \cdot \mathbf{x}_i| = \frac{|\mathbf{e}_i^T H^{-1} H \mathbf{e}_i|}{\|H\mathbf{e}_i\|_2 \|(H^{-1})^T \mathbf{e}_i\|_2} = \frac{1}{\|H\mathbf{e}_i\|_2 \|(H^{-1})^T \mathbf{e}_i\|_2}.$$

   Por otra parte, sabemos que

   $$\|H\mathbf{e}_i\|_2 \leq \|H\|_2 \|\mathbf{e}_i\|_2 = \|H\|_2,$$

   y también

   $$\|(H^{-1})^T \mathbf{e}_i\|_2 \leq \|(H^{-1})^T\|_2 \|\mathbf{e}_i\|_2 = \|(H^{-1})^T\|_2 = \|H^{-1}\|_2.$$

   Juntando todo tenemos finalmente que

   $$|s_i^{-1}| = \|H\mathbf{e}_i\|_2 \|(H^{-1})^T \mathbf{e}_i\|_2 \leq \|H\|_2 \|H^{-1}\|_2 = \kappa_2(H).$$

   □

2. Dada una matriz A, se tiene que $\kappa^A \leq \sum_{i=1}^{n} |s_i^{-1}|$. Para demostrar esta propiedad del número de condición κ^A vamos a utilizar la elección de la matriz H tal que las columnas de H son $\frac{\mathbf{x}_i}{\sqrt{s_i}}$ y las filas de H^{-1} están dadas por $\frac{\mathbf{y}_i^T}{\sqrt{s_i}}$.

   Demostración.

   Para tal elección de H tenemos que

   $$\kappa_2(H) = \|H\|_2 \|H^{-1}\|_2 \leq \|H\|_F \|H^{-1}\|_F = \left(\sum_{i=1}^{n} |s_i^{-1}| \right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} |s_i^{-1}| \right)^{\frac{1}{2}} = \sum_{i=1}^{n} |s_i^{-1}|,$$

   donde $\|\cdot\|_F$ denota la norma de Frobenius, y hemos utilizado que los autovectores son unitarios por lo que

   $$\|H\|_F = \left(\sum_{i,j=1}^{n} |h_{i,j}|^2 \right)^{\frac{1}{2}} = \left(\sum_{j=1}^{n} \frac{\|\mathbf{x}_j\|_2^2}{|s_j|} \right)^{\frac{1}{2}} = \left(\sum_{j=1}^{n} |s_j^{-1}| \right)^{\frac{1}{2}},$$

   y lo mismo se tiene para $\|H^{-1}\|_F$ □

Tras probar estas dos propiedades, podemos utilizar las cantidades $|s_i^{-1}|,\ \ i = 1, \ldots, n$, como números de condición de A con respecto al problema de autovalores.

Cabe destacar que en la práctica, para obtener una aproximación del valor de κ^A necesitamos encontrar aproximaciones al conjunto de autovectores de A, pero una vez conocidas aproximaciones a los autovectores es más fácil hallar estimaciones de los números de condición s_i, que de κ^A.

1.3.3. Propiedades de invarianza de los números de condición.

A continuación vamos a probar algunas propiedades de los n números de condición $|s_i^{-1}|$ así como de κ^A respecto de transformaciones unitarias de semejanza aplicadas a la matriz A.

1. Sean A y B dos matrices unitariamente semejantes. Entonces $\kappa^A = \kappa^B$.

   Demostración.

   Sea una matriz A, tal que $H^{-1}AH = \text{diag}(\lambda_i)$ y cuyo número de condición con respecto al problema de autovalores κ^A viene dado por $\kappa^A = \kappa_2(H) = \|H^{-1}\|_2\|H\|_2$.

   Si B es unitariamente semejante con A, existe U unitaria tal que $B = UAU^*$ (U^* es la matriz traspuesta conjugada de U). Entonces

   $$H^{-1}AH = H^{-1}U^*B\ U\ H = \text{diag}(\lambda_i)$$
   $$\Rightarrow (UH)^{-1}B(UH) = \text{diag}(\lambda_i),$$

   luego el número de condición κ^B de B cumple

   $$\kappa^B \leq \|(UH)^{-1}\|_2\|UH\|_2 = \|H^{-1}\|_2\|H\|_2 = \kappa^A,$$

   ya que por ser U una matriz unitaria, $\|U\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ para todo $\mathbf{x}$. Razonando de forma análoga intercambiando los papeles de A y B llegamos también a que $\kappa^A \leq \kappa^B$ y, por tanto, ambos números de condición tienen que coincidir.

   □

2. Sean A y B dos matrices unitariamente semejantes. Entonces, para cada $i = 1, \ldots, n$, se tiene $s_i' = s_i$, donde s_i' denota el correspondiente número de condición de la matriz B.

   Demostración.

   Si $\mathbf{x}_i$ e $\mathbf{y}_i$, $1 \leq i \leq n$, son autovectores unitarios por la derecha y por la izquierda, respectivamente, de la matriz A, entonces los respectivos autovectores de B están dados por

   $$\mathbf{x}_i' = U\mathbf{x}_i \ \text{ e } \ \mathbf{y}_i' = U\mathbf{y}_i,$$

   ya que

   $$A\mathbf{x}_i = \lambda_i\mathbf{x}_i \iff U^*BU\mathbf{x}_i = \lambda_i\mathbf{x}_i \iff BU\mathbf{x}_i = \lambda_iU\mathbf{x}_i.$$

   Por lo tanto, podemos concluir que

   $$s_i' = (\mathbf{y}_i')^T \cdot (\mathbf{x}_i') = \mathbf{y}_i^TU^*U\mathbf{x}_i = \mathbf{y}_i^T\mathbf{x}_i = s_i.$$

   □

1.3.4. Algunos ejemplos de matrices mal acondicionadas para el problema de autovalores

Mostramos en esta sección dos ejemplos de familias de matrices que están mal acondicionadas para el problema de autovalores, ilustrando algunos de los conceptos introducidos en las secciones anteriores.

Ejemplo 1.2.

Vamos a ver en primer lugar un ejemplo ilustrativo de las dificultades que pueden surgir en la práctica en el estudio del problema de perturbación para autovalores. En concreto, el siguiente ejemplo muestra que incluso si los autovalores de la matriz A son distintos y están bien separados, pueden estar muy mal acondicionados.

Sea A_n la matriz cuadrada de orden n definida por

$$A_n = \begin{pmatrix} n & n & 0 & \cdots & 0 & 0 \\ 0 & n-1 & n & \cdots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & n \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{1.34}$$

A_n es una matriz triangular superior, por lo que sus autovalores son los elementos de la diagonal, es decir, los números naturales $1, 2, \ldots, n$, todos ellos autovalores simples. Si se considera el caso $n = 20$ (la matriz A_{20} resultante se conoce como la *matriz de Wilkinson*) y si se modifica el elemento de la posición $(20, 1)$ pasando de 0 a ϵ, la ecuación característica se convierte en

$$\begin{aligned} \det(A_{20} + \epsilon B - \lambda I) =& (20-\lambda) \begin{vmatrix} 19-\lambda & 20 & \cdots & 0 & 0 \\ 0 & 18-\lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 2-\lambda & 20 \\ 0 & 0 & \cdots & 0 & 1-\lambda \end{vmatrix} \\ & - \epsilon \begin{vmatrix} 20 & 0 & \cdots & 0 & 0 \\ 19-\lambda & 20 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 20 & 0 \\ 0 & 0 & \cdots & 2-\lambda & 20 \end{vmatrix} = 0, \end{aligned}$$

es decir,

$$(20-\lambda)(19-\lambda)\cdot\cdots\cdot(1-\lambda) - \epsilon 20^{19} = 0, \tag{1.35}$$

donde (1.35) se ha obtenido desarrollando el determinante $\det(A_{20} + \epsilon B - \lambda I)$ por la primera columna.

Hemos visto en la Sección 1.1.1, que para ϵ suficientemente pequeño, la perturbación del autovalor $\lambda = r$ para $r = 1, 2, \ldots, 20$, se puede desarrollar en serie de potencias de ϵ,

como en (1.5), de la siguiente manera

$$\lambda_r(\epsilon) = r + k_r\epsilon + O(\epsilon^2), \quad \text{con} \quad |k_r| \leq \frac{20}{|s_r|},$$

donde la serie converge y podemos tomar $\lambda_r(\epsilon) \approx r + k_r\epsilon$.

Entonces, si escribimos para cada $r = 1, 2, \ldots, 20$, que $\lambda_r(\epsilon) - r \approx k_r\epsilon$, de la ecuación característica (1.35) vamos a poder obtener el valor de k_r. Para ello comenzamos observando que para cada r, por ser $\lambda_r(\epsilon)$ autovalor de $A + \epsilon B$, $\lambda_r(\epsilon)$ es raíz del polinomio característico (1.35). Sustituyendo $\lambda_r(\epsilon) = r + k_r\epsilon + O(\epsilon^2)$ en (1.35), tenemos que

$$(20 - r - k_r\epsilon)(19 - r - k_r\epsilon)\cdots\underbrace{(r - r - k_r\epsilon)}_{-k_r\epsilon}\cdots(1 - r - k_r\epsilon) + O(\epsilon^2) = \epsilon 20^{19}, \quad (1.36)$$

donde notamos que en el primer miembro de la igualdad (1.36) aparece el factor $-k_r\epsilon$, es decir,

$$-k_r\epsilon[\underbrace{(20 - r)(19 - r)\cdots(r + 1 - r)}_{(20-r)!}\underbrace{(r - 1 - r)\cdots(1 - r)}_{(-1)^{r+1}(r-1)!}] + O(\epsilon^2) = \epsilon 20^{19}. \quad (1.37)$$

Igualando en (1.37) los términos en ϵ y despejando k_r se obtiene

$$k_r = 20^{19}\frac{(-1)^r}{(20 - r)!(r - 1)!}.$$

La constante k_r que acabamos de obtener es grande para todos los valores de r, es decir, todos los autovalores de la matriz de Wilkinson están mal acondicionados. Los valores más pequeños (denominador más grande) se dan para $|k_1|$ y $|k_{20}|$, mientras que los mayores se dan para $|k_{10}|$ y $|k_{11}|$, como se observa en la Tabla 1.1.

Autovalor (r)	Número de condición $\lvert k_r\rvert$
1, 20	$4,31 \times 10^7$
2, 19	$8,19 \times 10^8$
3, 18	$7,37 \times 10^9$
4, 17	$4,18 \times 10^{10}$
5, 16	$1,67 \times 10^{11}$
6, 15	$5,01 \times 10^{11}$
7, 14	$1,17 \times 10^{12}$
8, 13	$2,17 \times 10^{12}$
9, 12	$3,26 \times 10^{12}$
10, 11	$3,98 \times 10^{12}$

Tabla 1.1: Valores del número de condición $k_r, r = 1, \ldots, 20$, de los autovalores de la matriz (1.34) para $n = 20$.

Observamos con esto que los factores $|k_r|$ obtenidos son tan grandes que la teoría lineal sólo puede aplicarse para valores muy pequeños de ϵ. De hecho, en la Tabla 1.2 vemos claramente que para $\epsilon = 10^{-10}$ y $n = 20$, solo da información significativa para el valor de $\lambda_r(\epsilon)$ para $r = 1, 2, 19, 20$.

Autovalor (r)	$\lvert k_r\rvert\epsilon$	$\lvert r-\lambda_r(\epsilon)\rvert$
1, 20	0,00431	0,00424
2, 19	0,082	0,109
3, 18	0,734	0,425
4, 17	4,176	1,088
5, 16	16,705	1,501
6, 15	50,116	1,951
7, 14	116,938	2,241
8, 13	217,171	2,532
9, 12	325,757	2,679
10, 11	398,147	2,779

Tabla 1.2: Comprobación de la validez de la aproximación de primer orden $\lambda_r(\epsilon) \approx r + k_r\epsilon$, para los autovalores de la matriz (1.34) con $n = 20$ y $\epsilon = 10^{-10}$.

Para $\epsilon = 10^{-n}$, $\quad n = 5, 6, \ldots, 10$, los autovalores de la matriz perturbada $A + \epsilon B$ se han obtenido utilizando la función `eig` de Matlab y están representados en la Figura 1.3. Los autovalores de la matriz A_{20} están dados por puntos rojos, mientras que los autovalores de la matriz perturbada van desde los puntos color azul oscuro para $\epsilon = 10^{-5}$, hasta los puntos de color cian para $\epsilon = 10^{-10}$. En la gráfica se observa la simetría de los autovalores de A en torno a 10,5. Además notamos que, mientras que los autovalores de A_{20} son los naturales $1, 2, \ldots, 20$, las matrices perturbadas tienen autovalores complejos.

La sensibilidad de los autovalores de la matriz A_{20} muestra que los valores $|s_i|$ deben ser pequeños. De hecho, el autovector $\mathbf{x}_r$ de A_{20} correspondiente a $\lambda = r$, para $r = 1, 2, \ldots, 20$, tiene componentes

$$\mathbf{x}_r = \left[1, \frac{20-r}{-20}, \frac{(20-r)(19-r)}{(-20)^2}, \ldots, \frac{(20-r)!}{(-20)^{20-r}}, 0, \ldots, 0\right]^T,$$

mientras que el correspondiente autovector por la izquierda es

$$\mathbf{y}_r = \left[0, \ldots 0, \frac{(r-1)!}{20^{r-1}}, \ldots, \frac{(r-1)(r-2)}{20^2}, \frac{r-1}{20}, 1\right]^T.$$

Estos vectores no están normalizados en norma $\|\cdot\|_2$, pero el producto $\mathbf{y}_r^T\mathbf{x}_r$ da una buena estimación de la magnitud de $|s_r|$, que viene dada por

$$|\mathbf{y}_r^T\mathbf{x}_r| = \frac{(20-r)!(r-1)!}{20^{19}},$$

cuyo inverso es precisamente $|k_r|$, tal como hemos calculado antes. Observamos que los correspondientes autovectores de A y de A^T son casi ortogonales.

Otra forma de visualizar el distinto acondicionamiento de los autovalores de la matriz A_{20} por medio de perturbaciones en el elemento $(20, 1)$ es tener en cuenta que para una perturbación ϵ, los autovalores son las raíces de la ecuación no lineal $f(\lambda) = 20^{19}\epsilon$, con

$$f(\lambda) = (20-\lambda)(19-\lambda)\cdots(1-\lambda).$$

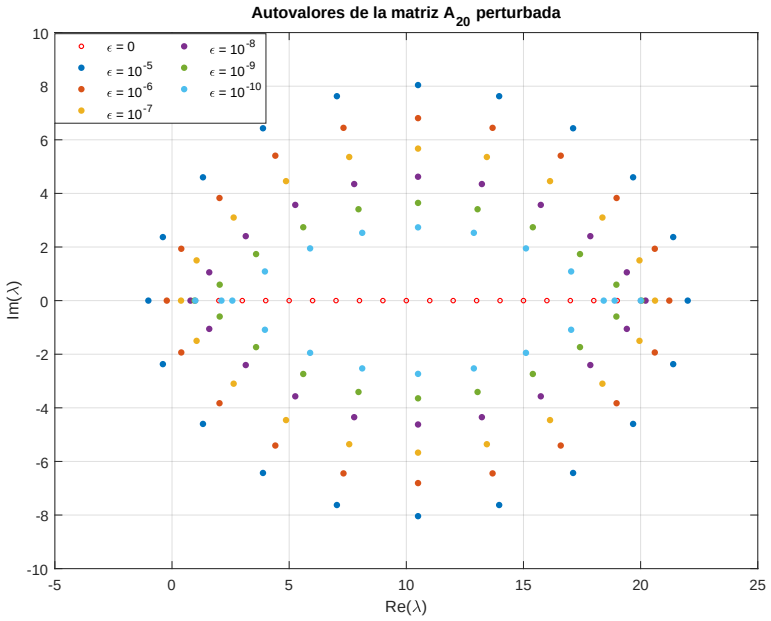

Figura 1.3: Autovalores $\lambda_r(\epsilon)$ de la matriz perturbada obtenida a partir de (1.34) con $n = 20$ para $\epsilon = 10^{-5}, 10^{-6}, \ldots, 10^{-10}$.

Si representamos gráficamente $f(\lambda)$, observamos que $f(\lambda)$ es simétrica en torno a $\lambda = 10{,}5$ y que presenta 9 máximos y 10 mínimos. Si vamos incrementando ϵ empezando desde 0 entonces las raíces 10 y 11 se acercan entre sí y coinciden en 10,5 cuando

$$\left(\frac{1}{2} \cdot \frac{3}{2} \cdots \frac{19}{2}\right)^2 = 20^{19}\epsilon,$$

lo que corresponde a un valor de aproximadamente $\epsilon \approx 7{,}8 \times 10^{-14}$.

Si seguimos incrementando ϵ llegamos a hacer coincidir las raíces 6 y 7, así como las raíces 14 y 15. Y podemos continuar así sucesivamente.

En la Figura 1.4 podemos visualizar la misma situación para el caso más sencillo en que la matriz A_n dada por (1.34) es de orden $n = 4$. En ella representamos $f(\lambda) - 4^3\epsilon$ para distintos valores de ϵ y observamos cómo van cambiando los puntos de corte con el eje horizontal.

