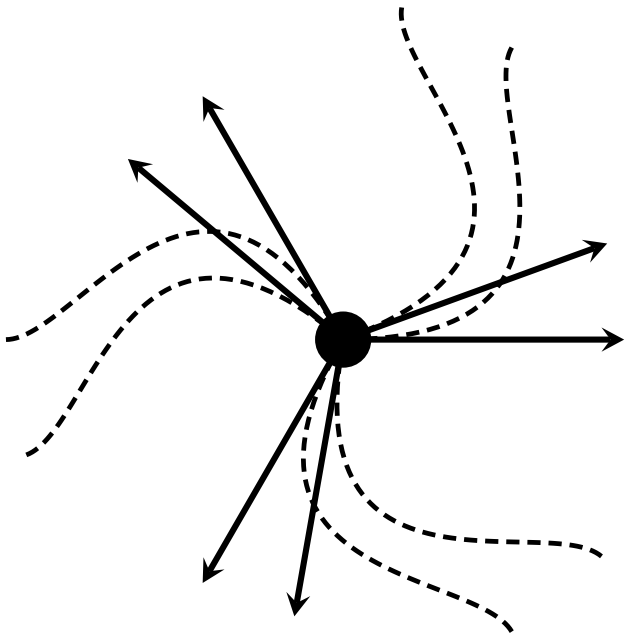
Tesis doctoral

Condicionamiento y alta precisión en problemas espectrales estructurados



Autor: María José Peláez Montalvo
Director: Julio Moro Carreño

Condicionamiento y alta precisión en problemas espectrales estructurados

Memoria que se presenta para optar al título de
Doctor en Matemáticas
por la Universidad Carlos III de Madrid

Autor: María José Peláez Montalvo
Director: Julio Moro Carreño

Programa de doctorado en Ingeniería Matemática
Departamento de Matemáticas
Escuela Politécnica Superior
Universidad Carlos III de Madrid
Marzo de 2007

A mis padres y hermanos,

en particular Fran y Rocío.

Agradecimientos

Un trabajo como es una tesis de 4 años es, sin duda, un paseo de conocimiento, aprendizaje y simbiosis de mucha gente. Primero, quiero agradecer a Julio la paciencia y la precisión que me ha enseñado en todo momento; su madurez científica que me ha transmitido, así como sus ánimos en los momentos no productivos y difíciles que se tienen en toda tesis. A Froilán y Juan Manuel, unas sinceras gracias por todo su apoyo y sus conversaciones siempre prácticas.

A Volker Merhmann y Daniel Kressner les agradezco su hospitalidad en mis estancias en Alemania y Suecia y mi más sincera admiración por el fervor que me han mostrado siempre en su visión de la ciencia[1]; a mis compañeros de doctorado por hacer más ameno mi camino, en especial a Alberto Portal por su ayuda siempre presente en temas computacionales y a Luis Lafuente que, aparte de las magníficas discusiones matemáticas mantenidas con él, me ha dado un baño de realidad y una amistad difícil de olvidar.

Y por último debo agradecer a todas esas personas que siempre han confiado en mí: a Juan Antonio, quien me ha enseñado en todo momento a ver la cara positiva; a Juan Pablo, que siempre ha hecho anteponer la razón a la derrota; a Mercedes, quien me ha dado paz y fuerzas en momentos difíciles de esta tesis; a Laura, por sus consejos y esos interesantes partidos de tenis que hemos compartido; a Pilar, quien ha sabido comprenderme y ponerse en mi piel en todo momento; a mi hermano Francisco, que siempre me dijo que mirase hacia delante y a mi madre, quien siempre me ha escuchado. Y, por supuesto, a todos los demás que no puedo nombrar en estas líneas porque, igual que un ordenador trabaja con precisión finita y esto conlleva en ocasiones errores de redondeo, estos agradecimientos también son finitos y con errores. A todos: gracias.

María
Madrid, Marzo del 2007.

[1]To Volker Merhmann and Daniel Kressner go my thanks for their hospitality during my visits to Germany and Sweden, and my most sincere admiration for the passion they have always shown me in their vision of science.

Este trabajo ha sido posible gracias a los proyectos del Ministerio de Educación y Ciencia BFM 2000-0008, BFM 2003-00223 y MTM 2006-05361.

Si he conseguido ver más lejos, es porque me he aupado en hombros de gigantes. No se lo que pareceré a los ojos del mundo, pero a los míos es como si hubiese sido un muchacho que juega en la orilla del mar y se divierte de tanto en tanto encontrando un guijarro más pulido o una concha más hermosa, mientras el inmenso océano de la verdad se extendía, inexplorado, frente a mí.

Isaac Newton

Índice general

1. Introducción **9**
- 1.1. Algoritmos estructurados . . . 9
- 1.2. Algoritmos estructurados y precisión . . . 11
  - 1.2.1. Números de condición estructurados . . . 13
  - 1.2.2. Algoritmos de alta precisión relativa . . . 14
- 1.3. Notación . . . 18

2. Números de condición espectrales estructurados **19**
- 2.1. Introducción . . . 19
  - 2.1.1. Preliminares . . . 23
- 2.2. Perturbaciones estructuradas genéricas: el caso matricial . . . 27
  - 2.2.1. Número de condición de Hölder estructurado para autovalores múltiples 27
  - 2.2.2. Matrices reales . . . 27
  - 2.2.3. Estructuras lineales . . . 29
  - 2.2.4. Matrices de Toeplitz y Hankel . . . 30
  - 2.2.5. Matrices simétricas, antisimétricas y hermíticas . . . 33
  - 2.2.6. Matrices J–simétricas y J–antisimétricas . . . 36
- 2.3. Perturbaciones estructuradas genéricas: problemas generalizados . . . 39
  - 2.3.1. Haces de matrices . . . 39
  - 2.3.2. Polinomios matriciales . . . 49
- 2.4. Perturbaciones estructuradas completamente no genéricas: el caso matricial 52
  - 2.4.1. Estructuras lineales completamente no genéricas . . . 54
  - 2.4.2. Estructuras y antiestructuras lineales . . . 56
  - 2.4.3. Números de condición estructurados vía el diagrama de Newton . . 60
- 2.5. Conclusiones y trabajos futuros . . . 74

3. Algoritmos espectrales de alta precisión relativa para matrices simétricas estructuradas **75**
- 3.1. Preliminares . . . 75
- 3.2. Factorización LDL^T por bloques de matrices simétricas . . . 81
- 3.3. Factorización con alta precisión para matrices simétricas DSTU y TSC . . 87
  - 3.3.1. Matrices DSTU . . . 87
  - 3.3.2. Matrices TSC . . . 96

3.4. Paso de LDL^T a RRD: Análisis de errores 100
3.5. Experimentos numéricos 104
3.5.1. Matrices DSTU 105
3.5.2. Matrices TSC 107
3.6. Conclusiones y trabajos futuros 109

Capítulo 1

Introducción

1.1. Algoritmos estructurados

En los inicios del Álgebra Lineal Numérica, como es natural, los algoritmos se diseñaron a fin de poder ejecutarse sobre la clase más amplia posible de matrices, idealmente sin restricción alguna. Los algoritmos más célebres del cálculo matricial, como eliminación gaussiana o el algoritmo QR se desarrollaron con la idea de aplicarlos a matrices completamente arbitrarias. En el curso de los años, estos algoritmos se han ido perfeccionando gradualmente, incorporando todo tipo de mejoras y modificándose en el sentido de hacerlos más rápidos, más robustos y más precisos. Gracias a ello disponemos a fecha de hoy de una panoplia de algoritmos de carácter general rápidos y fiables, ampliamente testados y que resuelven de manera satisfactoria gran parte de los problemas matriciales que se presentan en la práctica.