De esta manera concluimos que todos los autovalores de la matriz que hemos considerado están mal acondicionados, aunque algunos están peor acondicionados que otros.

Ejemplo 1.3.

Damos ahora un ejemplo de un conjunto de matrices que tienen autovalores con números de condición muy distintos. Las matrices tienen forma de Hessenberg superior

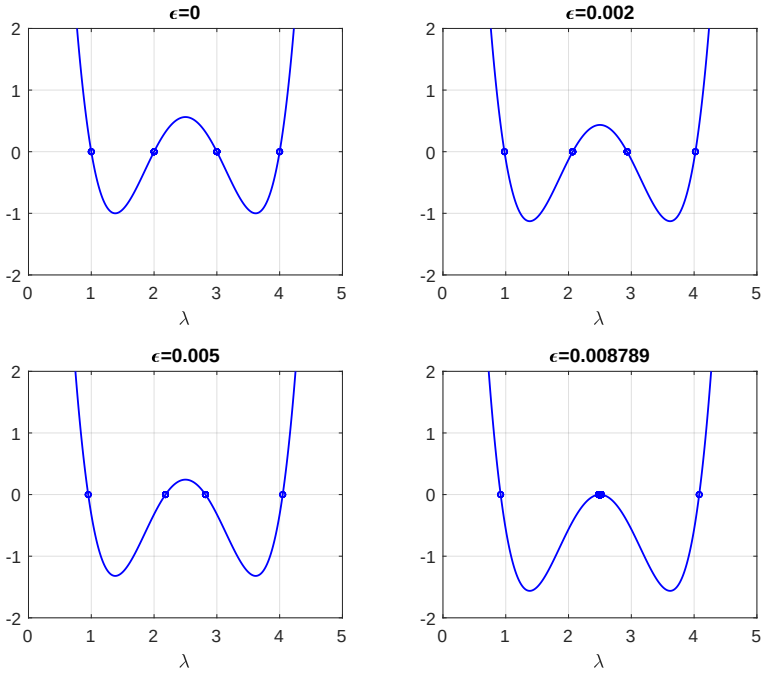

Figura 1.4: Raíces de la función $f(\lambda) - 4^3\epsilon$ para distintos valores de ϵ.

y están definidas por

$$C_n = \begin{pmatrix} n & n-1 & n-2 & \cdots & 3 & 2 & 1 \\ n-1 & n-1 & n-2 & \cdots & 3 & 2 & 1 \\ 0 & n-2 & n-2 & \cdots & 3 & 2 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 2 & 2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 1 \end{pmatrix}. \tag{1.38}$$

Todas las matrices de la forma (1.38) tienen determinante igual a 1 porque

$$\det(C_n) \underset{f_1 - f_2 \to f_1}{=} \begin{vmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ n-1 & n-1 & n-2 & \cdots & 3 & 2 & 1 \\ 0 & n-2 & n-2 & \cdots & 3 & 2 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 2 & 2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 1 \end{vmatrix} = 1 \cdot \det(C_{n-1}), \tag{1.39}$$

y a partir de (1.39) el resultado es evidente razonando por inducción.

Si en C_n sustituimos el primer elemento de la columna n-ésima que vale 1, por $1+\epsilon$,

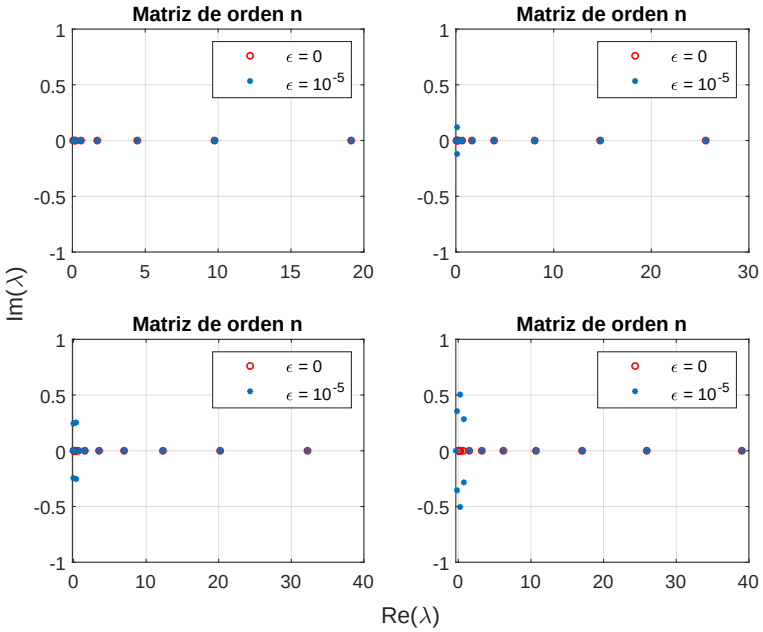

Figura 1.5: Acondicionamiento de los autovalores de matrices (1.38) para $n = 8, 10, 12$ y 14.

entonces el determinante pasa a ser

$$\det(C_n + \epsilon B) = \begin{vmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & \epsilon \\ n-1 & n-1 & n-2 & \cdots & 3 & 2 & 1 \\ 0 & n-2 & n-2 & \cdots & 3 & 2 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 2 & 2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 1 \end{vmatrix} =$$

$$= 1 \cdot \det(C_{n-1}) + (-1)^{n+1}\epsilon \begin{vmatrix} n-1 & n-1 & n-2 & \cdots & 3 & 2 \\ 0 & n-2 & n-2 & \cdots & 3 & 2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 2 & 2 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{vmatrix} =$$

$$= 1 + (-1)^{n+1}\epsilon(n-1)!.$$

Entonces, si tomamos, por ejemplo, $\epsilon = 10^{-10}$ y $n = 20$, tenemos que

- $\det(C_{20}) = 1$,
- $\det(C_{20} + \epsilon B) = 1 - 19! \cdot 10^{-10} \approx -1{,}216 \times 10^7$.

Como el determinante de una matriz es igual al producto de sus autovalores, para $\epsilon \neq 0$ al menos un autovalor de la matriz perturbada debe ser distinto del correspondiente autovalor de la matriz original.

A la vista del cambio que se produce en el valor del determinante de la matriz tras introducir la perturbación de tamaño ϵ, parece esperable que los autovalores de C_n sean muy sensibles a pequeños cambios en algunos elementos de la matriz.

De nuevo, utilizamos la función `eig` de Matlab para realizar un experimento numérico que nos permita ilustrar el mal acondicionamiento de los autovalores de las matrices de la forma (1.38). En concreto, en la Figura 1.5, representamos en rojo los autovalores de la matriz C_n, para $n = 8, 10, 12$ y 14, y en azul los autovalores correspondientes a la matriz perturbada $C_n + \epsilon B$, con $\epsilon = 10^{-5}$.

Las principales propiedades de las matrices de la forma (1.38) respecto del problema de pertubación son

- Los autovalores más grandes de la matriz (1.38) están bien acondicionados, y los más pequeños están muy mal acondicionados.
- Para $n = 12$, los primeros valores de $|s_i|$ son del orden de 10^0, mientras que los tres últimos son del orden de 10^{-7}. Esto es, $|s_i^{-1}| \approx 10^7$, $i = 10, 11, 12$.
- Incrementando el valor de n, los autovalores pequeños se van convirtiendo progresivamente en peor acondicionados.

Un mal acondicionamiento del tipo que acabamos de analizar en estos ejemplos es más importante que el asociado directamente con matrices que no diagonalizan. Las matrices no diagonalizables son casi inexistentes en el trabajo práctico. Si los elementos de una tal matriz son irracionales, o no pueden ser representados con todas sus cifras decimales en el ordenador que se esté usando, los errores de redondeo harán que los autovalores múltiples dejen de serlo. Por tanto, aunque la mayor parte de los resultados que se han presentado en esta sección son para matrices que diagonalizan, pueden ser ampliamente utilizados en la práctica puesto que las que sí son comunes aunque diagonalicen son las matrices con $|s_i| \ll 1$.

1.4. Teoría de perturbación para matrices reales simétricas

La teoría de perturbación para el problema de autovalores aplicada a una matriz A simétrica es más sencilla que para matrices generales, pues los dos conjuntos de autovectores por la derecha y por la izquierda de A, $\{\mathbf{x}_i\}_{i=1}^n$ e $\{\mathbf{y}_i\}_{i=1}^n$, serán idénticos, incluso si la matriz A posee algún autovalor múltiple. Por tanto, en el caso de una matriz A simétrica, las cantidades $s_i = \mathbf{y}_i^T \mathbf{x}_i$ verifican

$$|s_i| = 1 \ , \quad 1 \leq i \leq n,$$

y como hemos visto en la Sección 1.3.2 el problema de autovalores está siempre bien acondicionado. Algunos de los resultados que vamos a probar a continuación pueden ser

extendidos de manera inmediata a toda matriz A normal, aunque otros lo harán solo a matrices hermíticas, o no podrán extenderse.

Vamos a distinguir el caso en que la perturbación de la matriz A no es simétrica, del caso en que sí lo es.

1.4.1. Perturbación no simétrica

Vamos a enunciar y demostrar sendos resultados análogos a los obtenidos en el caso general (Teorema 1.3.1 y Teorema 1.3.2), para el caso de perturbación no simétrica de una matriz A simétrica.

Teorema 1.4.1 *Sea A una matriz real simétrica y sean $\{\lambda_1, \ldots, \lambda_n\}$ los autovalores de A. Sea λ un autovalor de la matriz perturbada $A+\epsilon B$, donde B no es simétrica. Entonces existe al menos un índice i para el que*

$$|\lambda - \lambda_i| \leq \epsilon \, \|B\|_2.$$

Demostración.

Como la matriz A es simétrica es diagonalizable, y por el Teorema 1.3.1, sabemos que cada autovalor λ de la matriz perturbada $A + \epsilon B$ está, al menos, en un disco dado por

$$|\lambda - \lambda_i| \leq \epsilon \kappa_2(H) \|B\|_2,$$

siendo H una matriz de autovectores de A. Por ser A una matriz simétrica sabemos también que H puede tomarse ortogonal, y en ese caso $\kappa_2(H) = 1$. Entonces

$$|\lambda - \lambda_i| \leq \epsilon \, \|B\|_2 < \epsilon \, n,$$

si $|b_{i,j}| < 1, \quad 1 \leq i \leq n.$

□

Por el Teorema 1.3.2, sabemos además, que si un conjunto dado por s de estos discos forma un dominio conexo aislado del resto, entonces hay exactamente s autovalores λ de la matriz perturbada en dicho dominio. Estos resultados son obviamente ciertos no sólo para matrices simétricas, sino para toda matriz A normal.

Un caso que requiere especial atención es aquel en que la matriz A es real y simétrica, y la matriz de perturbación ϵB también es real. Para esta situación podemos demostrar el siguiente resultado.

Teorema 1.4.2 *Sea A una matriz real y simétrica tal que todos sus autovalores $\{\lambda_1, \ldots, \lambda_n\}$ están separados por más de $2n\tilde{\epsilon}$, para cierto $\tilde{\epsilon} > 0$, esto es,*

$$|\lambda_i - \lambda_j| > 2n\tilde{\epsilon} \;\; \textit{para todo } i \neq j. \tag{1.40}$$

Entonces, una perturbación cualquiera de tamaño ϵ con $0 < \epsilon < \tilde{\epsilon}$ en cada elemento individual de A, mantiene los autovalores reales y simples. Es decir, si $|b_{ij}| < 1$ para $1 \leq i, j \leq n$ y $0 < \epsilon < \tilde{\epsilon}$, los autovalores de $A+\epsilon B$ son reales y simples, y sus autovectores forman una base de $\mathbb{R}^n$.

Demostración.

Puesto que la matriz perturbada $A + \epsilon B$ es real, sus autovalores complejos, en caso de tenerlos, serán pares conjugados.

Si los autovalores $\{\lambda_i\}$ de la matriz A verifican (1.40), como por el Teorema 1.4.1 tenemos que para cualquier autovalor λ de la matriz perturbada $A+\epsilon B$, existe al menos un índice i tal que $|\lambda_i - \lambda| \leq n\epsilon$. Entonces,

$$|\lambda - \lambda_j| > n\epsilon > 0 \ (j \neq i). \tag{1.41}$$

Deducimos de esta forma que el i-ésimo disco $|\lambda_i - \lambda| \leq n\epsilon$, está aislado del resto y contiene un único autovalor λ de $A + \epsilon B$. Este autovalor será por tanto real, ya que si fuese complejo, estaría también su complejo conjugado en el disco.

□

1.4.2. Perturbación simétrica

La mayoría de las técnicas numéricas que se utilizan para el cálculo de autovalores de matrices simétricas están basadas en el uso de transformaciones de semejanza ortogonales. La simetría exacta se conserva en las sucesivas matrices transformadas calculando únicamente la parte triangular superior y tomando los elementos inferiores a la diagonal iguales que los superiores. Por esta razón, nos vamos a centrar en esta subsección en perturbaciones simétricas de matrices simétricas.

Cuando las perturbaciones son simétricas, la matriz perturbada necesariamente tiene autovalores y un sistema de autovectores reales, y lo mismo es cierto para la matriz de perturbación. Es natural buscar relaciones entre autovalores de la matriz original, la matriz de perturbación, y la matriz perturbada.

Gran parte de los resultados que hemos obtenido hasta ahora requieren que las perturbaciones sean pequeñas. Los resultados que vamos a probar en esta subsección no están sometidos a tales limitaciones.

Esto nos permite deshacernos del parámetro ϵ y escribir simplemente

$$A = B + C \ \text{ con } A,\ B,\ C \text{ matrices reales y simétricas.}$$

Empezamos analizando algunos casos particulares sencillos para pasar después al caso general y, por último, a resultados basados en el principio del minimax.

Matrices simétricas obtenidas orlando una matriz diagonal.

Dada una matriz diagonal $\text{diag}(\alpha_1, \ldots, \alpha_{n-1})$, sea X una matriz simétrica de la forma

$$X = \left(\begin{array}{c|c} \alpha & \mathbf{a}^T \\ \hline \mathbf{a} & \text{diag}(\alpha_i) \end{array}\right), \quad \alpha \in \mathbb{R}, \mathbf{a} = [a_1, \ldots, a_{n-1}]^T \in \mathbb{R}^{n-1}. \tag{1.42}$$

Nuestro objetivo es relacionar los autovalores de X con los elementos diagonales $\alpha_1, \ldots, \alpha_{n-1}$.

En primer lugar, notamos que si $a_j = 0$ entonces α_j es autovalor de X con autovector asociado $\mathbf{e}_{j+1} \in \mathbb{R}^n$.

Si $\mathbf{a} = \mathbf{0}$, los autovalores de X son $\{\alpha_i\}$, $1 \leq i \leq n-1$, y α.

Si $\mathbf{a} \neq \mathbf{0}$, supongamos que $\mathbf{a}$ tiene sólo s componentes no nulas ($s \geq 1$). Podemos elegir una matriz P de permutación adecuada, que modifique sólo las $n-1$ últimas filas, tal que $P([\alpha, a_1, \ldots, a_{n-1}]^T) = [\alpha, b_1, \ldots, b_s, 0, \ldots, 0]^T$, con $b_i \neq 0$, $1 \leq i \leq s$. De ese modo obtenemos una matriz Y ortogonalmente semejante a X dada por

$$Y = P^T X P = \left(\begin{array}{c|c|c} \alpha & \mathbf{b}^T & \mathbf{0}^T \\ \hline \mathbf{b} & \mathrm{diag}(\beta_i) & O \\ \hline \mathbf{0} & O & \mathrm{diag}(\gamma_i) \end{array}\right),$$

con $\mathrm{diag}(\beta_i) \in \mathcal{M}_{s\times s}$, $\mathrm{diag}(\gamma_i) \in \mathcal{M}_{n-1-s\times n-1-s}$, y donde los conjuntos $\{\beta_i\}$, $1 \leq i \leq s$, y $\{\gamma_i\}$, $1 \leq i \leq n-1-s$, son una reordenación de los $\{\alpha_i\}$, $1 \leq i \leq n-1$.

Notar que si la matriz $\mathrm{diag}(\alpha_i)$ tiene autovalores múltiples, estos podrían aparecer simultáneamente formando parte de $\mathrm{diag}(\beta_i)$ y de $\mathrm{diag}(\gamma_i)$. Los autovalores de X son, por tanto, los $\{\gamma_i\}$, $1 \leq i \leq n-1-s$, junto con los autovalores de la matriz Z de tamaño $(s+1)\times(s+1)$ dada por

$$Z = \left(\begin{array}{c|c} \alpha & \mathbf{b}^T \\ \hline \mathbf{b} & \mathrm{diag}(\beta_i) \end{array}\right). \tag{1.43}$$

Planteamos el polinomio característico de Z, y desarrollando el determinante por la primera columna se obtiene

$$\det(Z - \lambda I) = \left|\begin{array}{c|c} \alpha - \lambda & \mathbf{b}^T \\ \hline \mathbf{b} & \mathrm{diag}(\beta_i - \lambda) \end{array}\right| = (\alpha - \lambda)\prod_{i=1}^{s}(\beta_i - \lambda)$$

$$+ \sum_{i=1}^{s}(-1)^i \cdot b_i \cdot \left|\begin{array}{c|c|c} b_1 \quad \cdots \quad b_{i-1} & b_i & b_{i+1} \quad \cdots \quad b_s \\ \hline \mathrm{diag}(\beta_j - \lambda)_{j=1}^{i-1} & \mathbf{0} & \\ \hline & \mathbf{0} & \mathrm{diag}(\beta_j - \lambda)_{j=i+1}^{s} \end{array}\right| = 0.$$

Desarrollando ahora cada determinante del sumatorio por su columna i-ésima, que

solo tiene un elemento no nulo, queda

$$\begin{aligned}\det(Z-\lambda I) &= (\alpha-\lambda)\prod_{i=1}^{s}(\beta_i-\lambda)+\sum_{i=1}^{s}(-1)^i\cdot b_i\cdot(-1)^{i+1}\cdot b_i\prod_{j\neq i}(\beta_j-\lambda)= \\ &= (\alpha-\lambda)\prod_{i=1}^{s}(\beta_i-\lambda)-\sum_{i=1}^{s}b_i^2\prod_{j\neq i}(\beta_j-\lambda)=0.\end{aligned} \tag{1.44}$$

Supongamos que entre los $\{\beta_i\}$, hay sólo t valores distintos, y supongamos también, sin pérdida de generalidad, que son $\beta_1>\beta_2>\cdots>\beta_t$, y que estos tienen multiplicidades $r_1,r_2,\ldots,r_t$, respectivamente, de manera que

$$r_1+r_2+\cdots+r_t=s.$$

Entonces, la ecuación característica (1.44) es

$$(\alpha-\lambda)\prod_{i=1}^{t}(\beta_i-\lambda)^{r_i}-\sum_{i=1}^{t}c_i^2\left((\beta_i-\lambda)^{r_i-1}\prod_{j=1,j\neq i}^{t}(\beta_j-\lambda)^{r_j}\right)=0, \tag{1.45}$$

donde $c_i^2=\sum_j b_j^2$ es la suma de los r_i valores b_j^2 asociados con β_i, de manera que $c_i>0$.

El polinomio característico tiene un factor $\prod_{i=1}^{t}(\beta_i-\lambda)^{r_i-1}$, por lo que β_i es autovalor de Z con multiplicidad r_i-1. Simplificando la ecuación (1.45), dividiendo entre el factor $\prod_{i=1}^{t}(\beta_i-\lambda)^{r_i}$, los restantes autovalores de Z son las raíces de

$$\alpha-f(\lambda)\equiv(\alpha-\lambda)-\sum_{i=1}^{t}c_i^2(\beta_i-\lambda)^{-1}=0,$$

que podemos escribir como $\alpha-f(\lambda)=0$ para

$$f(\lambda)=\lambda+\sum_{i=1}^{t}c_i^2(\beta_i-\lambda)^{-1}, \tag{1.46}$$

que tendrá una gráfica como la representada en la Figura 1.6, que se ha obtenido para un caso particular en que $c_i=1$ para todo $1\leq i\leq t$, y β_i recorre los valores $\{-2,-1,0,1,2\}$. De la gráfica y de (1.46) es evidente que las $t+1$ raíces de $\alpha=f(\lambda)$, que denotamos por $\delta_1,\delta_2,\ldots,\delta_{t+1}$ satisfacen las relaciones

$$\infty>\delta_1>\beta_1;\quad \beta_{i-1}>\delta_i>\beta_i,\quad i=1,\ldots,t;\quad \beta_t>\delta_{t+1}>-\infty. \tag{1.47}$$

Con todo ello, llegamos a que los n autovalores de la matriz X pueden clasificarse en 3 grupos, que son

(i) Los autovalores $\gamma_1,\gamma_2,\ldots,\gamma_{n-s-1}$ asociados a las componentes nulas del vector $\mathbf{a}$, que son iguales a los $n-s-1$ correspondientes elementos de $\{\alpha_i\}$.

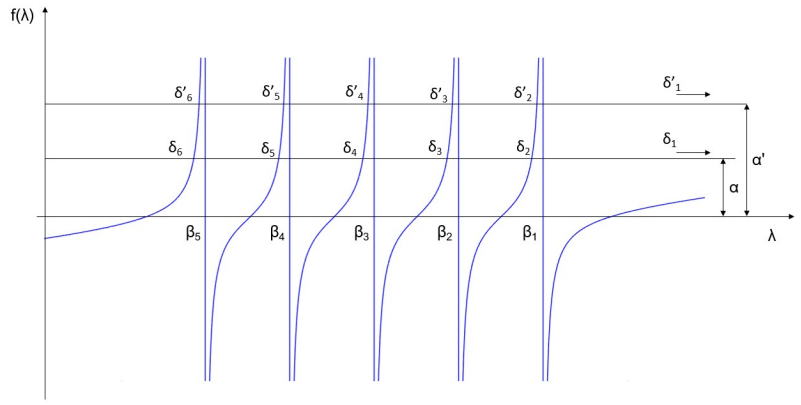

Figura 1.6: Autovalores de una matriz Z de la forma (1.43), para dos valores distintos de α y α' manteniendo fijo el vector $\mathbf{a}$.

(ii) $s-t$ autovalores iguales a β_i, con multiplicidad $r_i - 1$, $i = 1, ..., t$, que son iguales a otros $s-t$ elementos de $\{\alpha_i\}$.

(iii) $t+1$ autovalores iguales a δ_i que satisfacen las relaciones (1.47).

Notamos que sólo los elementos del conjunto (iii) dependen de α.

Teorema 1.4.3 *Si denotamos los autovalores de la matriz X dada por (1.42), como $\{\lambda_1, \ldots, \lambda_n\}$, dados en orden no creciente, entonces, si los $\{\alpha_i\}$ también están ordenados en orden no creciente se tiene que*

$$\lambda_1 \geq \alpha_1 \geq \lambda_2 \geq \cdots \geq \alpha_{n-1} \geq \lambda_n. \tag{1.48}$$

Los elementos $\{\alpha_1, \ldots, \alpha_{n-1}\}$ separan a los autovales $\{\lambda_1, \ldots, \lambda_n\}$, al menos en sentido débil.

Perturbación de la matriz X

Consideramos ahora los autovalores de la matriz X' obtenida a partir de la matriz dada por (1.42) reemplazando α por α' y suponiendo que $\alpha' > \alpha$. Los autovalores de X' serán iguales a los de X en lo que a los conjuntos (i) y (ii) se refiere.

Denotamos por $\delta'_1, \delta'_2, \ldots, \delta'_{t+1}$ a los autovalores de X' del conjunto (iii). Dichos autovalores son las raíces de

$$\alpha' - f(\lambda) = (\alpha' - \lambda) - \sum_{i=1}^{t} c_i^2 (\beta_i - \lambda)^{-1} = 0.$$

Derivando la función $f(\lambda)$ se tiene que

$$\frac{df(\lambda)}{d\lambda} = 1 + \sum_{i=1}^{t} \frac{c_i^2}{(\beta_i - \lambda)^2} > 1, \quad \lambda \neq \beta_i.$$

Esto implica, por tanto, que en cada intervalo (β_{i+1}, β_i), $f(\lambda)$ es creciente y si $f(\delta_i) = \alpha$ y $f(\delta_i') = \alpha'$, como $\alpha' > \alpha$ entonces $\delta_i' > \delta_i$.

Por otra parte, en este mismo intervalo, la pendiente en cada punto es mayor que 1, por tanto, se cumple que $\dfrac{\alpha' - \alpha}{\delta_i' - \delta_i} > 1$, es decir, $\delta_i' - \delta_i < \alpha' - \alpha$, $1 \leq i \leq t+1$.

Entonces, cada $\delta_i' - \delta_i$ está entre 0 y $\alpha' - \alpha$. Podemos definir ciertas cantidades m_i, $1 \leq i \leq t+1$, por

$$\delta_i' - \delta_i = m_i(\alpha' - \alpha), \ \text{ con } \ 0 < m_i < 1,$$

y se cumple que

$$\sum_i m_i = 1. \tag{1.49}$$

Para ver (1.49), como $m_i = \dfrac{\delta_i' - \delta_i}{\alpha' - \alpha}$, $1 \leq i \leq t+1$, y la suma de los autovalores δ_i es igual a la traza de la matriz Z descrita en (1.43), se tiene

$$\sum_{i=1}^{t+1} m_i = \frac{1}{\alpha' - \alpha}\left(\left(\sum_{i=1}^{t+1} \delta_i'\right) - \left(\sum_{i=1}^{t+1} \delta_i\right)\right) = \frac{1}{\alpha' - \alpha}\left(\operatorname{tr}(Z_{\alpha'}) - \operatorname{tr}(Z_\alpha)\right) = \frac{\alpha' - \alpha}{\alpha' - \alpha} = 1.$$

Si $\mathbf{a} = \mathbf{0}$, entonces $\delta_1' = \alpha'$ y $\delta_1 = \alpha$ y en consecuencia, $\delta_1' - \delta_1 = \alpha' - \alpha$.

Entonces podemos escribir en todos los casos

$$\begin{aligned} &\delta_i' - \delta_i = m_i(\alpha' - \alpha), \ \text{ con } \ 0 < m_i \leq 1, \\ &\sum_i m_i = 1. \end{aligned}$$

Como los otros autovalores de X y X' son iguales, podemos decir que hemos establecido una correspondencia entre los n autovalores de X a los que renombramos por $\{\lambda_1, \ldots, \lambda_n\}$ y los n autovalores de X', renombrados por $\{\lambda_1', \ldots, \lambda_n'\}$, de la siguiente manera

Teorema 1.4.4 *Sean $\{\lambda_1, \ldots, \lambda_n\}$ los n autovalores, en orden no creciente, de una matriz X dada por (1.42), y sean $\{\lambda_1', \ldots, \lambda_n'\}$ los autovalores de la matriz X' obtenida a partir de X reemplazando α por α', con $\alpha > \alpha'$. Entonces*

$$\begin{aligned} &\lambda_i' - \lambda_i = m_i(\alpha' - \alpha), \ \ \textit{con } \ 0 \leq m_i \leq 1, \quad \textit{y} \\ &\sum_i m_i = 1, \end{aligned} \tag{1.50}$$

donde es claro que $m_i = 0$ para los autovalores de los conjuntos (i) y (ii).

Una observación que será últil más adelante es que si las relaciones (1.50) se satisfacen para los conjuntos $\{\lambda_1, \ldots, \lambda_n\}$ y $\{\lambda_1', \ldots, \lambda_n'\}$ con algún orden concreto, entonces se satisfacen con mayor motivo cuando los $\{\lambda_i\}$ y los $\{\lambda_i'\}$ están cada uno dispuestos en orden no creciente.

Perturbaciones simétricas de rango 1

Consideramos en este apartado el caso particular en que

$$C = A + B,$$

donde A y B son simétricas y B tiene rango 1. Puesto que la matriz B tiene rango 1, si ρ denota su único autovalor no nulo, existe una matriz ortogonal R tal que

$$R^T B R = \left(\begin{array}{c|c} \rho & \mathbf{0}^T \\ \hline \mathbf{0} & O \end{array}\right), \quad R^T A R = \left(\begin{array}{c|c} \alpha & \mathbf{a}^T \\ \hline \mathbf{a} & A_{n-1} \end{array}\right),$$

donde O denota la matriz nula. La matriz A_{n-1} es simétrica y, por tanto, existe una matriz ortogonal S de orden $n-1$ con

$$S^T A_{n-1} S = \mathrm{diag}(\alpha_i).$$

La matriz Q definida por

$$Q = R \left(\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & S \end{array}\right),$$

verifica las propiedades siguientes:

- Q es ortogonal, por serlo R y S.
- Además,

$$\begin{aligned} Q^T(A+B)Q = Q^T A Q + Q^T B Q &= \left(\begin{array}{c|c} \alpha & \mathbf{b}^T \\ \hline \mathbf{b} & \mathrm{diag}(\alpha_i) \end{array}\right) \\ &+ \left(\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & S^T \end{array}\right) R^T B R \left(\begin{array}{c|c} 1 & \mathbf{0}^T \\ \hline \mathbf{0} & S \end{array}\right) \\ &= \left(\begin{array}{c|c} \alpha & \mathbf{b}^T \\ \hline \mathbf{b} & \mathrm{diag}(\alpha_i) \end{array}\right) + \left(\begin{array}{c|c} \rho & \mathbf{0}^T \\ \hline \mathbf{0} & 0 \end{array}\right), \end{aligned}$$

  donde $\mathbf{b} = S^T \mathbf{a}$.

Los autovalores de A y $A+B$ son, por tanto, los autovalores de

$$\left(\begin{array}{c|c} \alpha & \mathbf{b}^T \\ \hline \mathbf{b} & \mathrm{diag}(\alpha_i) \end{array}\right) \quad y \quad \left(\begin{array}{c|c} \alpha+\rho & \mathbf{b}^T \\ \hline \mathbf{b} & \mathrm{diag}(\alpha_i) \end{array}\right), \tag{1.51}$$

respectivamente. En este punto vemos claramente que podemos aplicar los resultados obtenidos en el apartado previo. En efecto, si denotamos por $\{\lambda_i\}$ y $\{\lambda'_i\}$, respectivamente, a los autovalores de las dos matrices de (1.51), ordenadas en orden no creciente, entonces satisfacen las siguientes relaciones

$$\begin{cases} \lambda'_i - \lambda_i = m_i\rho, \quad 0 \leq m_i \leq 1, \\ \sum m_i = 1. \end{cases}$$

Por tanto, al pasar de A a $A+B$ todos los autovalores de $A+B$ se obtienen sumando a los de A una cantidad entre 0 y ρ, siendo ρ el único autovalor no nulo de B.

De la nota final del apartado anterior y de la estructura de la matriz ortogonalmente semejante a A dada por (1.51) es claro que los autovalores del menor principal A_{n-1} de A separan a los de A según (1.48). Si A tiene un autovalor λ_i de multiplicidad k, entonces necesariamente A_{n-1} debe tener como autovalor a λ_i con multiplicidad k, $k+1$ o $k-1$.

1.4.3. Propiedades extremales de los autovalores de matrices simétricas

El contenido de esta subsección continúa dentro del caso de perturbaciones simétricas de matrices reales simétricas. No obstante, la relevancia e interés práctico de los resultados que se obtienen hacen que merecezca una consideración aparte.

Teorema 1.4.5 *Sea A una matriz real simétrica $n \times n$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ los autovalores de A, ordenados en orden no creciente, y $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$ los autovectores unitarios asociados. Entonces*

(i)

$$\lambda_1 = max\left\{ \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \quad : \quad \mathbf{x} \neq \mathbf{0} \right\}.$$

Además, el máximo se obtiene cuando $\mathbf{x}$ es el autovector de A asociado al autovalor λ_1.

(ii) De manera análoga,

$$\lambda_{s+1} = max\left\{ \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \quad : \quad \mathbf{x} \neq \mathbf{0}, \quad \mathbf{r}_1^T\mathbf{x} = \cdots = \mathbf{r}_s^T\mathbf{x} = 0 \right\},$$

y el máximo se obtiene cuando $\mathbf{x}$ es el autovector de A asociado a λ_{s+1}.

Demostración.

Por ser A una matriz real simétrica, la matriz R cuyas columnas son $\mathbf{r}_1, \ldots, \mathbf{r}_n$ es ortogonal y satisface

$$R^T AR = \text{diag}(\lambda_1, \ldots, \lambda_n).$$

Para cada vector $\mathbf{x} \in \mathbb{R}^n$, definimos

$$\mathbf{x} = R\mathbf{y} \iff \mathbf{y} = R^T\mathbf{x}. \tag{1.52}$$

Entonces,

$$\begin{cases} \mathbf{x}^T A\mathbf{x} = \mathbf{y}^T R^T AR\mathbf{y} = \mathbf{y}^T \mathrm{diag}(\lambda_1, \ldots, \lambda_n)\mathbf{y} = \displaystyle\sum_{i=1}^{n} \lambda_i y_i^2, \\ \mathbf{x}^T\mathbf{x} = \mathbf{y}^T R^T R\mathbf{y} = \displaystyle\sum_{i=1}^{n} y_i^2. \end{cases}$$

Como la expresión (1.52) establece una relación uno a uno entre los vectores $\mathbf{x}$ e $\mathbf{y}$ y $\mathbf{x}^T\mathbf{x} = \mathbf{y}^T\mathbf{y}$, el problema orginal es quivalente a encontrar el máximo valor de

$$\max\left\{ \sum_{i=1}^{n} \lambda_i y_i^2 : \ \sum_{i=1}^{n} y_i^2 = 1 \right\}. \tag{1.53}$$

Como los autovalores de A, $\{\lambda_1, \ldots, \lambda_n\}$, están en orden no creciente, entonces

$$\lambda_n \leq \sum_{i=1}^{n} \lambda_i y_i^2 \leq \lambda_1. \tag{1.54}$$

Además, el máximo de (1.53) se alcanza cuando $\displaystyle\sum_{i=1}^{n} \lambda_i y_i^2 = \lambda_1$, esto es, para $\mathbf{y} = \mathbf{e}_1$. El correspondiente vector $\mathbf{x}$ es la primera columna de la matriz R, que es por tanto, un autovector unitario de A asociado a λ_1. Notamos que si $\lambda_1 = \lambda_2 = \cdots = \lambda_r \neq \lambda_{r+1}$, entonces cualquier vector unitario $\mathbf{y}$ perteneciente al subespacio generado por $\{\mathbf{e}_1, \ldots, \mathbf{e}_r\}$ da el valor λ_1 en la suma considerada en (1.54). Análogamente, λ_n es el mínimo valor de $\mathbf{x}^T A\mathbf{x}$ bajo la misma condición, y se alcanza con $\mathbf{x} = \mathbf{r}_n$.