Sin embargo, estos algoritmos de carácter general, que llamaremos *convencionales* para distinguirlos de los algoritmos *estructurados*, no son siempre la mejor opción a la hora de resolver un problema: con frecuencia, las matrices u operadores lineales que aparecen en las aplicaciones no son arbitrarios, sino que suelen tener estructuras especiales, bien como consecuencia de las propiedades físicas subyacentes al modelo (simetrías, invariancias,...) o, por ejemplo, debido a la discretización de la que surge el problema lineal que se debe resolver. Si la dimensión n del problema no es demasiado grande y se dispone de un buen algoritmo convencional para resolver el problema, ignorar la estructura puede ser una buena opción. Sin embargo, cuando n es grande, o cuando hay que resolver muchos problemas similares de manera sucesiva, puede ser necesario sacarle partido a la estructura a efectos de reducir el coste computacional del algoritmo empleado.

La estructura de una clase de matrices puede manifestarse de maneras muy distintas. Puede venir dada, por ejemplo, por relaciones de igualdad entre los elementos de la matriz: relaciones de simetría o antisimetría con respecto a la diagonal o a la antidiagonal principal (es el caso de las matrices simétricas, las antisimétricas o las persimétricas), relaciones de igualdad a lo largo de diagonales o de antidigonales (matrices Toeplitz o Hankel), o puede que las repeticiones de elementos sigan otros patrones diferentes (matrices circulantes,

centrosimétricas, etc...). La estructura puede venir dada de forma implícita, como en el caso de sumas (Toeplitz más Hankel, por ejemplo), productos, o inversas de matrices con estructura dada. O puede venir impuesta de modo más sutil, como es el caso de las álgebras de Lie o de Jordan asociadas a productos escalares arbitrarios: un ejemplo paradigmático son las matrices hamiltonianas y las antihamiltonianas, que son, respectivamente, el álgebra de Lie y el álgebra de Jordan correspondientes al producto escalar indefinido asociado a la matriz

$$J_n = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}.$$

Otras clases de matrices estructuradas aparecen de manera natural en problemas de interpolación y aproximación de funciones (matrices de Cauchy, de Vandermonde, de Vandermonde polinómicas, de Pick,...), o al discretizar ecuaciones diferenciales (matrices banda: tridiagonales, pentadiagonales,...). También son relevantes en la práctica las estructuras dadas por la distribución de elementos nulos y no nulos en la matriz, o aquellas con gran número de elementos nulos en posiciones más o menos prefijadas (las matrices *sparse* en la terminología inglesa, que, a falta de mejor traducción, llamaremos huecas). El auge de los métodos de Krylov, una de las áreas de investigación más activas del Álgebra Lineal Numérica en los últimos años, se debe a que para matrices de dimensiones muy grandes, los algoritmos convencionales, típicamente de orden n^3, son demasiado costosos y deben buscarse algoritmos más baratos desde el punto de vista computacional. Para matrices huecas, un método de Krylov con un buen precondicionador es una alternativa mucho más rápida que los métodos convencionales.

En general, diremos que un conjunto de matrices $n \times n$ es una *clase estructurada de matrices* si existe un número $t < n^2$ tal que cualquier matriz del conjunto se puede describir como función de t parámetros o menos. Obviamente, el hecho de que las matrices dependan de un número pequeño de parámetros abre la posibilidad de diseñar algoritmos más rápidos que los convencionales. Uno de los ejemplos más familiares es el de las matrices simétricas: el hecho de que una matriz simétrica depende de aproximadamente la mitad de parámetros que una matriz no simétrica permite reducir el coste de eliminación gaussiana a la mitad en su versión simétrica, el algoritmo de Cholesky, y permite reducir a bastante menos de la mitad el coste del algoritmo QR de autovalores. Por tanto, hay un primer incentivo a la hora de buscar algoritmos estructurados, relacionado con el *coste computacional* del algoritmo y el *almacenamiento* de los datos: si la matriz depende de menos parámetros, las necesidades de almacenamiento se reducen y es probable que se pueda reducir también el número de operaciones que debe ejecutar el algoritmo. Este tipo de reducción es la que con frecuencia lleva a cabo cualquier usuario de MATLAB, aunque no sea consciente de ello: el algoritmo que MATLAB emplea para calcular los autovalores de una matriz, o para resolver un sistema lineal, es distinto según que la matriz sea simétrica o que no lo sea. Cuando MATLAB detecta que la matriz en cuestión es simétrica, emplea la versión estructurada (en este caso, simétrica) del algoritmo, aunque esta elección sea transparente para el usuario.

De manera muy laxa, llamaremos *algoritmo estructurado* a cualquier algoritmo ideado de manera específica para aplicarse a una clase concreta de matrices estructuradas, y que explote de alguna manera las propiedades especiales de dicha clase, bien para reducir el

coste computacional, o con otros propósitos relacionados con la precisión, que explicaremos a continuación.

Con una definición tan vaga no es fácil hacer una historia de los algoritmos matriciales estructurados, en parte porque los analistas numéricos han tenido siempre en cuenta la posibilidad de aprovechar la estructura para acelerar o simplificar sus algoritmos. Sin embargo, no fue hasta los años 80 que esos esfuerzos se convirtieron en una corriente reconocible dentro del cálculo matricial, corriente que llega hasta nuestros días. Como puede verse en la página web

`http://www.math.uconn.edu/ olshevsky/community.php`,

que enumera los integrantes de la *"Structured matrix community"*, muchos investigadores dedican sus esfuerzos a día de hoy al diseño y al análisis de algoritmos estructurados para resolver todo tipo de problemas matriciales. Sin embargo, las técnicas y la terminología que emplean, o los propósitos que rigen su investigación, son muy diversos. Muestra de ello es, por ejemplo, la diversidad de enfoques que se puede apreciar en los capítulos del libro [33]. Esta es otra dificultad añadida a la hora de trazar un panorama que recoja todas las aportaciones de los diferentes enfoques.

Aunque hay algoritmos rápidos anteriores para problemas estructurados, como el de Björck-Pereyra para resolver sistemas lineales de Vandermonde [7], quizá uno de los conceptos más exitosos a la hora de sistematizar el estudio de una amplia clase de problemas matriciales estructurados sea el de *desplazamiento*, introducido primeramente en [36, 53] en conexión con las matrices de Toeplitz, y extendido posteriormente a clases mucho más amplias de matrices y a diversas aplicaciones prácticas. Por dar una idea de la amplitud de problemas que se abarcan, este tipo de técnicas permite hallar algoritmos de orden $O(n^2)$ para resolver sistemas lineales[1] cuyas matrices son, no sólo Toeplitz, sino también productos T_1T_2, inversas T^{-1} u otras combinaciones $T_1 - T_2T_3^{-1}T_4$, $(T_1T_2)^{-1}T_3$, ... de matrices Toeplitz T_i. El concepto de desplazamiento ha dado lugar a todo un área de investigación, dedicada a obtener *algoritmos rápidos para clases de matrices con desplazamiento de rango finito* (véanse [40, 54, 33] para más detalles y para una panorámica de este activo campo de investigación).