Consideramos ahora el mismo problema de maximización sujeto a las restricciones adicionales de que $\mathbf{x}$ sea ortogonal a $\mathbf{r}_i$, para $1 \leq i \leq s$. De las relaciones

$$0 = \mathbf{r}_i^T\mathbf{x} = \mathbf{r}_i^T R\mathbf{y} = \mathbf{e}_i^T R^T R\mathbf{y} = \mathbf{e}_i^T\mathbf{y}, \quad 1 \leq i \leq n,$$

vemos que la correspondiente restricción en $\mathbf{y}$ es que tenga un cero en sus s primeras componentes. Entonces, el máximo valor de $\mathbf{x}^T A\mathbf{x}$ sujeto a estas restricciones es λ_{s+1}, y este se alcanza para $\mathbf{y} = \mathbf{e}_{s+1}$. El correspondiente vector $\mathbf{x}$ es ahora la $(s+1)-$ésima columna de la matriz R.

□

A la hora de poner en práctica esta caracterización de los autovalores de una matriz simétrica surge un inconveniente. La determinación de cada λ_s depende del conocimiento de los autovectores de A asociados a $\lambda_1, \lambda_2, \ldots, \lambda_{s-1}$.

Vamos a tratar de encontrar ahora una caracterización que nos permita obtener los autovalores de una matriz A simétrica, y que no presente la desventaja que acabamos de mencionar. Es decir, que la obtención del autovalor λ_i no precise de haber calculado previamente $\lambda_1, \ldots, \lambda_{i-1}$.

Teorema 1.4.6 (Teorema de Courant-Fischer) *Sea A una matriz simétrica y $\{\lambda_i\}$ el conjunto de autovalores de A, ordenados de manera no creciente. Consideramos el máximo valor de $\mathbf{x}^T A\mathbf{x}$ sujeto a las condiciones*

$$\begin{cases} \mathbf{x}^T\mathbf{x} = 1, \\ \mathbf{p}_i^T\mathbf{x} = 0, \ \ \mathbf{p}_i \neq \mathbf{0}, \ \ i = 1, 2, \ldots, s, \ \ \textit{con } s < n, \end{cases} \tag{1.55}$$

donde los $\mathbf{p}_i$ *son vectores no nulos cualesquiera. Entonces, el mínimo valor que toma dicho máximo para todas las posibles elecciones de los* s *vectores* $\mathbf{p}_i$ *es* λ_{s+1}.

Demostración.

Para todo vector $\mathbf{x}$, que satisface $\mathbf{x}^T\mathbf{x} = 1$, por el Teorema 1.4.5 sabemos que $\lambda_n \leq \mathbf{x}^T A\mathbf{x} \leq \lambda_1$. De manera que $\mathbf{x}^T A\mathbf{x}$ está acotado, y su máximo sujeto a las condiciones dadas por (1.55) es claramente función de las $n \times s$ componentes de los vectores $\mathbf{p}$.

La pregunta que nos hacemos es cuál es el mínimo valor que toma dicho máximo para todas las posibles elecciones de los s vectores $\mathbf{p}$.

Como antes, para resolverlo, vamos a trabajar con el cambio de variable dado por $\mathbf{y} = R^T\mathbf{x}$. Las relaciones (1.55) se convierten en

$$\begin{cases} \mathbf{y}^T\mathbf{y} = 1, \\ \mathbf{q}_i^T\mathbf{y} = 0, \text{ donde } \mathbf{q}_i = R^T\mathbf{p}_i \neq \mathbf{0}, \ \ i = 1, 2, \ldots, s; \text{ con } s < n. \end{cases}$$

Consideramos una elección particular de vectores $\mathbf{p}_1, \ldots, \mathbf{p}_s$. Tal elección da el conjunto de los correspondientes $\mathbf{q}_1, \ldots, \mathbf{q}_s$. Las n variables y_i (componentes del vector $\mathbf{y}$) satisfacen s ecuaciones lineales homogéneas

$$\mathbf{q}_i^T\mathbf{y} = 0, \ \ i = 1, \ldots, s.$$

Si añadimos las relaciones $y_{s+2} = y_{s+3} = \cdots = y_n = 0$, entonces tenemos un total de $n-1$ ecuaciones homogéneas en las n incógnitas $y_1, \ldots, y_n$, y por tanto, hay al menos una solución no nula $(y_1, \ldots, y_s, y_{s+1}, 0, \ldots, 0)$ del sistema, que puede ser normalizada para que se verifique que $\sum_{i=1}^{s+1} y_i^2 = 1$. Con esta elección de $\mathbf{y}$ tenemos

$$\mathbf{y}^T \operatorname{diag}(\lambda_i)\mathbf{y} = \sum_{i=1}^{s+1} \lambda_i y_i^2 \geq \lambda_{s+1}.$$

Esto demuestra que para cualquier elección de $\{\mathbf{p}_i\}$ siempre habrá un vector $\mathbf{y}$, y por tanto un vector $\mathbf{x}$, para los que

$$\mathbf{x}^T A\mathbf{x} = \mathbf{y}^T \operatorname{diag}(\lambda_i)\mathbf{y} \geq \lambda_{s+1}.$$

Es decir,

$$\text{máx}\{\mathbf{x}^T A\mathbf{x}\} \geq \lambda_{s+1},$$

para cualquier elección de los s vectores $\{\mathbf{p}_i\}$. Esto significa que

$$\text{minmax}\{\mathbf{x}^T A\mathbf{x}\} \geq \lambda_{s+1}. \tag{1.56}$$

Por otra parte, si tomamos la elección de los vectores $\{\mathbf{p}_i\}$ como $\mathbf{p}_i = R\mathbf{e}_i, \quad 1 \leq i \leq s$, entonces $\mathbf{q}_i = \mathbf{e}_i, \quad 1 \leq i \leq s$, y las relaciones (1.55) se convierten en

$$y_i = 0, \ \ i = 1, 2, \ldots, s.$$

Por tanto, para cualquier $\mathbf{y}$ sujeto a estas relaciones particulares tenemos

$$\mathbf{x}^T A\mathbf{x} = \mathbf{y}^T \operatorname{diag}(\lambda_i)\mathbf{y} = \sum_{i=s+1}^{n} \lambda_i y_i^2 \leq \lambda_{s+1}.$$

Los resultados obtenidos implican que para esta elección de los vectores $\{\mathbf{p}_i\}$,

$$\text{máx}\{\mathbf{x}^T A\mathbf{x}\} = \lambda_{s+1},$$

y que este valor se alcanza para la elección del vector $\mathbf{y}$, dada por

$$y_{s+1} = 1; \;\; y_i = 0 \;\text{ para }\; i \neq s+1.$$

De esta manera, hemos encontrado una elección de los vectores $\{\mathbf{p}_i\}$ para la que se tiene $\text{máx}(\mathbf{x}^T A\mathbf{x}) = \lambda_{s+1}$. Por lo tanto

$$\text{minmax}\{\mathbf{x}^T A\mathbf{x}\} \leq \lambda_{s+1}. \tag{1.57}$$

De (1.56) y (1.57) deducimos directamente que

$$\text{minmax}\{\mathbf{x}^T A\mathbf{x}\} = \lambda_{s+1},$$

cuando $\mathbf{x}$ está sujeto a las condiciones dadas por (1.55).

□

El resultado que hemos obtenido se conoce habitualmente como *caracterización minimax* de los autovalores de una matriz simétrica. Cabe destacar que hemos probado que existe al menos un conjunto de vectores $\{\mathbf{p}_i\}$ y un vector $\mathbf{x}$ correspondiente para los cuales se alcanza el valor minimax.

De una manera análoga podemos probar que

$$\lambda_s = \text{maxmin}\{\mathbf{x}^T A\mathbf{x}\} \;\text{ para }\; \mathbf{x}^T\mathbf{x} = 1, \;\; \mathbf{p}_i^T\mathbf{x} = 0 \;\; i = 1, 2, \ldots, n-s.$$

Notar que en ambas caracterizaciones los conjuntos de vectores $\{\mathbf{p}_i\}$ incluyen aquellos en los que algunos de los $\{\mathbf{p}_i\}$ son iguales, aunque en general, la minimización del máximo ocurrirá para un conjunto de vectores $\{\mathbf{p}_i\}$ todos ellos distintos.

Si tenemos algún vector $\mathbf{p}_i$ repetido s veces, podemos afirmar que

$$\lambda_{s+1} \leq \max(\mathbf{x}^T A\mathbf{x}) \;\text{ para }\; \mathbf{x}^T\mathbf{x} = 1, \;\; \mathbf{p}_i^T\mathbf{x} = 0 \;\; i = 1, 2, \ldots, s,$$

donde los $\{\mathbf{p}_i\}$ pueden ser, o no, distintos.

1.4.4. Autovalores de la suma de dos matrices simétricas

La caracterización minimax de los autovalores de una matriz simétrica que se ha visto en el Teorema 1.4.6, puede utilizarse para establecer una relación entre los autovalores de dos matrices simétricas A, B y los autovalores de la matriz suma $C = A + B$. Esta relación será útil para probar resultados posteriores.

Teorema 1.4.7 *Sean A, B dos matrices reales simétricas, y sea $C = A + B$. Denotamos por $\{\alpha_i\}$, $\{\beta_i\}$ y $\{\gamma_i\}$ los autovalores de A, B y C, respectivamente, donde los tres conjuntos se suponen ordenados de forma no creciente. Entonces*

$$\alpha_s + \beta_n \leq \gamma_s \leq \alpha_s + \beta_1, \quad 1 \leq s \leq n. \tag{1.58}$$

Demostración.

Por el Teorema 1.4.6, sabemos que para $1 \leq s \leq n$,

$$\begin{cases} \gamma_s = \text{minmax}(\mathbf{x}^T C\mathbf{x}), \text{ cuando} \\ \mathbf{x}^T\mathbf{x} = 1, \ \mathbf{p}_i^T\mathbf{x} = 0, \ i = 1, 2, \ldots, s-1. \end{cases} \tag{1.59}$$

Así para una elección particular de los vectores $\{\mathbf{p}_i\}$, tenemos para todo $\mathbf{x}$ que verifique las condiciones impuestas en (1.59), que

$$\gamma_s \leq \max(\mathbf{x}^T C\mathbf{x}) = \max(\mathbf{x}^T A\mathbf{x} + \mathbf{x}^T B\mathbf{x}). \tag{1.60}$$

Si U es la matriz ortogonal tal que $U^T AU = \text{diag}(\alpha_1, \ldots, \alpha_n)$, entonces tomando $\mathbf{p}_i = U\mathbf{e}_i$, las relaciones que aparecen en (1.59) son

$$0 = \mathbf{p}_i^T\mathbf{x} = \mathbf{e}_i^T\mathbf{y} = y_i \ \ i = 1, 2, \ldots, s-1,$$

donde $\mathbf{y} = U^T\mathbf{x}$ y, en consecuencia, $\mathbf{y}^T\mathbf{y} = 1$. Con esta elección de los $\{\mathbf{p}_i\}$ las primeras $s-1$ componentes de $\mathbf{y}$ son nulas, y de (1.60) tenemos

$$\gamma_s \leq \max(\mathbf{x}^T A\mathbf{x} + \mathbf{x}^T B\mathbf{x}) = \max\left(\sum_{i=s}^{n} \alpha_i y_i^2 + \mathbf{x}^T B\mathbf{x}\right).$$

Ahora bien, $\displaystyle\sum_{i=s}^{n} \alpha_i y_i^2 \leq \alpha_s$ y $\mathbf{x}^T B\mathbf{x} \leq \beta_1 \ \ \forall \mathbf{x}$, por lo que

$$\left(\sum_{i=s}^{n} \alpha_i y_i^2 + \mathbf{x}^T B\mathbf{x}\right) \leq \alpha_s + \beta_1,$$

para todo $\mathbf{x}$ correspondiente a esta elección de los vectores $\{\mathbf{p}_i\}$. Por tanto, el máximo de $\mathbf{x}^T C\mathbf{x}$ sobre tales vectores $\mathbf{x}$, no es mayor que $\alpha_s + \beta_1$ y llegamos a que

$$\gamma_s \leq \alpha_s + \beta_1. \tag{1.61}$$

Como $A = C + (-B)$ y los autovalores de $-B$ en orden no creciente son $-\beta_n, -\beta_{n-1}$, $\ldots, -\beta_1$, la aplicación del resultado que acabamos de probar nos lleva a que

$$\alpha_s \leq \gamma_s + (-\beta_n),$$

luego

$$\gamma_s \geq \alpha_s + \beta_n. \tag{1.62}$$

Entonces de (1.61) y (1.62) obtenemos que, cuando se suma una matriz simétrica B a una matriz simétrica A, todos sus autovalores cambian en una cierta cantidad que se sitúa entre el autovalor más pequeño y el autovalor más grande de B.

□

El resultado anterior es de gran interés en la práctica. Frecuentemente la matriz B será pequeña, (una perturbación) y la única información que tendremos sobre ella será alguna cota superior de sus elementos o quizá una cota superior para alguna norma. Si,

por ejemplo, tenemos como en la Sección 1.1.1, $|b_{ij}| \leq \epsilon$, entonces $-n\epsilon \leq \beta_n \leq \beta_1 \leq n\epsilon$ y, en consecuencia

$$|\gamma_r - \alpha_r| \leq n\epsilon. \tag{1.63}$$

Notemos que (1.63) es más fuerte que el resultado establecido en el Teorema 1.4.2 para una perturbación no simétrica ya que ahora no hay restricciones en las separaciones de los autovalores α_i, β_i y γ_i. Anteriormente solo fuimos capaces de probar la relación (1.63) cuando los autovalores α_i estaban todos separados por más de $2n\epsilon$ (ver (1.41)).

El Teorema 1.4.7 extiende el resultado probado analíticamente en la Sección 1.4.2 para una perturbación B simétrica de rango 1. Como ya se dijo antes, los resultados probados no son ciertos únicamente para pertubaciones pequeñas, y tampoco se ven afectados por la multiplicidad de los autovalores de las matrices A, B o C.

Teorema 1.4.8 (Generalización del Teorema minimax.) *Sean A, B dos matrices reales simétricas, y sea $C = A + B$. Denotamos por $\{\alpha_i\}$, $\{\beta_i\}$ y $\{\gamma_i\}$ los autovalores de A, B y C, respectivamente, donde los tres conjuntos se suponen ordenados de forma no creciente. Entonces,*

$$\gamma_{r+s-1} \leq \alpha_r + \beta_s \quad \textit{para } r + s - 1 \leq n. \tag{1.64}$$

Demostración.

Antes de demostrar el teorema, notamos que (1.64) es una generalización de (1.58). El Teorema 1.4.6 nos permite afirmar que existe al menos una elección de los vectores $\{\mathbf{p}_i\}$ para la que

$$\max(\mathbf{x}^T A\mathbf{x}) = \alpha_r \ \text{ cuando } \mathbf{p}_i^T\mathbf{x} = 0, \ \ i = 1, 2, \ldots, r-1,$$

y un conjunto de vectores $\{\mathbf{q}_i\}$ para los que

$$\max(\mathbf{x}^T B\mathbf{x}) = \beta_s \ \text{ cuando } \mathbf{q}_i^T\mathbf{x} = 0, \ \ i = 1, 2, \ldots, s-1.$$

Consideremos ahora el conjunto de vectores $\mathbf{x}$ que satisfacen

$$\begin{cases} \mathbf{p}_i^T\mathbf{x} = 0 \ \ i = 1, 2, \ldots, r-1, \\ \mathbf{q}_i^T\mathbf{x} = 0 \ \ i = 1, 2, \ldots, s-1. \end{cases} \tag{1.65}$$

Dicho conjunto es no vacío ya que el número total de ecuaciones es $r+s-2 < r+s-1 \leq n$ esto es, $r + s - 2 \leq n - 1$.

Para un vector $\mathbf{x}$ de dicho conjunto tenemos que

$$\mathbf{x}^T C\mathbf{x} = \mathbf{x}^T A\mathbf{x} + \mathbf{x}^T B\mathbf{x} \leq \alpha_r + \beta_s.$$

Entonces,

$$\max(\mathbf{x}^T C\mathbf{x}) \leq \alpha_r + \beta_s,$$

para todo $\mathbf{x}$ que satisface las $r + s - 2$ ecuaciones lineales (1.65).