1.2. Algoritmos estructurados y precisión

Además de obtener algoritmos más rápidos, hay otros incentivos de peso, relacionados con la *precisión en el cálculo de las soluciones*, para emplear algoritmos estructurados. Es precisamente a estos aspectos de precisión a los que se dedica fundamentalmente la presente memoria. En primer lugar, algunas estructuras imponen fuertes restricciones sobre la naturaleza de las soluciones. Los autovalores de una matriz hamiltoniana real, por ejemplo, son simétricos con respecto a los ejes real y complejo, esto es, si λ es autovalor, entonces también lo son $-\lambda$, $\overline{\lambda}$ y $-\overline{\lambda}$. Si para calcular los autovalores de la matriz aplicamos el

[1] Más en concreto, el coste es $O(rn^2)$, siendo r el rango de desplazamiento de la matriz.

método QR, un algoritmo de Krylov o cualquier otro algoritmo que ignore la estructura hamiltoniana, los autovalores calculados perderán esta simetría debido a la influencia de los errores de redondeo. Esto dificulta, por ejemplo, el identificar los autovalores imaginarios puros de una matriz hamiltoniana, algo que resulta crucial tanto para calcular el radio de estabilidad de una matriz [47] como para calcular la norma H_∞ de sistemas lineales invariantes en el tiempo [9]. Sin embargo, si en lugar de un algoritmo convencional se emplea un algoritmo que respete la estructura hamiltoniana, la simetría de los autovalores se preserva. En resumen, suele ocurrir que las respuestas obtenidas de un algoritmo que preserva la estructura contienen una información cualitativa más acorde con la naturaleza del sistema físico subyacente que las soluciones calculadas por métodos convencionales.

Esta idea puede precisarse en lenguaje matemático por medio del *análisis regresivo de errores*. Por simplicidad, lo haremos, como antes, aplicado al ejemplo de las matrices hamiltonianas reales. Se dice que un algoritmo de autovalores es *regresivamente estable* si, para cada matriz A, los autovalores de A calculados por el algoritmo son los autovalores *exactos* de una matriz ligeramente perturbada

$$A + E \tag{1.1}$$

(esto es, la matriz E es pequeña, por ejemplo en norma, comparada con la matriz A). A la matriz E se le llama matriz de *errores regresivos*. Si para calcular los autovalores de una matriz hamiltoniana real H empleamos un método regresivamente estable (y prácticamente todos los algoritmos convencionales lo son), los autovalores calculados por dicho método serán los autovalores exactos de una matriz perturbada $H+E$, donde la matriz E acumula, por así decirlo, todos los errores debidos al redondeo en las operaciones aritméticas que requiere el proceso de cálculo de autovalores. Si, además, el algoritmo en cuestión preserva la estructura hamiltoniana, el error regresivo E será también hamiltoniano y real. En tal caso, se dice que el método es *regresivamente estable en sentido fuerte* [12] y, al ser la suma $H+E$ también hamiltoniana, la simetría en los autovalores se conservará. Por el contrario, si el método no preserva la estructura, el error regresivo E deja de ser hamiltoniano y, como consecuencia, la simetría se pierde.

Idealmente, un algoritmo estructurado debería cumplir tres condiciones:

- ser regresivamente estable en sentido fuerte,
- ser capaz de calcular los autovalores de cualquier matriz que pertenezca a la clase estructurada en cuestión, y
- no ser más costoso computacionalmente que un método convencional.

Para algunas clases, como la de matrices simétricas reales o la de las antihamiltonianas (véase, por ejemplo, [4, §2.4–2.5]), existen métodos de autovalores que cumplen los tres requisitos. Sin embargo, no siempre es fácil desarrollar métodos que satisfagan las tres condiciones, especialmente la primera. Valga como ejemplo de ello el caso hamiltoniano: la historia de los algoritmos estructurados para el problema hamiltoniano se abre con el

trabajo de Paige y van Loan [75] en los años 80. Desde entonces se han propuesto diversos métodos, ninguno de ellos totalmente satisfactorio: los primeros algoritmos, como el *square-reduced method* de van Loan [65], trabajan de forma implícita con la matriz al cuadrado, por lo que preservan la estructura de la matriz al cuadrado, no la de la matriz original. Esto conlleva calcular los autovalores con la mitad de la precisión con la que lo haría un algoritmo que trabaje directamente con la matriz. Posteriormente, Benner et al. [5] evitan esa pérdida de precisión, aunque preservando igualmente la estructura de la matriz al cuadrado. Finalmente, algoritmos más recientes, como el propuesto en [6], sólo se aplican al caso en que *ninguno* de los autovalores de la matriz es imaginario puro. Por tanto, aunque los algoritmos ya existentes son capaces de resolver de manera satisfactoria gran parte de los problemas que se plantean en la práctica, la cuestión de hallar un algoritmo "ideal" para el problema hamiltoniano de autovalores sigue aún abierta.

1.2.1. Números de condición estructurados

Otra observación importante es que, además de preservar propiedades cualitativas como las simetrías, *es de esperar que un algoritmo regresivamente estable en sentido fuerte sea más preciso que uno convencional*. Ello se debe a que si el método es regresivamente estable en sentido fuerte, el conjunto de posibles errores regresivos E en (1.1) debidos al redondeo se restringe a la clase de matrices estructuradas en cuestión. Al ser menor el conjunto de perturbaciones admisibles, el peor efecto posible sobre la precisión de los autovalores *a causa de errores estructurados* puede ser mucho menor que el peor efecto posible debido a errores arbitrarios, sin estructurar. Esto abre la posibilidad de que un método que preserve la estructura alcance una precisión mucho mayor que un algoritmo convencional, si es que los autovalores son mucho más sensibles a perturbaciones no estructuradas que a las estructuradas. Dado que la sensibilidad de los autovalores de una matriz se mide a través de su *número de condición*, es de la mayor importancia el poder comparar, para las distintas estructuras, el número de condición estructurado y el no estructurado, a fin de detectar aquellos casos en que un algoritmo estructurado pueda ser significativamente más preciso que un algoritmo convencional. Varios autores, como Higham y Higham [42], Rump [81, 82], o Tisseur [91] se han ocupado de esta cuestión, pero *siempre para autovalores simples*. En [42, 91], por ejemplo, se obtienen expresiones explícitas del número de condición estructurado de un autovalor simple cuando la clase estructurada de matrices es un subespacio lineal. Este tipo de resultados fue extendido por Karow, Kressner y Tisseur en [55] a estructuras no lineales bajo ciertas condiciones de regularidad. Rump, por su parte, hace en [81, 82] un estudio exhaustivo del número de condición estructurado de autovalores simples y del pseudoespectro para diversas estructuras, tanto lineales como no lineales. Sorprendentemente, en la mayor parte de los casos investigados, el condicionamiento estructurado difiere poco del no estructurado. Los únicos casos en los que se sabe que puede haber una diferencia significativa son el caso antisimétrico complejo [82], el de matrices con ciertas distribuciones de ceros [73], y el caso simpléctico [55].

Nuestra intención en el capítulo 2 de esta memoria es explorar el caso de autovalores múltiples. En primer lugar, definiremos con todo detalle los números de condición, tanto

estructurados como no estructurados, para autovalores múltiples, posiblemente defectivos de matrices. Para ello nos basaremos en la definición de número de condición propuesta en [72, §4], que a su vez se sigue de la teoría de perturbación de Lidskii [64, 96], que describe los desarrollos asintóticos en potencias fraccionarias que son de esperar cuando se perturba un autovalor múltiple. Una vez dispongamos de un concepto de número de condición estructurado, obtendremos expresiones explícitas del mismo para diversas clases de matrices estructuradas, e incluso para algunas clases estructuradas de *haces de matrices*: dadas dos matrices cuadradas A y B de la misma dimensión, consideraremos el problema de determinar escalares λ y vectores $x \neq 0$ tales que

$$Ax = \lambda Bx. \tag{1.2}$$

También trataremos, aunque muy tangencialmente, los problemas polinómicos de autovalores en la sección 2.3.2.

1.2.2. Algoritmos de alta precisión relativa

En el último capítulo de la memoria nos ocuparemos de ciertos algoritmos estructurados de autovalores que se han venido desarrollando en los últimos veinte años para matrices simétricas, y que podemos englobar bajo la denominación de algoritmos *de alta precisión relativa*. El nombre se debe a que si A es la matriz simétrica cuyos autovalores queremos calcular, E es el error regresivo como en (1.1), y suponemos que el cociente de normas $\|E\|/\|A\|$ es del orden de una cierta cantidad ϵ (por ejemplo, la unidad de redondeo de la aritmética finita empleada por un ordenador), entonces lo más que se puede asegurar para los autovalores calculados es que se obtienen con *alta precisión absoluta*, esto es, que el error absoluto $|\lambda - \widehat{\lambda}|$ verifica

$$|\lambda - \widehat{\lambda}| \leq \eta \, \|A\|,$$

para una cierta cantidad η del orden de ϵ, donde λ es un autovalor exacto y $\widehat{\lambda}$ es la aproximación de λ que nos proporciona el algoritmo. Esto *no garantiza que todos* los autovalores de la matriz se calculen con un mínimo de cifras significativas correctas: la anterior cota absoluta se traduce en la cota relativa

$$\frac{|\lambda - \widehat{\lambda}|}{|\lambda|} \leq \eta \, \frac{|\lambda_{\text{máx}}|}{|\lambda|}$$

donde $\lambda_{\text{máx}}$ es el autovalor de módulo máximo de la matriz. Por tanto, los únicos autovalores para los que se puede garantizar un error relativo de orden ϵ son aquéllos de tamaño similar al de λ_{max}. Sin embargo, hay en las aplicaciones numerosos ejemplos de sistemas con escalas múltiples, cuya matriz tiene autovalores de tamaños muy dispares y en los que, de hecho, son los autovalores más pequeños los que interesa calcular con mayor precisión. En tales situaciones *los algoritmos al uso proporcionan con frecuencia aproximaciones que no concuerdan en ninguna cifra significativa con los autovalores de la matriz e incluso difieren en órdenes de magnitud y/o signo con los autovalores exactos.*

Estas dificultades han motivado que desde finales de los años 80 se haya desarrollado una intensa investigación en los **algoritmos de alta precisión relativa**, que garantizan cotas de error de la forma

$$\frac{|\lambda - \widehat{\lambda}|}{|\lambda|} \leq \widetilde{\eta} \qquad \forall \lambda, \tag{1.3}$$

con $\widetilde{\eta} = O(\epsilon)$, esto es, *que calculan con la misma precisión los autovalores grandes que los pequeños*. En la actualidad estos algoritmos existen sólo para ciertas clases muy particulares de matrices simétricas, así como para el problema relacionado de la descomposición en valores singulares (DVS) de matrices cualesquiera. A fecha de hoy sólo se conoce un algoritmo de alta precisión relativa para autovalores de matrices no simétricas: el algoritmo para matrices totalmente no negativas desarrollado por Koev [57] (véase también Koev & Dopico [58]).

Una historia resumida de los algoritmos de alta precisión relativa comienza con el cálculo de autovalores y autovectores de ciertas matrices tridiagonales simétricas [52], y el problema relacionado del cálculo de valores y vectores singulares de matrices bidiagonales [23, 35]. Estos estudios proporcionan algoritmos no sólo precisos sino muy eficientes, que en la actualidad son los utilizados por defecto para el problema bidiagonal en la librería numérica LAPACK [1]. Los algoritmos propuestos por Dhillon y Parlett [26, 27, 28], por ejemplo, son la base del algoritmo MRRR, el más eficiente a fecha de hoy[2] para el problema espectral tridiagonal simétrico. Dicho algoritmo está implementado en LAPACK y alcanza una precisión absoluta similar a la del algoritmo QR.

Para matrices densas los logros iniciales más importantes fueron: algoritmos para matrices escaladas diagonalmente dominantes (matrices que reescaladas tienen diagonal principal dominante) [2], para matrices acíclicas (matrices tales que el grafo bipartito que representa el patrón de ceros de la matriz es acíclico) [25], para matrices definidas positivas bien condicionadas tras escalamiento diagonal simétrico [24] y para ciertos tipos de matrices simétricas indefinidas muy difíciles de caracterizar a través de su factor polar definido positivo [95, 94].

En cada uno de estos trabajos pioneros se desarrollaba un algoritmo distinto, que requería un análisis de errores diferente y una teoría de perturbaciones propia. El trabajo unificador de Demmel et al. [22] establece que, para todas las clases de matrices para las que hoy en día existen algoritmos de alta precisión relativa para calcular la DVS, dicha descomposición se puede calcular en dos fases:

1. Calcular una descomposición de la matriz $A = XDY^T$ con D diagonal y X, Y bien condicionadas. Dicha descomposición se calcula usando diferentes versiones, *adaptadas a la estructura de la matriz* y, en ocasiones muy sofisticadas, del algoritmo de eliminación Gaussiana con pivote completo.

2. Aplicar un algoritmo de tipo Jacobi a la matriz factorizada XDY^T.

[2] El algoritmo MRRR es el único método para el problema espectral tridiagonal simétrico con coste $O(n^2)$, donde n es la dimensión de la matriz

Dado que el algoritmo de la segunda etapa es siempre el mismo, las diferencias para cada clase estructurada de matrices se reducen a *cómo calcular la factorización* XDY^T. La clave es evitar el efecto de *cancelación* en todas y cada una de las operaciones aritméticas del proceso de factorización. Para ello se modifica adecuadamente el algoritmo de eliminación gaussiana, haciendo uso de las propiedades especiales de la clase de matrices sobre la que se está trabajando para evitar las operaciones potencialmente peligrosas a efectos de cancelación. Hay que resaltar que en [22] no sólo se unifican la teoría y los algoritmos previamente existentes, sino que se introducen nuevas clases de matrices para las que la DVS se puede calcular con alta precisión relativa. Así se identificó la colección de matrices más amplia, hasta esa fecha, para las que esto era posible.