Por lo tanto,

$$\gamma_{r+s-1} = \text{minmax}(\mathbf{x}^T C\mathbf{x}) \leq \alpha_r + \beta_s,$$

ya que la minimización se hace sobre todos los conjuntos de vectores $\mathbf{x}$ que satisfacen las $r+s-2$ ecuaciones (1.65).

□

Como una ilustración más de la utilidad del Teorema de caracterización minimax de los autovalores de una matriz simétrica, vamos a probar que los autovalores $\{\lambda'_1,\ldots,\lambda'_{n-1}\}$ del menor principal A_{n-1} de una matriz real y simétrica $A\in\mathcal{M}_{n\times n}$ separan a los autovalores $\{\lambda_1,\ldots,\lambda_n\}$ de A. Este resultado generaliza al que ya fue probado en la Sección 1.4.2 para matrices simétricas obtenidas orlando una matriz diagonal.

Teorema 1.4.9 (Teorema de separación) *Sea A una matriz simétrica y $\{\lambda_1,\ldots,\lambda_n\}$ el conjunto de autovalores de A. Si A_{n-1} es el menor principal de orden $n-1$ de la matriz A, y $\{\lambda'_1,\ldots,\lambda'_{n-1}\}$ son los autovalores de A_{n-1}, entonces*

$$\lambda_{s+1}\le\lambda'_s\le\lambda_s,\quad 1\le s\le n-1. \tag{1.66}$$

Demostración

El conjunto de los valores que toma $\tilde{\mathbf{x}}^TA_{n-1}\tilde{\mathbf{x}}$ para todo vector unitario $\tilde{\mathbf{x}}\in\mathbb{R}^{n-1}$ es el mismo que toma $\mathbf{x}^TA\mathbf{x}$ para todo vector unitario $\mathbf{x}\in\mathbb{R}^n$ con $x_n=0$ ($\mathbf{x}=[\tilde{\mathbf{x}}^T,0]^T$). Entonces,

$$\begin{cases}\lambda'_s=\text{minmax}(\mathbf{x}^TA\mathbf{x}),\ \ \mathbf{x}^T\mathbf{x}=1\\ x_n=0;\ \ \mathbf{p}_i^T\mathbf{x}=0,\ \ i=1,2,\ldots s-1.\end{cases} \tag{1.67}$$

Ahora bien, λ'_s se alcanzará para algún conjunto concreto de vectores $\{\mathbf{p}_i\}$ en (1.67), y para este conjunto de $\{\mathbf{p}_i\}$, λ'_s es el máximo valor de $\mathbf{x}^TA\mathbf{x}$ sujeto a las s ecuaciones lineales de la segunda línea de (1.67). Por otra parte, λ_{s+1} es el mínimo valor que toma dicho máximo sujeto a s ecuaciones lineales cualesquiera. Entonces, es claro que

$$\lambda_{s+1}\le\lambda'_s. \tag{1.68}$$

Consideramos ahora cualquier conjunto de $s-1$ vectores $\{\mathbf{p}_i\}$. Denotamos por $f(\mathbf{p}_i)$, al máximo valor de $\mathbf{x}^TA\mathbf{x}$ para vectores unitarios $\mathbf{x}$ sujetos a las relaciones lineales correspondientes a los vectores $\mathbf{p}_i$, $i=1,2,\ldots,s-1$. Sea $f_{n-1}(\mathbf{p}_i)$, el máximo de $\mathbf{x}^TA\mathbf{x}$ sujeto a la relación extra $x_n=0$. Entonces,

$$f_{n-1}(\mathbf{p}_i)\le f(\mathbf{p}_i),$$

y, tomando el mínimo entre todos los conjuntos de vectores $\{\mathbf{p}_i\}$,

$$\min\{f_{n-1}(\mathbf{p}_i)\}\le\min\{f_n(\mathbf{p}_i)\},$$

implicando que

$$\lambda'_s\le\lambda_s. \tag{1.69}$$

Juntando (1.68) y (1.69) obtenemos el resultado (1.66) que estamos buscando.

□

Los resultados vistos hasta ahora en la Sección 1.4 se extienden inmediatamente a matrices hermíticas en general, sin más que sustituir el superíndice ^{T} por ^{H} en los resultados y demostraciones.

El teorema que presentamos a continuación permite relacionar la perturbación de los autovalores de una matriz simétrica, con la norma de Frobenius de la matriz de perturbación, que seguimos considerando también simétrica.

Teorema 1.4.10 (Teorema de Wielandt-Hoffman) *Sean A, B dos matrices reales simétricas, y sea $C = A + B$. Denotamos por $\{\alpha_i\}$, $\{\beta_i\}$ y $\{\gamma_i\}$ los autovalores de A, B y C, respectivamente, donde los tres conjuntos se suponen ordenados de forma no creciente. Entonces*

$$\sum_{i=1}^{n}(\gamma_i - \alpha_i)^2 \leq \|B\|_F^2 = \sum_{i=1}^{n}\beta_i^2.$$

Demostración.

Para la demostración de este teorema será de interés recordar dos resultados previos que están enunciados como Lema A.2.1 y Lema A.2.2 en la Sección A.2 del Apéndice.

Sean U_1 y U_2 matrices ortogonales tales que $U_1^T B U_1 = \mathrm{diag}(\beta_1, \ldots, \beta_n)$ y $U_2^T C U_2 = \mathrm{diag}(\gamma_1, \ldots, \gamma_n)$. Entonces,

$$\begin{aligned}\mathrm{diag}(\beta_1, \ldots, \beta_n) =& U_1^T B U_1 = U_1^T (C - A) U_1 = U_1^T \left(U_2\, \mathrm{diag}(\gamma_1, \ldots, \gamma_n)\, U_2^T - A\right) U_1 \\ =& U_1^T U_2 [\mathrm{diag}(\gamma_1, \ldots, \gamma_n) - U_2^T A U_2]\, U_2^T U_1.\end{aligned}$$

Tomando normas en la igualdad anterior y teniendo en cuenta de nuevo que la norma de Frobenius no cambia si se multiplica por una matriz ortogonal, obtenemos

$$\sum_{i=1}^{n}\beta_i^2 = \|\mathrm{diag}(\gamma_i) - U_2^T A U_2\|_F^2. \tag{1.70}$$

Consideramos ahora el conjunto de valores que toma la función definida por

$$f(U) = \|\mathrm{diag}(\gamma_i) - U^T A U\|_F^2,$$

para toda matriz U ortogonal.

La ecuación (1.70) muestra que la cantidad $\displaystyle\sum_{i=1}^{n}\beta_i^2$ pertenece a este conjunto. Dicho conjunto de valores es acotado y tiene una cota superior K_u, y una cota inferior K_l, ambas finitas, que se alcanzan para alguna matriz ortogonal U, ya que $f(U)$ es una función continua en el conjunto compacto de las matrices ortogonales.

Veamos que la cota inferior K_l debe alcanzarse para una matriz U tal que $U^T A U$ es diagonal.

En general, entre los autovalores $\{\gamma_1, \ldots, \gamma_n\}$, de la matriz C tendremos r valores distintos que denotamos por $\{\delta_1 > \delta_2 > \cdots > \delta_r\}$. Entonces, escribimos

$$\mathrm{diag}(\gamma_1, \ldots, \gamma_n) = \begin{pmatrix} \delta_1 I & & & \\ & \delta_2 I & & \\ & & \ddots & \\ & & & \delta_r I \end{pmatrix}, \tag{1.71}$$

donde cada submatriz I tiene el orden adecuado a la multiplicidad del correspondiente autovalor δ_i. Consideramos $U^T A U$ dividida en cajas de las mismas dimensiones que

$\mathrm{diag}(\gamma_1, \ldots, \gamma_r)$ en (1.71),

$$U^T AU = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1r} \\ X_{21} & X_{22} & \cdots & X_{2r} \\ \cdots & \cdots & \cdots & \cdots \\ X_{r1} & X_{r2} & \cdots & X_{rr} \end{pmatrix} \equiv X.$$

Vamos a probar que solo se puede alcanzar la cota K_l para una matriz U cuyos bloques no diagonales son todos nulos.

Sea x un elemento no nulo en la fila p y columna q de $U^T AU$, que corresponde a un bloque X_{ij}, $i \neq j$ (no diagonal). Entonces, los elementos en las intersecciones de las filas y columnas p y q de las matrices $\mathrm{diag}(\gamma_1, \ldots, \gamma_n)$ y $U^T AU$, son

$$\mathrm{diag}(\gamma_1, \ldots, \gamma_n): \quad \begin{array}{l} \\ \text{fila } p \\ \\ \text{fila } q \\ \\ \end{array} \begin{pmatrix} & \text{col } p & & \text{col } q & \\ & \vdots & & \vdots & \\ \cdots & \delta_i & \cdots & 0 & \cdots \\ & \vdots & & \vdots & \\ \cdots & 0 & \cdots & \delta_j & \cdots \\ & \vdots & & \vdots & \end{pmatrix},$$

$$X = U^T AU: \quad \begin{array}{l} \\ \text{fila } p \\ \\ \text{fila } q \\ \\ \end{array} \begin{pmatrix} & \text{col } p & & \text{col } q & \\ & \vdots & & \vdots & \\ \cdots & a & \cdots & x & \cdots \\ & \vdots & & \vdots & \\ \cdots & x & \cdots & b & \cdots \\ & \vdots & & \vdots & \end{pmatrix}.$$

Veamos que en ese caso podemos tomar una matriz ortogonal S, correspondiente a una rotación en el plano (p, q), tal que

$$g(S) = \|\mathrm{diag}(\gamma_1, \ldots, \gamma_n) - S^T U^T AUS\|_F^2 - \|\mathrm{diag}(\gamma_1, \ldots, \gamma_n) - U^T AU\|_F^2 < 0,$$

es decir,

$$\|\mathrm{diag}(\gamma_1, \ldots, \gamma_n) - S^T U^T AUS\|_F^2 < \|\mathrm{diag}(\gamma_1, \ldots, \gamma_n) - U^T AU\|_F^2,$$

y, por tanto, el último término no es cota inferior del conjunto de valores que estamos considerando.

En efecto, sabemos que la matriz S en el plano (p, q) debe ser de la forma

$$S = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix},$$

para algún θ. Denotamos por a', b' y x' a los elementos correspondientes de $S^T U^T AUS$ en las intersecciones de las filas y columnas p y q.

Como $\|S^TU^TAUS\|_F^2 = \|U^TAU\|_F^2$, de la definición de norma de Frobenius vemos que $\|A+B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 + 2\sum_{i,j=1}^n a_{ij}b_{ij}$. Entonces,

$$\begin{aligned} g(S) &= \|\mathrm{diag}(\gamma_1,\ldots,\gamma_n)\|_F^2 + \|S^TU^TAUS\|_F^2 - 2\sum_{i=1}^n \gamma_i (S^TU^TAUS)_{ii} \\ &\quad - \|\mathrm{diag}(\gamma_1,\ldots,\gamma_n)\|_F^2 - \|U^TAU\|_F^2 + 2\sum_{i=1}^n \gamma_i (U^TAU)_{ii} \\ &= -2\delta_i(S^TU^TAUS_{pp} - 2\delta_j(S^TU^TAUS_{qq}) + 2\delta_i(U^TAU)_{pp} + 2\delta_j(U^TAU)_{qq}) \\ &= -2a'\delta_i - 2b'\delta_j + 2a\delta_i + 2b\delta_j = 2(a-a')\delta_i + 2(b-b')\delta_j, \end{aligned}$$

donde todos los demás términos se cancelan. El ángulo de la rotación θ satisface

$$\begin{cases} a' = a\cos^2\theta - 2x\cos\theta\sin\theta + b\sin^2\theta, \\ b' = a\sin^2\theta + 2x\cos\theta\sin\theta + b\cos^2\theta. \end{cases}$$

Por lo tanto, podemos escribir tras algunos cálculos

$$\begin{aligned} h(\theta) \equiv g(S) &= 2\delta_i[(a-b)\sin^2\theta + x\sin 2\theta] + 2\delta_j[(b-a)\sin^2\theta - x\sin 2\theta] \\ &= P\sin^2\theta + Q\sin 2\theta, \end{aligned} \tag{1.72}$$

con $P = 2(a-b)(\delta_i - \delta_j)$ y $Q = 2x(\delta_i - \delta_j)$. Observamos que $Q \neq 0$ puesto que hemos supuesto que $x \neq 0$ y sabemos que $\delta_i - \delta_j \neq 0$ $(i \neq j)$. Entonces

$$\frac{dh(\theta)}{d\theta} = Psen(2\theta) + 2Qcos(2\theta)\Big|_{\theta=0} = 2Q,$$

y como $h(0) = 0$ y $h'(0) \neq 0$ se pueden elegir ángulos θ_1 y θ_2 para los cuales $h(\theta_1)\cdot h(\theta_2) < 0$, es decir, rotaciones S_1 y S_2 de ángulos θ_1 y θ_2, respectivamente, para las cuales $g(S_1)\cdot g(S_2) < 0$. En particular, existe un ángulo θ para el cual $g(S) < 0$.

Hemos probado, por tanto, que $f(U)$ no puede ser máximo o mínimo para ninguna matriz U para la que X no sea diagonal por bloques ($X_{ij} = 0,\ \ i \neq j$).

Supongamos que U es una matriz ortogonal para la que $f(U)$ alcanza su valor mínimo. Debe ser entonces

$$diag(\gamma_1,\ldots,\gamma_n) - U^TAU = \begin{pmatrix} \delta_1 I & & & \\ & \delta_2 I & & \\ & & \ddots & \\ & & & \delta_r I \end{pmatrix} - \begin{pmatrix} X_{11} & & & \\ & X_{22} & & \\ & & \ddots & \\ & & & X_{rr} \end{pmatrix}.$$

Si Q_i son matrices ortogonales tales que $Q_i^TX_{ii}Q_i = D_i$, con D_i diagonal, entonces

la matriz Q formada poniendo en su diagonal los bloques Q_i es ortogonal y verifica que

$$Q^T[\mathrm{diag}(\gamma_1,\ldots,\gamma_n) - U^TAU]Q = Q^T\mathrm{diag}(\gamma_1,\ldots,\gamma_n)Q - Q^TU^TAUQ$$
$$= \begin{pmatrix} \delta_1 I & & & \\ & \delta_2 I & & \\ & & \ddots & \\ & & & \delta_r I \end{pmatrix} - \begin{pmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_r \end{pmatrix}$$

$$= \mathrm{diag}(\gamma_1,\ldots,\gamma_n) - Q^TU^TAUQ.$$

Tomando normas en la igualdad obtenida queda $f(UQ) = f(U)$, de donde deducimos que el mínimo debe alcanzarse siempre para una matriz UQ que reduce A a forma diagonal.

Los elementos de D_i son los $\{\alpha_1,\ldots,\alpha_n\}$ en algún orden, y los $\{\delta_1,\ldots,\delta_r\}$, con su multiplicidad apropiada, son los $\{\gamma_1,\ldots,\gamma_n\}$. Entonces tenemos

$$\min_U \ \ f(U) = \sum_{i=1}^{n}(\gamma_i - \alpha_{p_i})^2,$$

donde $(p_1, p_2, \ldots, p_n)$ es una permutación de $(1, 2, \ldots, n)$.

El último paso es probar que el mínimo ocurre para $p_i = i$ para todo i. Escribimos $x = \sum_{i=1}^{n}(\gamma_i - \alpha_{p_i})^2$ para alguna permutación particular.

Si $p_1 = 1$, α_1 y γ_1 ya están emparejados. Si $p_1 \neq 1$, suponemos que es $p_s = 1$ (esto es, el sumando es $(\gamma_s - \alpha_1)^2$). Si intercambiamos p_1 y p_s en la permutación, el cambio en x está dado por

$$(\gamma_1 - \alpha_1)^2 + (\gamma_s - \alpha_s)^2 - (\gamma_1 - \alpha_{p_1})^2 - (\gamma_s - \alpha_1)^2 = -2(\gamma_s - \gamma_1)(\alpha_{p_1} - \alpha_1) \leq 0,$$

y la suma disminuye. De manera análoga, manteniendo α_1 en la primera posición y emparejando α_2 con γ_2 vemos de nuevo que la suma no crece. Finalmente, llegamos a que cada α_i está emparejado con el correspondiente γ_i, y la suma no es mayor que su valor inicial.

El mínimo valor es entonces $f(U) = \sum_{i}^{n}(\gamma_i - \alpha_i)^2$.

Como en (1.70) vimos que $\sum_{i=1}^{n}\beta_i^2 = f(U)$ para alguna matriz ortogonal U, entonces

$$\|B\|_F^2 = \sum_{i=1}^{n}\beta_i^2 \geq \sum_{i=1}^{n}(\alpha_i - \gamma_i)^2.$$

□

La prueba puede extenderse fácilmente a matrices hermíticas. Además, el resultado también es cierto para matrices normales.

Capítulo 2

Pseudoespectro de una matriz

Tradicionalmente, el análisis de los modelos lineales se ha basado en el estudio de los autovalores con un resultado satisfactorio para muchos problemas de matemáticas, ciencias e ingeniería. En especial, las técnicas de autovalores se han aplicado con éxito en campos como la acústica, la mecánica cuántica, la mecánica de fluidos o el análisis numérico. No obstante, la mayoría de los problemas que la teoría de autovalores es capaz de resolver con éxito tienen en común el hecho de que las matrices y operadores que los describen son normales, es decir, poseen una base de autovectores ortogonales.

Cuando una matriz, o un operador, carece de una base de autovectores ortogonales, el estudio de sus autovalores no proporciona una imagen completa de ella. Por ejemplo, la no normalidad puede asociarse con un comportamiento transitorio que difiere totalmente del comportamiento asintótico sugerido por los autovalores, que puede manifestarse, por ejemplo, en la convergencia lenta de procesos iterativos, o en la proximidad a la inestabilidad.

Entre las numerosas herramientas que se han propuesto para describir la no normalidad y analizar sus efectos, se incluyen herramientas clásicas de la teoría de matrices y operadores, como el rango numérico, los ángulos entre subespacios invariantes y los números de condición de los valores propios. En este capítulo vamos a centrarnos en el estudio del pseudoespectro, cuya utilidad se ha probado con éxito en gran variedad de problemas [12].

Vamos a estudiar, a continuación, un problema sencillo que nos servirá como motivación para la definición del pseudoespectro de una matriz.

Supongamos que debemos resolver la siguiente ecuación

$$A\mathbf{u} = z\mathbf{u} + \mathbf{v}, \tag{2.1}$$

donde z no es un autovalor de la matriz A. El objetivo es obtener soluciones que sean estables respecto de pequeñas perturbaciones en $\mathbf{v}$ o en A. Consideramos primero la solución $\mathbf{u}'$ al problema

$$A\mathbf{u}' = z\mathbf{u}' + \mathbf{v}', \tag{2.2}$$

donde suponemos $\|\mathbf{v} - \mathbf{v}'\| < \epsilon$. Entonces, para la diferencia entre las soluciones de (2.2)

y (2.1) se tiene

$$\|\mathbf{u}' - \mathbf{u}\| \leq \epsilon \, \|(zI - A)^{-1}\|,$$

para cualquier norma matricial derivada de una norma vectorial. El valor de $\|(zI-A)^{-1}\|$ puede ser grande, incluso para valores de z alejados del espectro de A.

Si pasamos al estudio de una perturbación de la matriz A, de la forma $A + \epsilon B$ tal que $\|B\| < 1$, y buscamos la solución $\mathbf{u}''$ del problema

$$(A + \epsilon B)\mathbf{u}'' = z\mathbf{u}'' + \mathbf{v},$$

llegamos, aplicando resultados conocidos sobre normas matriciales derivadas de normas vectoriales, a que

$$\|\mathbf{u}'' - \mathbf{u}\| \leq \frac{\epsilon \|(zI - A)^{-1}\|}{1 - \epsilon \, \|(zI - A)^{-1}\|} \|(zI - A)^{-1}\| \, \|\mathbf{v}\|.$$

De nuevo, una buena estimación del error $\|\mathbf{u}'' - \mathbf{u}\|$ requiere que el valor de $\epsilon \, \|(zI - A)^{-1}\|$ sea pequeño.

2.1. Definiciones y equivalencia

En primer lugar, vamos a dar varias definiciones del pseudoespectro de una matriz $A \in \mathcal{M}_{n\times n}(\mathbb{C})$. Posteriormente, probaremos que de hecho, son todas equivalentes, por lo que la definicón de pseudoespectro no es ambigua.

Definición 2.1.1 *Dada una matriz $A \in \mathcal{M}_{n\times n}(\mathbb{C})$, y dado $\epsilon > 0$ arbitrario, se llama $\epsilon-$pseudoespectro de A al conjunto de los $z \in \mathbb{C}$ tales que*

$$\|(zI - A)^{-1}\| > \epsilon^{-1}. \tag{2.3}$$

Denotamos por $\sigma_\epsilon(A)$ al $\epsilon-$pseudoespectro de A.

Para cada $z \in \mathbb{C}$, la matriz $(zI - A)^{-1}$ se llama *matriz resolvente* de A en z y el $\epsilon-$pseudoespectro de A es el subconjunto abierto del plano complejo limitado por la curva de nivel ϵ^{-1} de la norma de la resolvente. A lo largo de todo el capítulo vamos a utilizar el convenio de que

$$\|(zI - A)^{-1}\| = \infty, \quad \text{para } z \in \sigma(A), \tag{2.4}$$

donde $\sigma(A)$ es el espectro de la matriz A. Con dicho convenio, observamos que con esta definición de pseudoespectro de A, se tiene $\sigma(A) \subseteq \sigma_\epsilon(A)$ para todo $\epsilon > 0$.

De manera intuitiva podemos pensar que son los valores z cercanos a un autovalor de la matriz A, aquellos para los que $\|(zI-A)^{-1}\|$ se hace grande. Para una matriz normal, cuando $\|\cdot\| = \|\cdot\|_2$ esta intuición es correcta. No obstante, para matrices que no son normales, y están lejos de serlo, puede ocurrir que la norma $\|(zI - A)^{-1}\|$ sea grande incluso cuando $z \notin \sigma(A)$ y está alejado de dicho conjunto, y más generalmente sucede para matrices que satisfacen $\|A^{-1}\| \gg 1$ o $\kappa(A) = \|A\|\|A^{-1}\| \gg 1$.

La segunda definición que damos de pseudoespectro de una matriz A está basada en la conexión entre la norma de la resolvente y la teoría de perturbación para autovalores.

Definición 2.1.2 *Dada una matriz $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ y dado $\epsilon > 0$ arbitrario, se llama $\epsilon-$pseudoespectro de A al conjunto de los $z \in \mathbb{C}$ tales que*

$$z \in \sigma(A + E) \tag{2.5}$$

para alguna matriz $E \in \mathcal{M}_{n\times n}(\mathbb{C})$ con $\|E\| < \epsilon$.

Es decir, definimos ahora el $\epsilon-$pseudoespectro de A como el conjunto de números complejos que son autovalores de alguna matriz perturbada $A + E$ con $\|E\| < \epsilon$.

De las dos definiciones dadas hasta ahora, se deduce fácilmente que los $\epsilon-$pseudoespectros asociados con diferentes valores de ϵ están relacionados de manera que

$$\sigma_{\epsilon_1}(A) \subseteq \sigma_{\epsilon_1}(A), \text{ si } 0 < \epsilon_1 \leq \epsilon_2, \tag{2.6}$$

y que la intersección de todos los pseudoespectros es el espectro, esto es

$$\cap_{\epsilon>0}\sigma_\epsilon(A) = \sigma(A). \tag{2.7}$$

Continuamos con la tercera definición de pseudoespectro.

Definición 2.1.3 *Dada una matriz $A \in \mathcal{M}_{n\times n}(\mathbb{C})$, y dado $\epsilon > 0$ arbitrario, se llama $\epsilon-$pseudoespectros de A al conjunto de los $z \in \mathbb{C}$ tales que*

$$\|(zI - A)\mathbf{v}\| < \epsilon \tag{2.8}$$

para algún $\mathbf{v} \in \mathbb{C}^n$, con $\|\mathbf{v}\| = 1$.

Con esta definición, cada elemento z del $\epsilon-$pseudoespectro de A se llama *$\epsilon-$pseudoautovalor* de A, y el vector $\mathbf{v}$ correspondiente se llama *$\epsilon-$pseudoautovector* de A.

Probaremos a continuación la equivalencia de las tres definiciones de pseudoespectro dadas en lo que llevamos de sección.

Teorema 2.1.1 *Para toda matriz $A \in \mathcal{M}_{n\times n}(\mathbb{C})$, las Definiciones 2.1.1, 2.1.2 y 2.1.3 para el $\epsilon-$pseudoespectro de A son equivalentes.*

Demostración.

Notamos en primer lugar que para los autovalores $z \in \sigma(A)$, la equivalencia entre las tres definiciones es evidente. Vamos, por tanto, a probarlo para $z \notin \sigma(A)$.

Comenzamos con la implicación $2 \Rightarrow 3$.

Si z es tal que $z \in \sigma(A+E)$ para alguna matriz $E \in \mathcal{M}_{n\times n}(\mathbb{C})$ con $\|E\| < \epsilon$, entonces existe un vector no nulo $\mathbf{v} \in \mathbb{C}^n$, que podemos suponer normalizado, es decir $\|\mathbf{v}\| = 1$, tal que

$$(A + E)\mathbf{v} = z\mathbf{v}. \tag{2.9}$$

Entonces, $\|(zI - A)\mathbf{v}\| = \|E\mathbf{v}\| \leq \|E\|\|\mathbf{v}\| < \epsilon$.

Vamos ahora con $3 \Rightarrow 1$.

Dado $\epsilon > 0$, sea $z \in \mathbb{C}$ tal que $\|(zI - A)\mathbf{v}\| < \epsilon$ para algún $\mathbf{v} \in \mathbb{C}^n$, con $\|\mathbf{v}\| = 1$.

Sea $\mathbf{u} \in \mathbb{C}^n$, el vector con $\|\mathbf{u}\| = 1$, y tal que $(zI - A)\mathbf{v} = s\,\mathbf{u}$. Como $s = \|(zI - A)\,\mathbf{v}\|$, entonces, $s < \epsilon$. Si $s = 0$, es claro que $z \in \sigma(A) \subseteq \sigma_\epsilon(A)$. Si $s \neq 0$, se tiene

$$(zI - A)^{-1}\, s\,\mathbf{u} = \mathbf{v},$$

o, equivalentemente,

$$(zI - A)^{-1}\,\mathbf{u} = s^{-1}\mathbf{v}. \tag{2.10}$$

Tomando normas en (2.10), llegamos a que

$$\|(zI - A)^{-1}\| \geq s^{-1} > \epsilon^{-1}. \tag{2.11}$$

Finalmente, probamos $1 \Rightarrow 2$.

Dado $\epsilon > 0$, sea $z \in \mathbb{C}$ tal que $\|(zI - A)^{-1}\| > \epsilon^{-1}$. Entonces, $\|(zI - A)^{-1}\,\mathbf{u}\| > \epsilon^{-1}$ para algún vector $\mathbf{u}$ con $\|\mathbf{u}\| = 1$. Por lo tanto, tomando $\mathbf{v}$ con $\|\mathbf{v}\| = 1$ tal que

$$(zI - A)^{-1}\mathbf{u} = s^{-1}\mathbf{v}, \tag{2.12}$$

es claro que $s < \epsilon$. Si probamos que existe una matriz E, tal que $\|E\| = s$ y $E\,\mathbf{v} = s\,\mathbf{u}$, entonces tendremos que

$$(A + E)\,\mathbf{v} = z\,\mathbf{v} - s\,\mathbf{u} + s\,\mathbf{u} = z\,\mathbf{v},$$

y habremos probado (2.5), puesto que por (2.12) sabemos que

$$s\mathbf{u} = (zI - A)\mathbf{v} \iff A\mathbf{v} = z\mathbf{v} - s\mathbf{u}.$$

De hecho, E puede tomarse de rango 1, de la forma $E = s\mathbf{u}\mathbf{w}^T$ para algún $\mathbf{w} \in \mathbb{C}^n$ con $\mathbf{w}^T\,\mathbf{v} = 1$. Podemos distinguir dos casos, en función de la norma matricial con la que estemos trabajando.

• Norma $\|\cdot\|_2$: si $\mathbf{w} = \mathbf{v}$ entonces, $\mathbf{v}^T\mathbf{v} = \|\mathbf{v}\|^2 = 1$, y se tiene $E\mathbf{v} = s\mathbf{u}\mathbf{v}^T\mathbf{v} = s\mathbf{u}$.

• Norma $\|\cdot\|$ general: la existencia de un vector $\mathbf{w}$ que satisface las condiciones pedidas puede interpretarse como la existencia de una función lineal $\mathcal{L}$ en $\mathbb{C}^n$ con $\|\mathcal{L}\,\mathbf{v}\| = 1$, $\|\mathcal{L}\| = 1$. El resultado está garantizado por el teorema de Hahn-Bannach [8].

□

Hasta ahora, hemos tratado la definición de pseudoespectro tomando una norma matricial $\|\cdot\|$ arbitraria derivada de una norma vectorial. No obstante, podemos preguntarnos por las propiedades que se tienen si trabajamos con la norma $\|\cdot\|_2$. Antes de nada, notamos que, para poder hacer esta elección, debemos restringirnos al caso en que $\mathbb{C}^n$ está dotado del producto interno

$$(\mathbf{u}, \mathbf{v}) = \mathbf{u}^*\,\mathbf{v},$$

y $\|\cdot\|_2$ es la norma matricial asociada a la norma vectorial definida por

$$\|\mathbf{v}\| = \sqrt{\mathbf{v}^*\,\mathbf{v}}.$$

De las asignaturas de Análisis Numérico cursadas durante el grado, son conocidos algunos resultados sobre esta norma matricial. Entre ellos, podemos destacar el siguiente

$$\|A\|_2 = \mathrm{s}_{\max}(A),$$

donde $\mathrm{s}_{\max}(A)$ denota el máximo valor singular [6] de la matriz A. Es claro, por tanto, que

$$\|A^{-1}\|_2 = [\mathrm{s}_{\min}(A)]^{-1}.$$

En el contexto en que estamos interesados, dada una matriz A, y $z \in \mathbb{C}$, entonces

$$\|(zI - A)^{-1}\|_2 = [\mathrm{s}_{\min}(zI - A)]^{-1}.$$

Definición 2.1.4 *Dada una matriz $A \in \mathcal{M}_{n\times n}(\mathbb{C})$, y dado $\epsilon > 0$ arbitrario, si consideramos la norma matricial $\|\cdot\|_2$, el $\epsilon-$pseudoespectro de A es el conjunto de los $z \in \mathbb{C}$ tales que*

$$s_{min}(zI - A) < \epsilon, \tag{2.13}$$

donde $s_{min}(zI - A)$, es el mínimo valor singular de $zI - A$.

La equivalencia con las definiciones 2.1.1, 2.1.2 y 2.1.3 anteriores, aplicadas al caso en que consideramos la norma matricial $\|\cdot\|_2$ es clara. Por un lado, sabemos que

$$\|(zI - A)^{-1}\|_2 = \frac{1}{s_{\min}(zI - A)}, \tag{2.14}$$

lo que nos da directamente la equivalencia de la Definición 2.1.4 con la Definición 2.1.1. Además, en la demostración del Teorema 2.1.1, vemos que la matriz E puede tomarse de rango 1 como $E = s\,\mathbf{u}\mathbf{v}^T$, donde s es el menor valor singular de la matriz $zI - A$, y $\mathbf{u}$, $\mathbf{v}$ son los vectores singulares asociados por la izquierda y por la derecha, respectivamente.

Una vez establecidas las diferentes definiciones de pseudoespectro de una matriz, así como la equivalencia entre todas ellas, en el siguiente resultado se recogen algunas de las propiedades más importantes, y elementales, del pseudoespectro.

Teorema 2.1.2 (Propiedades del pseudoespectro) *Sea $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ y sea $\epsilon > 0$. Entonces,*

(i) El $\epsilon-$pseudoespectro de la matriz A, $\sigma_\epsilon(A)$, es un conjunto no vacío, abierto y acotado del plano complejo. Además, $\sigma_\epsilon(A)$ tiene, como máximo, n componentes conexas y cada una contiene uno o más autovalores de A.

(ii) Si $\|\cdot\| = \|\cdot\|_2$, entonces $\sigma_\epsilon(A^*) = \overline{\sigma_\epsilon(A)}$.

(iii) Si $\|\cdot\| = \|\cdot\|_2$, entonces $\sigma_\epsilon(A_1 \oplus A_2) = \sigma_\epsilon(A_1) \cup \sigma_\epsilon(A_2)$.

(iv) Para cualquier $c \in \mathbb{C}$, $\sigma_\epsilon(A + cI) = c + \sigma_\epsilon(A)$.

(v) Para cualquier $\beta \in \mathbb{C}$ no nulo, $\sigma_{|\beta|\epsilon}(\beta A) = \beta\,\sigma_\epsilon(A)$.

En (ii), la matriz A^* denota la transpuesta conjugada de A. Por otro lado, en (iii), $A_1 \oplus A_2$ denota la suma directa de dos matrices cuadradas de igual orden, definida por

$$A_1 \oplus A_2 = \left(\begin{array}{c|c} A_1 & O \\ \hline O & A_2 \end{array}\right)$$

Demostración

(*i*) La demostración, cuyo contenido se escapa de los objetivos de este trabajo, se puede encontrar en [2].

(*ii*) Sea $\|\cdot\| = \|\cdot\|_2$, y sea $z \in \sigma_\epsilon(A)$. De la Definición 2.1.2, sabemos que existe una matriz E, con $\|E\|_2 < \epsilon$ tal que $z \in \sigma(A+E)$. Tomando la traspuesta conjugada tenemos que

$$\overline{A}^T + \overline{E}^T = (\overline{A+E})^T,$$

luego

$$\sigma(\overline{A}^T + \overline{E}^T) = \sigma(\overline{A+E}) = \sigma((\overline{A+E})^T) = \overline{\sigma(A+E)},$$

y $\overline{z} \in \sigma_\epsilon(\overline{A}^T)$, ya que $\|\overline{E}^T\|_2 = \|E\|_2 < \epsilon$.