Las ideas de [22] fueron trasladadas al problema simétrico de autovalores por Dopico, Molera y Moro [30], en el sentido de demostrar que para todas las matrices simétricas, definidas o indefinidas, incluidas en las clases identificadas en [22], se pueden calcular sus autovalores y autovectores con alta precisión relativa. La idea es añadir una tercera fase al esquema de [22] que permite calcular autovalores y autovectores a partir de valores y vectores singulares. Sin embargo, el algoritmo de [30] no aprovecha al máximo la simetría de la matriz. En particular, comienza hallando una factorización $A = XDY^T$ mediante eliminación Gaussiana con pivote completo, que para matrices simétricas indefinidas *no respeta la simetría de la matriz*, y no permite en consecuencia los ahorros tanto de memoria como de operaciones esperables en un problema simétrico.

En el capítulo 3 de la presente memoria demostramos que, para dos clases específicas de matrices simétricas (las matrices DSTU y las TSC), la factorización no simétrica XDY^T de la primera etapa del algoritmo en [30] se puede reemplazar por una factorización simétrica XDX^T sin perder por ello precisión, ni en la factorización inicial ni en el cálculo subsiguiente de autovalores y autovectores. De hecho, se llega a la factorización XDX^T pasando por una factorización LDL^T por bloques que, convenientemente adaptada en cada caso a las propiedades específicas de la estructura, demostraremos que se calcula con error relativo pequeño *componente a componente*. Aunque este tipo tan fuerte de precisión se pierde al emplear después rotaciones de Givens para llegar a XDX^T, la precisión que se mantiene (columna a columna en norma) es suficiente para garantizar la alta precisión relativa de la segunda etapa del algoritmo espectral. Tras discutir en la sección 3.3 cómo se puede adaptar el algoritmo de factorización a cada una de las clases de matrices (DSTU y TSC), la sección 3.4 contiene un análisis de errores detallado al respecto. Finalmente, en la sección 3.5 se incluyen experimentos numéricos que confirman la alta precisión de los algoritmos propuestos. Otras clases de matrices para las que se puede obtener tanto factorizaciones simétricas como autovalores y autovectores con alta precisión relativa son las matrices de Cauchy (escaladas diagonalmente o no), las de Vandermonde y las matrices totalmente no negativas parametrizadas mediante su factorización bidiagonal [29].

Antes de terminar esta introducción queremos hacer notar que los dos tipos de incentivo que hemos mencionado a la hora de emplear algoritmos estructurados (*rapidez/almacenamiento* por un lado, y *precisión/estabilidad* por otro) no están descorrelacionados, ni son el objetivo primordial de comunidades científicas disjuntas: por un lado, se está haciendo un esfuerzo por estabilizar algoritmos rápidos ya existentes, por ejemplo mediante técnicas

especiales de pivotaje (véanse los artículos de *survey* [74, 10]). Por otro lado, los algoritmos de alta precisión relativa que acabamos de describir, cuya segunda etapa es un algoritmo de tipo Jacobi, pueden verse acelerados en un futuro muy próximo como consecuencia de las nuevas implementaciones del algoritmo de Jacobi, debidas principalmente a Drmač [31, 32], que comienzan a ser competitivas frente a algoritmos tradicionalmente considerados más rápidos, en particular frente al algoritmo QR.

Concluimos esta introducción mencionando los artículos a que ha dado lugar la investigación conducente a esta memoria: el contenido del capítulo 2 corresponde a los artículos [59, 77, 78]. El artículo [59], escrito en colaboración con Daniel Kressner, se encuentra en proceso de revisión por la revista *SIAM Journal on Matrix Analysis and Applications*, y el artículo [78] se someterá en breve a la revista *Linear Algebra and its Applications*. La nota [77], cuyos resultados están contenidos en [59], se publicó en la colección de actas *PAMM (Proceedings in Applied Mathematics and Mechanics)* de la *Sociedad Alemana de Matemática Aplicada y Mecánica (GAMM)*.

El contenido del capítulo 3 corresponde al artículo [76], que ya ha sido publicado en la revista *SIAM Journal on Matrix Analysis and Applications*.

1.3. Notación

Para comodidad del lector ofrecemos la siguiente tabla de notaciones, que se seguirá a lo largo de toda la memoria.

Símbolo	Significado
$\mathbb{C}^{n\times n}$	conjunto de matrices cuadradas de orden n con coeficientes complejos
$\mathbb{R}^{n\times n}$	conjunto de matrices cuadradas de orden n con coeficientes reales
$\mathbb{C}^p$	conjunto de vectores columna con p componentes complejas
$\mathbb{R}^p$	conjunto de vectores columna con p componentes reales
$\oplus_p$	suma directa de p matrices
$\otimes$	producto de Kronecker
$0(\cdot)$	O-grande de Landau
$o(\cdot)$	o-pequeña de landau
$>>$	mucho mayor
$<<$	mucho menor
$\equiv$	equivalente
$\not\equiv$	no equivalente
A^T	traspuesta de la matriz A, es decir, $(A^T)_{ij} = A_{ji}$
A^H	traspuesta conjugada de la matriz A, es decir, $(A^H)_{ij} = \overline{A_{ji}}$
$\sigma_i(A)$	i-ésimo mayor valor singular de la matriz A
$\|\cdot\|_F$	norma de Frobenius: $\|\|A\|\|_F = \sqrt{\sum_{i=1}^{n} \|a_{ij}^2\|}$
$\|\cdot\|_2$	norma espectral: $\|A\|_2 = \sigma_1(A)$
$\sigma(A)$	espectro de la matriz A
$\rho(A)$	radio espectral de la matriz A
I_n	matriz identidad de orden n
F_n	matriz antidiagonal con unos sobre la antidiagonal principal (*flip matrix*)
Σ_n	matriz antidiagonal con signos alternos sobre la antidiagonal principal

Por último, vaya de antemano la disculpa por el empleo de algunos términos y siglas, como *underflow* o *TSC*, que conservaremos en inglés, debido a que traducirlos conlleva o bien una pérdida grande de precisión o el empleo de una expresión poco natural en español.

Capítulo 2

Números de condición espectrales estructurados

2.1. Introducción

Como se señaló en el capítulo introductorio, los *números de condición* miden en general la sensibilidad de la solución de un problema frente a cambios infinitesimales en los datos de entrada. Un análisis regresivo de errores lleva en general a cotas de error del tipo

$$\text{(error en la solución)} \leq \text{(número de condición)} \times \text{(error regresivo)},$$

donde el error regresivo se define como en (1.1). Esta es la motivación de tratar de estimar tanto el número de condición del problema como el error regresivo correspondiente a cada algoritmo regresivamente estable. Además, como se ha visto también en el Capítulo 1, cuando la matriz del problema tiene una estructura particular puede ser ventajoso el incorporar esa estructura a los algoritmos de resolución. Un análisis de errores que incorpore la estructura requiere en primer lugar una estimación del *número de condición estructurado*, esto es, una estimación de lo sensibles que son las soluciones del problema con respecto a perturbaciones estructuradas. A ello se dedica el presente capítulo para el caso de algoritmos espectrales.