(*iii*) Tomando la norma $\|\cdot\|_2$, la igualdad se deduce directamente de

$$\sigma(A_1 \oplus A_2) = \sigma(A_1) \cup \sigma(A_2). \tag{2.15}$$

Probamos la igualdad (2.15) por doble contención. En primer lugar, sean λ y μ autovalores de A_1 y A_2 respectivamente. Entonces, existen sendos vectores propios asociados, $\mathbf{u}_1$, $\mathbf{u}_2 \in \mathbb{C}^n$ tales que

$$A_1\,\mathbf{u}_1 = \lambda\,\mathbf{u}_1,$$
$$A_2\,\mathbf{u}_2 = \mu\,\mathbf{u}_2.$$

Entonces,

$$(A_1 \oplus A_2)\begin{pmatrix}\mathbf{u}_1\\ \hline \mathbf{0}\end{pmatrix} = \left(\begin{array}{c|c} A_1 & O \\ \hline O & A_2\end{array}\right)\begin{pmatrix}\mathbf{u}_1\\ \hline \mathbf{0}\end{pmatrix} = \begin{pmatrix}\lambda\mathbf{u}_1\\ \hline \mathbf{0}\end{pmatrix} = \lambda\begin{pmatrix}\mathbf{u}_1\\ \hline \mathbf{0}\end{pmatrix},$$

y

$$(A_1 \oplus A_2)\begin{pmatrix}\mathbf{0}\\ \hline \mathbf{u}_2\end{pmatrix} = \left(\begin{array}{c|c} A_1 & O \\ \hline O & A_2\end{array}\right)\begin{pmatrix}\mathbf{0}\\ \hline \mathbf{u}_2\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \hline \mu\mathbf{u}_2\end{pmatrix} = \mu\begin{pmatrix}\mathbf{0}\\ \hline \mathbf{u}_2\end{pmatrix},$$

de forma que los vectores $\mathbf{w}_1$ y $\mathbf{w}_2 \in \mathbb{C}^{n+n}$ definidos por

$$\mathbf{w}_1 = \begin{pmatrix}\mathbf{u}_1\\ \hline \mathbf{0}\end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix}\mathbf{0}\\ \hline \mathbf{u}_2\end{pmatrix},$$

son autovectores de la matriz $A_1 \oplus A_2$ con autovalores asociados λ y μ, respectivamente.

Vamos ahora con la segunda contención. Sea $\gamma \in \sigma(A_1 \oplus A_2)$, entonces existe un vector $\mathbf{w} \in \mathbb{C}^{n+n}$ tal que

$$\left(\begin{array}{c|c} A_1 & O \\ \hline O & A_2\end{array}\right)\begin{pmatrix}\mathbf{w}_1\\ \hline \mathbf{w}_2\end{pmatrix} = \gamma\begin{pmatrix}\mathbf{w}_1\\ \hline \mathbf{w}_2\end{pmatrix},$$

donde $\mathbf{w}_1$ y $\mathbf{w}_2$ denotan los vectores dados por las primeras n componentes, y las siguientes n, respectivamente, de $\mathbf{w}$, y n es la dimensión de las matrices A_1 y A_2. Entonces, es claro que $\gamma \in \sigma(A_1) \cup \sigma(A_2)$.

(*iv*) y (*v*) Vamos a demostrar las dos últimas propiedades probando directamente, por doble contención, la igualdad

$$\sigma_{|\beta|\,\epsilon}(\beta A + cI) = c + \beta\sigma_\epsilon(A).$$

Antes de comenzar con la prueba, notamos que podemos suponer $\beta \neq 0$, pues en otro caso, los conjuntos se reducen a $\{c\}$.

En primer lugar, sea $z \in \sigma_\epsilon(A)$. Por la Definición 2.1.2, existe una matriz E con $\|E\| < \epsilon$, tal que $z \in \sigma(A+E)$. Entonces, $cI + \beta z \in \sigma(cI + \beta A + \beta E)$, donde $\|\beta E\| \leq |\beta|\|E\| < |\beta|\epsilon$. De nuevo por la Definición 2.1.2, se sigue que

$$cI + \beta z \in \sigma_{|\beta|\epsilon}(cI + \beta A).$$

Para ver la segunda contención, sea $\omega \in \sigma_{|\beta|\epsilon}(cI + \beta A)$. Por la Definición 2.1.2 existe una matriz F con $\|F\| < \epsilon|\beta|$ tal que $\omega \in \sigma(cI + \beta A + F)$.

Entonces, $\omega - c \in \sigma(\beta A + F)$, de donde se sigue que

$$\frac{\omega - c}{\beta} \in \sigma(A + \frac{1}{\beta}F),$$

y

$$\|\frac{1}{\beta}F\| = \frac{1}{|\beta|}\|F\| < \epsilon.$$

Por la Definición 2.1.2, se tiene que $z = \frac{1}{\beta}(\omega - c) \in \sigma_\epsilon(A)$, y entonces, llegamos a que

$$\omega = c + \beta z \in c + \beta\sigma_\epsilon(A).$$

□

2.2. Matrices Normales

Antes de enunciar el resultado principal de esta sección sobre el pseudoespectro de una matriz normal, vamos a recordar algunos conceptos y resultados sobre matrices unitarias y matrices normales.

Definición 2.2.1 *Se dice que una matriz $U \in \mathcal{M}(\mathbb{C})$ es unitaria, si su inversa coincide con su traspuesta conjugada. Es decir,*

$$U^* = U^{-1}.$$

De la Definición 2.2.1 se deduce fácilmente la invariancia del pseudoespectro frente a transformaciones de semejanza unitaria.

Lema 2.2.1 *Sea A una matriz compleja, y sea U una matriz unitaria. Entonces, considerando la norma matricial $\|\cdot\| = \|\cdot\|_2$,*

$$\sigma_\epsilon(A) = \sigma_\epsilon(UAU^*), \tag{2.16}$$

para todo $\epsilon > 0$.

Demostración

Si U es una matriz unitaria, entonces tenemos que

$$(zI - UAU^*) = (zUU^* - UAU^*) = U(zI - A)U^*.$$

Por lo tanto,

$$(zI - UAU^*)^{-1} = [U(zI - A)U^*]^{-1} = U(zI - A)^{-1}U^*. \tag{2.17}$$

De (2.17) y del hecho de que la norma espectral es invariante por transformaciones unitarias,deducimos que

$$\|(zI - UAU^*)^{-1}\|_2 = \|U(zI - A)^{-1}U^*\|_2 = \|(zI - A)^{-1}\|_2, \quad z \in \mathbb{C}. \tag{2.18}$$

De (2.18), utilizando la Definición 2.1.1 para el pseudoespectro, se sigue directamente la igualdad (2.16).

□

Definición 2.2.2 *Se dice que una matriz $A \in \mathcal{M}(\mathbb{C})$ es normal, si tiene un conjunto completo de autovectores ortonormales. Es decir, si es diagonalizable y la matriz de autovectores es unitaria.*

$$U^* = U^{-1}.$$

De la Definición 2.2.2 se deduce el siguiente resultado.

Lema 2.2.2 *Sea A una matriz normal, y sea V matriz unitaria de autovectores de A. Entonces, considerando la norma matricial $\|\cdot\| = \|\cdot\|_2$,*

$$\sigma_\epsilon(A) = \sigma_\epsilon(V^*AV) = \sigma_\epsilon(D) = \sigma_\epsilon \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}, \tag{2.19}$$

para todo $\epsilon > 0$.

Demostración

Si A es una matriz normal, entonces existe una matriz de autovectores de A, que es unitaria. Es decir,

$$D = V^*AV = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix},$$

donde V es una matriz unitaria. Por lo tanto, aplicando el Lema 2.2.1, llegamos a la igualdad (2.19).

□

Sea A una matriz normal con autovalores $\{\lambda_1, \ldots, \lambda_n\}$. Dado $\epsilon > 0$, el $\epsilon-$pseudoespectro de A está dado por

$$\sigma_\epsilon(A) = \bigcup_{i=1}^{n} B(\lambda_i, \epsilon), \tag{2.20}$$

donde $B(\lambda_i, \epsilon)$ denota la bola abierta del plano complejo, con centro λ_i y radio ϵ.

De manera equivalente, se cumple que

$$\|(zI - A)^{-1}\|_2 = \|(zI - V)^{-1}\|_2 = \left[s_{\min}(zI - D)\right]^{-1} = \frac{1}{\text{dist}(z, \sigma(A))}, \tag{2.21}$$

donde $\text{dist}(z, \sigma_(A))$ denota la distancia usual de un punto z, al conjunto $\sigma_(A)$ en el plano complejo.

En el enunciado y la demostración del teorema sobre pseudoespectro de matrices normales, vamos a utilizar la siguiente notación.

- $\Delta_\epsilon := B(0, \epsilon) = \{z \in \mathbb{C} : \ |z| < \epsilon\}$.
- $\sigma(A) + \Delta_\epsilon = \{z \in \mathbb{C} : \ z = z_1 + z_2, \ z_1 \in \sigma(A), \ |z_2| < \epsilon\}$
  $= \{z \in \mathbb{C} : \text{dist}(z, \sigma(A)) < \epsilon\}$.

Teorema 2.2.1 (Pseudoespectro de una matriz normal) *Sea A una matriz compleja $n \times n$. Entonces*

$$\sigma(A) + \Delta_\epsilon \subseteq \sigma_\epsilon(A), \tag{2.22}$$

para todo $\epsilon > 0$. Además, si A es una matriz normal y tomamos $\|\cdot\| = \|\cdot\|_2$, se tiene la igualdad

$$\sigma(A) + \Delta_\epsilon = \sigma_\epsilon(A), \tag{2.23}$$

para todo $\epsilon > 0$. Recíprocamente, si $\|\cdot\| = \|\cdot\|_2$ y $\sigma(A) + \Delta_\epsilon = \sigma_\epsilon(A)$, etonces A es una matriz normal.

Demostración

Vamos a demostrar la primera contención (2.22).

Sea z un autovalor de la matriz A y $\delta \in \Delta_\epsilon$ Entonces, $z + \delta$ es autovalor de $A + \delta I$ y basta tomar $E = \delta I$ en la Definición 2.1.2 para ver que $z + \delta \in \sigma_\epsilon(A)$ y que, en consecuencia, $\sigma(A) + \Delta_\epsilon \subseteq \sigma_\epsilon(A)$.

Pasamos a demostrar ahora que si A es una matriz normal, entonces se tiene la igualdad (2.23).

Si A es una matriz normal, y $\{\lambda_1, \ldots, \lambda_n\}$ son sus autovalores, entonces, ya hemos probado en (2.21) que

$$\|(zI - A)^{-1}\|_2 = \frac{1}{\text{dist}(z, \sigma(A))}.$$

De la Definición 2.1.1 se sigue que $z \in \sigma_\epsilon(A)$ si, solo si, $\text{dist}(z, \sigma(A)) < \epsilon$. Con lo que queda probada la igualdad (2.23).

Finalmente, la demostración del recíproco se puede encontrar en [2]. La prueba requiere una colección de resultados previos que no ha parecido oportuno incluir en la memoria.

□

Teorema 2.2.2 (Teorema de Bauer-Fike) *Sea A una matriz compleja diagonalizable, y sea H una matriz de autovectores de A. Entonces,*

$$\sigma(A) + \Delta_\epsilon \subseteq \sigma_\epsilon(A) \subseteq \sigma(A) + \Delta_{\epsilon\kappa_2(H)}, \tag{2.24}$$

para todo $\epsilon > 0$.

Demostración

La primera contención ya ha sido demostrada en el Teorema 2.2.1.

Veamos ahora que $\sigma_\epsilon(A) \subseteq \sigma(A) + \Delta_{\epsilon\kappa_2(H)}$. Sea $z \in \sigma_\epsilon(A)$. Por la Definición 2.1.1, sabemos que

$$\begin{aligned}\frac{1}{\epsilon} < \|(zI - A)^{-1}\|_2 &= \|H\,(zI - V)^{-1}\,H^{-1}\|_2 \leq \kappa_2(H)\|(zI - V)^{-1}\|_2 \\ &= \frac{\kappa_2(H)}{\mathrm{s}_{\min}(zI - V)^{-1}} = \frac{\kappa_2(H)}{\mathrm{dist}(z, \sigma(A)}.\end{aligned}$$

Por lo tanto,

$$\mathrm{dist}(z, \sigma(A) < \epsilon\kappa_2(H),$$

y

$$z \in \sigma(A) + \Delta_{\epsilon\kappa_2(H)}.$$

□

2.3. Algunos ejemplos de pseudoespectros

En esta sección se va a calcular el pseudoespectro de algunas matrices, utilizando el programa Matlab.

En [12] se describen varios procedimientos para representar el $\epsilon-$pseudoespectro de una matriz. En este trabajo hemos utilizado solo uno de ellos, que consiste en tomar un mallado suficientemente fino de una región adecuada del plano complejo y obtener, para cada nodo z del mismo, el menor valor singular que proporciona la descomposición en valores singulares de la matriz $zI - A$ (se ha utilizado la función `svd` de Matlab). Por la Definición 2.1.4, sabemos que

$$\sigma_\epsilon(A) = \{z \in \mathbb{C} : s_{\min}(zI - A) < \epsilon\},$$

y para mostrar el contorno de $\sigma_\epsilon(A)$ para los valores de ϵ en los que estamos interesados se han obtenido las correspondientes curvas de nivel de la superficie que resulta de representar en $\mathbb{R}^3$ la función $s_{\min}(zI - A)$ frente a z.

Ejemplo 2.1.

En primer lugar, consideramos la matriz compleja A (ver [12]) de dimensiones 3×3 dada por

$$A = \begin{pmatrix} -1 & 6 & 0 \\ 0 & i & 8 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}. \tag{2.25}$$

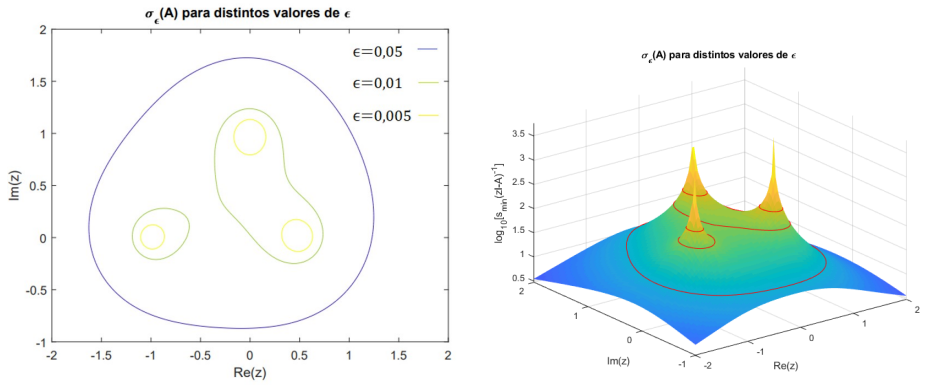

Figura 2.1: $\sigma_\epsilon(A)$ para A definida por (2.25) y valores de $\epsilon = 5 \times 10^{-2}, 10^{-2}, 5 \times 10^{-3}$.

La matriz A no es normal y sus autovalores están dados por $\{-1, i, \frac{1}{2}\}$. El resultado del cálculo numérico del $\epsilon-$pseudoespectro de A, para los valores de $\epsilon = 5 \times 10^{-2}$, 10^{-2}, 5×10^{-3} está representado en la Figura 2.1. En la gráfica izquierda se puede ver el contorno de $\sigma_\epsilon(A)$ para los tres valores de ϵ considerados. En la gráfica derecha aparece la representación tridimensional del $\epsilon-$pseudoespectro de A, utilizando escala logarítmica en el eje vertical, y en rojo se han destacado las curvas de nivel que delimitan $\sigma_\epsilon(A)$ para los mismos valores de ϵ.

Con este ejemplo tan sencillo podemos comprobar que, de acuerdo con el Teorema 2.2.1, como la matriz A no es normal se tiene

$$\sigma_\epsilon(A) \neq \sigma(A) + \Delta_\epsilon.$$

Es decir, el $\epsilon-$pseudoespectro de A no está formado únicamente por la unión de las bolas abiertas del plano complejo con centro λ_i y radio ϵ.

Ejemplo 2.2.

Tomamos, como en la Subsección 1.3.4, la matriz A_n definida por

$$A_n = \begin{pmatrix} n & n & 0 & \cdots & 0 & 0 \\ 0 & n-1 & n & \cdots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & n \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \tag{2.26}$$

para $n = 20$. Los autovalores de A_{20} son $\{1, 2, \ldots, 20\}$, están dibujados en la gráfica izquierda de la Figura 2.2 con puntos negros. Tal como se comprobó en el Capítulo 1, podemos ver en la gráfica derecha de la misma figura que todos los autovalores de la matriz de Wilkinson están mal acondicionados. En concreto, los autovalores intermedios (10 y 11) son los que están peor acondicionados. El máximo de la superficie de la Figura 2.2 está centrado en ellos.