Si nos centramos en el problema del cálculo de autovalores *simples* de una matriz, hay numerosos resultados en la literatura relacionados con números de condición y errores regresivos, tanto estructurados como no estructurados [16, 39, 42, 55, 73, 77]. A continuación resumiremos algunos de estos resultados. Sea λ un *autovalor simple* de una matriz $A \in \mathbb{C}^{n\times n}$. Se sabe que λ es una función diferenciable de los elementos de la matriz A y que el autovalor perturbado $\widehat{\lambda}$ de la matriz perturbada $A + \epsilon E$ admite un desarrollo de la forma

$$\widehat{\lambda} = \lambda + \frac{y^H E x}{\|x\|_2 \, \|y\|_2}\epsilon + O(\epsilon^2), \quad \epsilon \to 0, \tag{2.1}$$

donde x e y son, respectivamente, autovectores derecho e izquierdo de la matriz A correspondientes al autovalor λ, normalizados por la condición $|y^H x| = 1$. Así, el número de

condición absoluto del autovalor λ, definido como

$$\kappa(A,\lambda) = \lim_{\epsilon\to 0} \sup_{\substack{\|E\|\leq 1 \\ E\in\mathbb{C}^{n\times n}}} \frac{|\widehat{\lambda}-\lambda|}{\epsilon}, \tag{2.2}$$

viene dado por $\kappa(A,\lambda) = \|x\|_2 \|y\|_2$ para cualquier norma unitariamente invariante $\|\cdot\|$. Una de las matrices de perturbación que alcanza la máxima variación es la matriz de rango uno

$$E = \frac{yx^H}{\|x\|_2 \|y\|_2}, \tag{2.3}$$

conocida como *perturbación de Wilkinson*. A fin de incorporar la estructura al concepto de número de condición debemos considerar el caso en que las matrices de perturbación E pertenecen a un conjunto $\mathbb{S} \subset \mathbb{C}^{n\times n}$ de matrices estructuradas. El número de condición estructurado de un autovalor simple se define como

$$\kappa(A,\lambda;\mathbb{S}) = \lim_{\epsilon\to 0} \sup_{\substack{\|E\|\leq 1 \\ E\in\mathbb{S}}} \frac{|\widehat{\lambda}-\lambda|}{\epsilon}, \tag{2.4}$$

Obviamente,

$$\kappa(A,\lambda;\mathbb{S}) \leq \kappa(A,\lambda),$$

puesto que el conjunto de perturbaciones admisibles en $\kappa(A,\lambda;\mathbb{S})$ es un subconjunto del conjunto admisible para $\kappa(A,\lambda)$. La pregunta que surge de modo natural es en qué situaciones será $\kappa(A,\lambda;\mathbb{S})$ mucho más pequeño que $\kappa(A,\lambda)$ o, lo que es lo mismo, cuándo será el autovalor λ mucho menos sensible a perturbaciones estructuradas que a perturbaciones arbitrarias. Para muchas estructuras $\mathbb{S}$ la respuesta es negativa, en el sentido de que el número de condición estructurado $\kappa(A,\lambda;\mathbb{S})$ es, salvo un factor de tamaño moderado, igual al número de condición no estructurado $\kappa(A,\lambda)$. Esto se demuestra en [16] para $\mathbb{S} = \mathbb{R}^{n\times n}$ y en [55, 82] para matrices antisimétricas reales, matrices de Hankel y Toeplitz, hamiltonianas, persimétricas, circulantes, ortogonales y unitarias. También hay ejemplos relevantes para los cuales $\kappa(A,\lambda;\mathbb{S}) \ll \kappa(A,\lambda)$; estos ejemplos corresponden a matrices antisimétricas complejas [82], ciertas cero–estructuras (esto es, ciertas matrices con ceros en posiciones prefijadas) [73] o matrices simplécticas [55].

El condicionamiento de autovalores múltiples, sin embargo, no ha sido tratado hasta el momento en la literatura. Aunque hay varios conceptos similares de número de condición no estructurado para autovalores múltiples [17, 72], esas definiciones no se han trasladado al contexto estructurado hasta muy recientemente [59]. La definición en [59], que veremos más adelante como Definición 2.2.1, se basa en la teoría de perturbación de Lidskii [64] para autovalores múltiples, eventualmente defectivos

Antes de recordar la teoría de Lidskii, nótese que la definición (2.2) no es válida para autovalores múltiples. Por ejemplo, la matriz

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

tiene un único autovalor nulo doble. Si se perturba de la forma

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \epsilon \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

entonces

$$\frac{|\widehat{\lambda} - \lambda|}{\epsilon} = \sqrt{\epsilon}/\epsilon = \sqrt{\epsilon}^{-1},$$

luego

$$\lim_{\epsilon \to 0} \sup_{\substack{\|E\| \leq 1 \\ E \in \mathbb{C}^{n \times n}}} \frac{|\widehat{\lambda} - \lambda|}{\epsilon} = \infty.$$

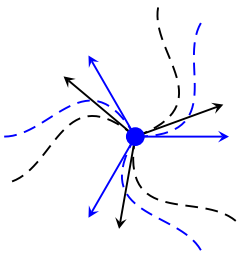
En general, es bien sabido que si λ es un autovalor múltiple de multiplicidad algebraica m, entonces $\widehat{\lambda}$ no se desarrolla de la forma (2.1), sino que pueden aparecer desarrollos de Puiseux en potencias fraccionarias (véase [3, §9.3.1] o [56, §II.1.2]). Más en concreto, el autovalor λ se bifurca en m autovalores perturbados $\widehat{\lambda}_k$ admitiendo cada uno un desarrollo

$$\widehat{\lambda}_k = \lambda + \alpha_k^{\gamma_k} \epsilon^{\gamma_k} + o(\epsilon), \qquad k = 1, \ldots, m, \tag{2.5}$$

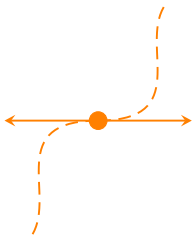
en potencias fraccionarias de ϵ con $\alpha_k > 0$ y $0 < \gamma_k \leq 1$. Bajo ciertas condiciones de genericidad sobre la matriz de perturbación E, la teoría de perturbación de Lidskii [64] asegura que cada bloque de Jordan de tamaño $n_j \times n_j$ asociado al autovalor λ da lugar genéricamente a n_j autovalores perturbados con desarrollo fraccional (2.5) para $\gamma_k = 1/n_j$. Por ejemplo, sea una matriz A con un autovalor λ de multiplicidad algebraica 9 y sea

$$\left(\begin{array}{ccc|ccc|cc} \lambda & 1 & & & & & & \\ & \lambda & 1 & & & & & \\ & & \lambda & & & & & \\ \hline & & & \lambda & 1 & & & \\ & & & & \lambda & 1 & & \\ & & & & & \lambda & & \\ \hline & & & & & & \lambda & 1 \\ & & & & & & & \lambda \\ \hline & & & & & & & \lambda \end{array}\right)$$

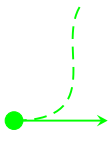
su correspondiente forma de Jordan. Si perturbamos esta matriz, en general los autovalores perturbados se bifurcan del modo que se muestra en las siguientes figuras:



Dos bloques 3×3.



Un bloque 2×2.



Un bloque 1×1.

donde el círculo representa el autovalor múltiple y las flechas las direcciones tangentes al movimiento de los autovalores perturbados.

Motivados por los desarrollos (2.5), en [72, §4] se define el *número de condición de Hölder* para un autovalor λ como un par

$$\kappa(A, \lambda) = (n_1, \alpha), \tag{2.6}$$

donde n_1 es el tamaño del mayor bloque de Jordan de A asociado a λ y la cantidad $\alpha^{1/n_1} > 0$ es el valor más grande posible del coeficiente ϵ^{1/n_1} entre todas las perturbaciones $E \in \mathbb{C}^{n\times n}$ con $||E|| \leq 1$ (véase la Definición 2.1.3 más adelante). Nótese que, al ser $1/n_1$ el menor exponente posible en los desarrollos (2.5), se tiene que

$$\alpha^{1/n_1} = \lim_{\epsilon\to 0} \sup_{\substack{\|E\|\leq 1 \\ E\in\mathbb{C}^{n\times n}}} \max_{k=1,\ldots,m} \frac{|\widehat{\lambda}_k - \lambda|}{\epsilon^{1/n_1}}. \tag{2.7}$$

(véase [17, p. 156] para una definición similar de número de condición de autovalores múltiples).

Llegados a este punto debemos hacer notar que, para ciertas perturbaciones no genéricas, el valor de γ_k en (2.5) puede ser mayor que $1/n_1$ *para todo autovalor perturbado* $\widehat{\lambda}_k$. Por ejemplo, consideremos el siguiente ejemplo, sacado de [97, §2.22]. Sea la matriz perturbada

$$A + \epsilon E = \left(\begin{array}{ccc|cc} 0 & 1 & 0 & & \\ & 0 & 1 & & \\ & & 0 & \epsilon & \\ \hline & & & 0 & 1 \\ \epsilon & & & & 0 \end{array}\right), \tag{2.8}$$

cuyo polinomio característico es $\epsilon^2 - \lambda^5$. En este caso, $\gamma_k = 2/5$ para todo $\widehat{\lambda}_k$ en (2.5) mientras que $1/n_1 = 1/3$.

Estos casos no genéricos suelen aparecer cuando la perturbación E pertenece a un conjunto $\mathbb{S}$ de matrices con una estructura dada. Fijada una estructura $\mathbb{S}$, queremos definir un *número de condición de tipo Hölder estructurado* de la forma

$$\kappa(A, \lambda; \mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}}),$$

donde $1/n_{\mathbb{S}}$ es el menor exponente posible de ϵ en (2.5) y la cantidad $\alpha^{1/n_{\mathbb{S}}} > 0$ es el valor más grande posible en el coeficiente $\epsilon^{1/n_{\mathbb{S}}}$ para toda perturbación $E \in \mathbb{S}$ (véase la Definición 2.2.1 más adelante).

En general, suele ocurrir que $n_{\mathbb{S}} = n_1$, pero existen estructuras tales que $n_{\mathbb{S}}$ es menor que n_1. Por ejemplo si $\mathbb{S}$ es el conjunto de matrices antisimétricas complejas, la matriz

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ -1 & 0 & -i & 0 \\ 0 & i & 0 & i \\ -1 & 0 & -i & 0 \end{pmatrix} \in \mathbb{S} \tag{2.9}$$

tiene un único autovalor $\lambda = 0$ de multiplicidad geométrica dos y mayor bloque de Jordan de tamaño tres, esto es, $n_1 = 3$. Puede comprobarse fácilmente que para cualquier matriz de perturbación $E \in \mathbb{S}$ la matriz perturbada $A + \epsilon E$ tiene autovalores de orden $O(\epsilon^{1/2})$. En otras palabras, se tiene $n_{\mathbb{S}} = 2 < n_1 = 3$.

En casos como éste en los que, dada una matriz y uno de sus autovalores, todas las perturbaciones dentro de una clase estructurada dén lugar a perturbaciones (2.5) con $\gamma_k > 1/n_1$ (y, por tanto, sea $n_{\mathbb{S}} < n_1$), diremos que las perturbaciones en $\mathbb{S}$ son *completamente no genéricas* (en inglés, *fully nongeneric*). Más adelante dividiremos nuestro estudio en dos casos: llamaremos *genérico* al caso en que $n_{\mathbb{S}} = n_1$ y caso *completamente no genérico* a aquél en que $n_{\mathbb{S}} < n_1$.

Finalmente, hacemos notar que, como se verá en §2.3.1 y §2.3.2, se pueden definir también números de condición de tipo Hölder para problemas generalizados de autovalores, tales como *haces de matrices* $A - \lambda B$ con $A, B \in \mathbb{C}^{n\times n}$, o *polinomios matriciales* de la forma $P(\lambda) = \lambda^m A_m + \lambda^{m-1} A_{m-1} + \ldots + \lambda A_1 + A_0 \quad \text{con} \quad A_i \in \mathbb{C}^{n\times n}$.

2.1.1. Preliminares

En esta sección resumiremos los resultados de la teoría de Lidskii [64, 72] que conducen al número de condición (2.6). Además, haremos unos breves comentarios sobre *formas canónicas estructuradas* [89, 71], que usaremos con frecuencia en todo este capítulo.

Sea λ un autovalor de la matriz $A \in \mathbb{C}^{n\times n}$ y sea n_1 el tamaño del mayor bloque de Jordan correspondiente a dicho autovalor λ. La forma canónica de Jordan de la matriz A se puede escribir como

$$\left(\begin{array}{c|c} J & 0 \\ \hline 0 & \widetilde{J} \end{array}\right) = \left(\begin{array}{c} Q \\ \hline \widetilde{Q} \end{array}\right) A \left(\begin{array}{c|c} P & \widetilde{P} \end{array}\right) = P_A^{-1} A P_A, \tag{2.10}$$

donde

$$\left(\begin{array}{c} Q \\ \hline \widetilde{Q} \end{array}\right) \left(\begin{array}{c|c} P & \widetilde{P} \end{array}\right) = I \tag{2.11}$$

y J contiene todos los bloques de Jordan de tamaño $n_1 \times n_1$ asociados a λ. Más concretamente, se tiene que

$$J = \mathrm{diag}(\Gamma_1^1, \ldots, \Gamma_1^{r_1}), \qquad \Gamma_1^1 = \cdots = \Gamma_1^{r_1} = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix} \in \mathbb{C}^{n_1\times n_1}. \tag{2.12}$$

El bloque $\widetilde{J}$ contiene tanto los bloques de Jordan de tamaño menor que n_1 asociados a λ como los bloques de Jordan asociados a autovalores distintos de λ.