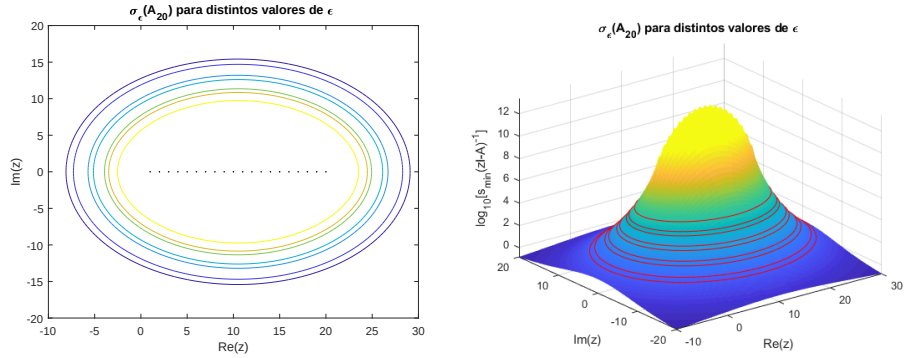

Figura 2.2: $\sigma_\epsilon(A_{20})$ para la matriz A_{20} definida por (2.26) y valores de $\epsilon = 10^{-1}, 5\times10^{-2}, 10^{-2}, 5\times10^{-3}, 10^{-3}, 5\times10^{-4}, 10^{-4}$.

Ejemplo 2.3.

Tomamos, de nuevo como en la Subsección 1.3.4, la matriz C_n definida por

$$C_n = \begin{pmatrix} n & n-1 & n-2 & \cdots & 3 & 2 & 1 \\ n-1 & n-1 & n-2 & \cdots & 3 & 2 & 1 \\ 0 & n-2 & n-2 & \cdots & 3 & 2 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 2 & 2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 1 \end{pmatrix}. \tag{2.27}$$

para $n = 12$. En este caso, comprobamos en el Capítulo 1 que los autovalores más grandes de la matriz (2.27) están bien acondicionados. No ocurre lo mismo para los autovalores más pequeños. Se puede observar una clara correspondencia entre la Figura 1.5 y la gráfica izquierda de la Figura 2.3 que permite comprender mejor el concepto de pseudoespectro de una matriz y su utilidad.

En la gráfica derecha de la Figura 2.3, en la que aparece la representación tridimensional del $\epsilon-$pseudoespectro de C_{12}, para los valores de $\epsilon = 10^{-1}, 5\times10^{-2}, 10^{-2}, 5\times10^{-3}, 10^{-3}, 5\times10^{-4}, 10^{-4}$, y en rojo se destacan las curvas de nivel que delimitan $\sigma_\epsilon(A)$ para los mismos valores de ϵ, vemos claramente que a medida que disminuye el tamaño ϵ de la perturbación, son los autovalores más pequeños los que pueden sufrir una mayor variación.

Ejemplo 2.4.

Tomamos como ejemplo ahora la matriz de Toeplitz tomada de [12], de orden $n = 64$.

$$T_{64} = \begin{pmatrix} 0 & 1 & & & \\ \frac{1}{4} & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{4} & 0 & 1 \\ & & & \frac{1}{4} & 0 \end{pmatrix}, \tag{2.28}$$

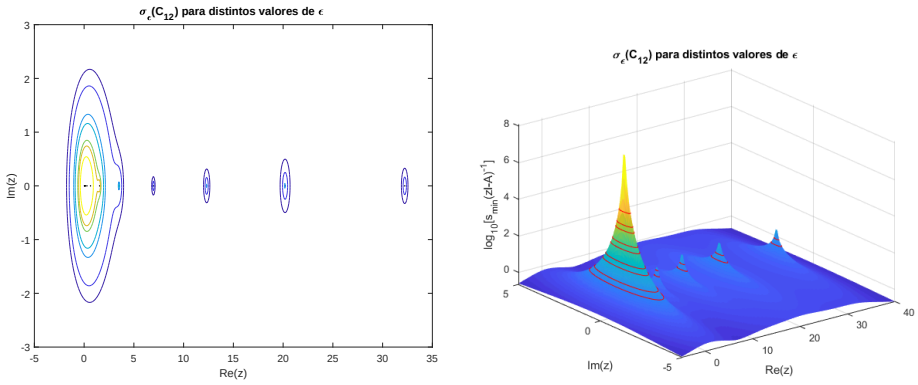

Figura 2.3: $\sigma_\epsilon(C_{12})$ para la matriz C_{12} definida por (2.27) y valores de $\epsilon = 10^{-1}, 5\times10^{-2}, 10^{-2}, 5\times10^{-3}, 10^{-3}, 5\times10^{-4}, 10^{-4}$.

La matriz (2.28) puede simetrizarse mediante una transformación de semejanza. Para un orden n arbitrario, sea $D = \mathrm{diag}(2, 4, \ldots, 2^n)$. Entonces

$$S = DT_nD^{-1} = \begin{pmatrix} 0 & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{2} & 0 & \frac{1}{2} \\ & & & \frac{1}{2} & 0 \end{pmatrix}. \tag{2.29}$$

Los autovalores de T_n y de S coinciden y están dados por

$$\lambda_i = \cos\left(\frac{k\pi}{n+1}\right), \qquad 1 \le i \le n,$$

es decir, el espectro de T_n está dado por un conjunto de n números reales en el intervalo $(-1, 1)$. Por su parte, el $\epsilon-$pseudoespectro de T_n consiste en la región del plano complejo delimitada por la elipse dada por la imagen de $|z| = \epsilon^{\frac{1}{n}}$ por la aplicación $f(z) = z^{-1} + 1/4z$. En la gráfica izquierda de la Figura 2.4 se observa la frontera de $\sigma_\epsilon(T_{64})$ para los valores de $\epsilon = 10^{-1}, 5\times10^{-2}, 10^{-2}, 5\times10^{-3}, 10^{-3}, 5\times10^{-4}, 10^{-4}$. Los puntos negros denotan los autovalores de la matriz T_{64}.

En la gráfica derecha de la Figura 2.4, en la que, de nuevo, aparece la representación tridimensional del $\epsilon-$pseudoespectro de T_{64} para los valores de ϵ considerados, y en rojo se destacan las curvas de nivel que delimitan $\sigma_\epsilon(A)$ para los mismos valores de ϵ, apreciamos que los autovalores de T_{64} están muy mal acondicionados. En este caso, el máximo de la superficie se alcanza en todos los autovalores de T_{64}. Para comprobar el mal acondicionamiento de la matriz T_n general, podemos considerar una perturbación de T_n que consiste en sumar una cantidad ϵ al elemento $(n, 1)$. Los autovalores de la matriz

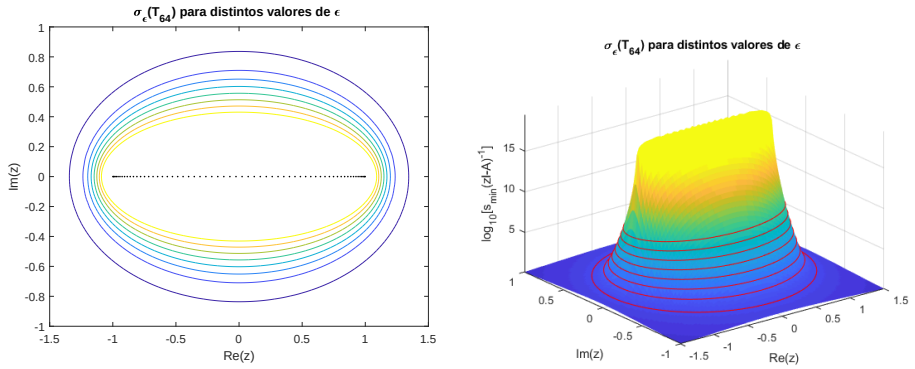

Figura 2.4: $\sigma_\epsilon(T_{64})$ para la matriz T_{64} definida por (2.28) y valores de $\epsilon = 10^{-1}, 5\times10^{-2}, 10^{-2}, 5\times10^{-3}, 10^{-3}, 5\times10^{-4}, 10^{-4}$.

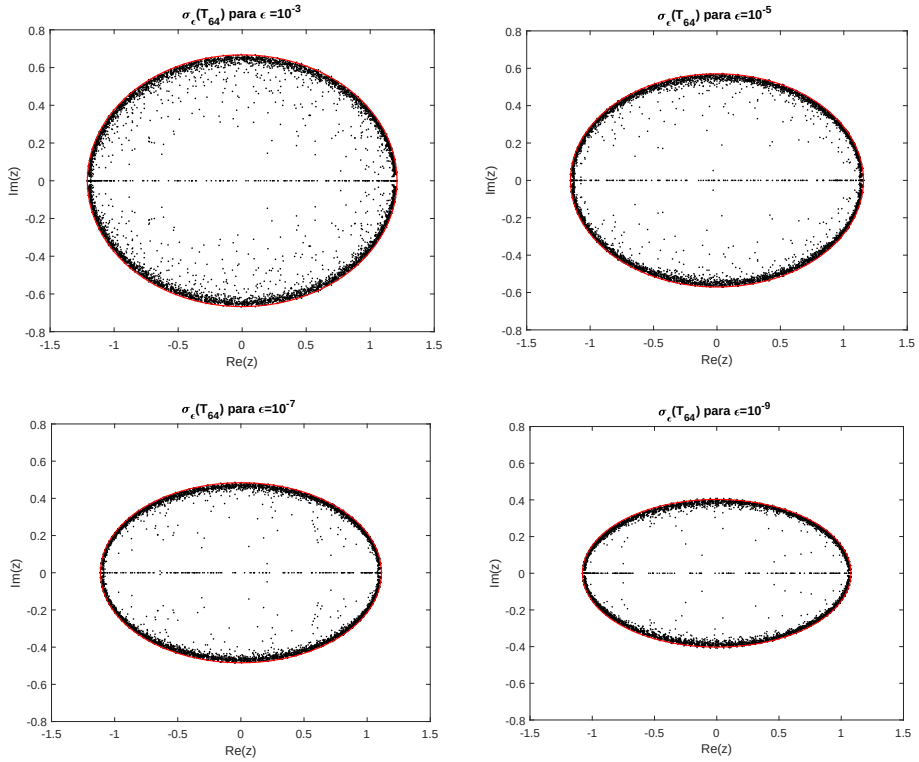

Figura 2.5: Autovalores de $N = 100$ matrices $T_{64} + \epsilon B$, para valores de $\epsilon = 10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}$. Podemos visualizar el $\sigma_\epsilon(T_{64})$.

perturbada coinciden con los de la matriz

$$D(T_n + \epsilon B)D^{-1} = \begin{pmatrix} 0 & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{2} & 0 & \frac{1}{2} \\ 2^{n-1}\epsilon & & & \frac{1}{2} & 0 \end{pmatrix}.$$

Entonces, los autovalores de $T_n + \epsilon B$ coinciden con los de una matriz obtenida al añadir la perturbación $2^{n-1}\epsilon$ a un elemento de la matriz simétrica DT_nD^{-1}. Acabamos de comprobar que el pseudoespectro permite obtener una caracterización más completa de las matrices que la que da el espectro de la misma.

Utilizando la Definición 2.1.2, podemos describir otra forma de obtener una aproximación al $\epsilon-$pseudoespectro de una matriz [12]. El método consiste en generar un número N suficientemente grande de matrices aleatorias B tales que $|b_{i,j}| < 1$, $1 \leq i, j \leq n$, y representar en el plano complejo los autovalores de las N matrices perturbadas $A + \epsilon B$ resultantes. Este procedimiento se ha utilizado con la matriz T_{64} definida en (2.28) para generar la Figura 2.5. Los puntos representados en ella no deben salirse de la región delimitada por las correspondientes curvas de nivel que aparecen destacadas con una línea roja en la Figura 2.5.

Bibliografía

[1] T. M. Apóstol, *Análisis Matemático*, Editorial Reverté, 1976.

[2] G. Armentia (2015), *Pseudoespectros de matrices*. Tesis doctoral, Universidad del País Vasco.

[3] P.A. Clement, *A class of triple-diagonal matrices for test purposes*, SIAM Review 1, 1959. (pp. 50-52)

[4] R. M. Corless & N. Fillion, *A Graduate Introduction to Numerical Methods (from the Viewpoint of Backward Error Analysis)*, Springer, 2013.

[5] E. Goursat, *Cours d'Analyse Mathématique (Tomo 2)*, Gauthier-Villars, 1942.

[6] R. A. Horn & C. R. Johnson, *Matrix Analysis (2^{nd} Edition)*, Cambridge University Press, 2013.

[7] P. D. Lax, *Linear algebra and its applications*, John Wiley & Sons, 2007.

[8] B. V. Limaye, *Functional Analysis*, Wiley Eastern, 1981.

[9] Y. Ma & A. Edelman, *Nongeneric eigenvalue perturbations of Jordan blocks*, Linear Algebra Appl, Vol. 273, pp. 45-63, 1998.

[10] L. Merino & E. Santos, *Álgebra lineal con métodos elementales*, Ediciones Paraninfo, 2016.

[11] J. W. S. Rayleigh, *The theory of Sound*, Vol. 1, London: Macmillan, 1894. (pp. 114-118)

[12] L. N. Trefethen & M. Embree, *Spectra and pseudoespectra: the behavior of nonnormal matrices and operators*, Princeton University Press, 2005.

[13] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, 1999.

Apéndice A

Apéndice

A.1. Algunos resultados utilizados para la Sección 1.1.1

Sea $f(x, y)$ definida por

$$f(x, y) = y^n + p_{n-1}(x)y^{n-1} + p_{n-2}(x)y^{n-2} + \cdots + p_1(x)y + p_0(x),$$

donde los $p_i(x)$ son polinomios en x para $1 \leq i \leq n-1$. De la teoría de funciones algebraicas conocemos [5] los siguientes resultados.

Teorema A.1.1 (Teorema 1 sobre funciones algebraicas.) *Sea $y_i(0)$ una raíz simple de $f(0, y) = 0$. Entonces existe $\delta_i > 0$ tal que $f(x, y) = 0$ posee una raíz simple $y_i(x)$ definida por*

$$y_i(x) = y_i(0) + p_{i1}x + p_{i2}x^2 + \cdots,$$

donde la serie de la derecha converge para $|x| < \delta_i$.

Nota:

(i) Solo se requiere que la raíz $y_i(0)$ sea simple. No se impone nada sobre la multiplicidad de las $n-1$ raíces restantes de $f(0, y) = 0$.

(ii) La serie del término de la derecha puede ser finita.

(iii) Verifica $y_i(x) \longrightarrow y_i(0)$ cuando $x \longrightarrow 0$

Teorema A.1.2 (Teorema 2 sobre funciones algebraicas.) *Si $y_1(0) = y_2(0) = \cdots = y_m(0)$ es una raíz de multiplicidad m de $f(0, y) = 0$ entonces, existe $\delta > 0$ tal que cuando $|x| < \delta$, existen m raíces de $f(x, y) = 0$ que se pueden agrupar en r conjuntos con $m_i \quad 1 \leq i \leq r$ raíces cada uno, de manera que las m_i raíces del i-ésimo grupo son los m_i valores de la serie*

$$y_1(0) + p_{i1}z + p_{i2}z^2 + \cdots$$

correspondientes a los m_i valores de z definidos por

$$z = x^{\frac{1}{m_i}}.$$

A.2. Lemas utilizados en la demostración del Teorema 1.4.10

Lema A.2.1 *Si una matriz real y simétrica X tiene autovalores $\{\lambda_1, \ldots, \lambda_n\}$, entonces*

$$\|X\|_F^2 = \sum_{i=1}^{n} \lambda_i^2.$$

Demostración

Por ser $\{\lambda_1, \ldots, \lambda_n\}$ los autovalores de la matriz simétrica X, existe una matriz U ortogonal tal que $U^T X U = diag(\lambda_1, \ldots, \lambda_n)$. Tomando la norma de Frobenius en la igualdad anterior, utilizando que U y U^T son matrices ortogonales y recordando que la norma de Frobenius es invariante por transformaciones ortogonales, llegamos a

$$\|X\|_F^2 = \|U^T X U\|_F^2 = \|\text{diag}(\lambda_1, \ldots, \lambda_n)\|_F^2 = \sum_{i=1}^{n} \lambda_i^2.$$

□

Lema A.2.2 *Si $\{U_k\}_{k=1}^{\infty}$ es una sucesión infinita de matrices ortogonales, entonces existe una subsucesión $\{U_{s_k}\}_{k=1}^{\infty}$, tal que*

$$lim_{k\to\infty} U_{s_k} = U,$$

con U ortogonal.

Demostración

La prueba de este lema se sigue de la versión n^2-dimensional del *Teorema de Bolzano-Weierstrass* [1], pues todos los elementos de las matrices U_k están entre $[-1, 1]$.

La matriz U debe ser ortogonal, pues

$$U_{s_k}^T U_{s_k} = I, \qquad k \geq 1,$$

y tomando límite llegamos a

$$\lim_{i\to\infty} U_{s_k}^T U_{s_k} = U^T U = I.$$

□