Las columnas de la matriz P forman r_1 cadenas de Jordan linealmente independientes de longitud n_1, cada una de las cuales comienza con un autovector derecho de A asociado

a λ. Reunimos todos estos autovectores derechos en una matriz X de tamaño $n \times r_1$

$$X = \left(\; Pe_1, \; Pe_{n_1+1}, \; \ldots, \; Pe_{(r_1-1)n_1+1} \; \right), \tag{2.13}$$

donde e_i denota la columna i-ésima de la matriz identidad de orden n. Análogamente, reunimos en una matriz Y los autovectores izquierdos de las r_1 cadenas de Jordan linealmente independientes que aparecen en la matriz Q, es decir,

$$Y = \left(\; Q^H e_{n_1}, \; Q^H e_{2n_1}, \; \ldots, \; Q^H e_{r_1 n_1} \; \right). \tag{2.14}$$

Nótese que la relación (2.11) implica que $Y^H X = I_{r_1 \times r_1}$ si $n_1 = 1$ e $Y^H X = 0_{r_1 \times r_1}$ en otro caso. Con esta notación, el siguiente resultado es una versión simplificada del resultado fundamental de [64]:

Teorema 2.1.1 ([64, 72]) *Sea $A \in \mathbb{C}^{n\times n}$, sea λ un autovalor de A y sea (2.10) una forma de Jordan de A. Sean las matrices X e Y definidas como en (2.13) y (2.14), respectivamente, y supongamos que $Y^H EX$ es invertible para una cierta $E \in \mathbb{C}^{n\times n}$. Entonces existen $n_1 r_1$ autovalores $\widehat{\lambda}_k$ de la matriz perturbada $A+\epsilon E$ que admiten un desarrollo en potencias fraccionarias de la forma*

$$\widehat{\lambda}_k = \lambda + (\xi_k)^{1/n_1} \epsilon^{1/n_1} + o(\epsilon^{1/n_1}), \quad k = 1, \ldots, r_1, \tag{2.15}$$

donde $\xi_1, \ldots, \xi_{r_1}$ son los autovalores de la matriz $Y^H EX$.

Nótese que las columnas de las matrices X e Y son linealmente independientes, luego la matriz $Y^H EX$ será no singular para una perturbación genérica $E \in \mathbb{C}^{n\times n}$. En el caso no genérico en que $Y^H EX$ es singular, los desarrollos (2.15) siguen siendo válidos, aunque sólo parcialmente. En concreto, se tiene el siguiente resultado.

Lema 2.1.2 ([72, Pg. 809]) *Sea λ un autovalor de la matriz $A \in \mathbb{C}^{n\times n}$, y sea (2.10) una forma canónica de Jordan de A. Sean las matrices X e Y definidas como en (2.13) y (2.14), respectivamente. Sea $E \in \mathbb{C}^{n\times n}$ una perturbación tal que $Y^H EX$ es singular. Entonces cada uno de los $\beta < r_1$ autovalores no nulos $\xi_1, \ldots, \xi_\beta$ de $Y^H EX$ da lugar a n_1 autovalores perturbados de la forma (2.15). Los $r_1 - \beta$ autovalores nulos restantes corresponden a autovalores perturbados con desarrollos en los que el exponente principal es estrictamente menor que $1/n_1$.*

El Teorema 2.1.1 implica que, para valores de ϵ suficientemente pequeños, el peor caso en cuanto a la magnitud de la perturbación en los autovalores corresponde al mayor autovalor en valor absoluto de la matriz $Y^H EX$. Esto motiva la siguiente definición de número de condición.

Definición 2.1.3 ([72, Definition 4.1]) *Sea λ un autovalor de la matriz $A \in \mathbb{C}^{n\times n}$, sea (2.10) una forma canónica de Jordan de A y sean X e Y matrices definidas como en (2.13) y (2.14). El* **número de condición (absoluto) Hölder** *de λ se define como*

$\kappa(A,\lambda) = (n_1, \alpha)$, *donde* n_1 *es el tamaño del mayor bloque de Jordan asociado al autovalor* λ *y*

$$\alpha = \sup_{\substack{\|E\|\leq 1 \\ E\in\mathbb{C}^{n\times n}}} \rho(Y^H E X), \tag{2.16}$$

donde $\rho(\cdot)$ *denota el radio espectral de una matriz.*

La cadena de desigualdades $\rho(Y^H E X) = \rho(EXY^H) \leq \|EXY^H\|_2 \leq \|XY^H\|_2$ para la norma $\|\cdot\|_2$ nos dice que para alcanzar la igualdad deberíamos construir una matriz de perturbación E tal que $\rho(Y^H E X) = \|XY^H\|_2$. El siguiente lema caracteriza algunas de estas perturbaciones.

Lema 2.1.4 *Sea*

$$XY^H = U\Sigma V^H$$

una descomposición en valores singulares, esto es, $U \in \mathbb{C}^{n\times r_1}$, $V \in \mathbb{C}^{n\times r_1}$ *con columnas ortonormales y* $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{r_1})$ *con* $\sigma_1 \geq \cdots \sigma_{r_1} \geq 0$. *Consideramos* $E = VDU^H$ *con* $D = \mathrm{diag}(1, \delta_2, \ldots, \delta_{r_1})$ *tal que* $\delta_j \leq 1,\ j = 2, \ldots, r_1$. *Entonces* $\rho(EXY^H) = \|XY^H\|_2$.

Demostración:
El resultado se sigue de que

$$\rho(EXY^H) = \rho(VD\Sigma V^H) = \rho(D\Sigma) = \|D\Sigma\|_2 = \|\Sigma\|_2 = \|XY^H\|_2.$$

□

Nótese que la definición de α en (2.16) depende de la norma $\|\cdot\|$ usada. Para cualquier norma unitariamente invariante, si $D = \mathrm{diag}(1, 0, \ldots, 0)$ y $E = VDU^H$ en el Lema 2.1.4, se tiene $\|E\| = 1$ y por tanto

$$\alpha = \|XY^H\|_2\,,$$

resultado que ya se demostró en [72]. Para otras normas, otras elecciones de D son posibles. Para la norma espectral, por ejemplo, cualquier perturbación E definida como en el Lema 2.1.4, cumple que $\|E\|_2 = 1$. En particular, si tomamos $D = \Sigma/\sigma_1$, la matriz

$$E = \frac{YX^H}{\|XY^H\|_2}, \tag{2.17}$$

que llamaremos *perturbación de Lidskii*, resulta ser una generalización de la perturbación de Wilkinson (2.3), en el sentido de que es una matriz de rango bajo que realiza el máximo efecto sobre λ de entre todas las perturbaciones de norma 1. Este tipo de perturbaciones será útil al probar que el número de condición estructurado y el no estructurado coinciden para la norma dos. Otras clases de perturbaciones que también se usarán son las siguientes.

Lema 2.1.5 *Sea* λ *autovalor de una matriz* $A \in \mathbb{C}^{n\times n}$ *con forma de Jordan* (2.10) *y sean* X *e* Y *las matrices de autovectores de* A *definidas en* (2.13) *y* (2.14)*, respectivamente. Sean* $u_1, v_1 \in \mathbb{C}^{n\times n}$*, respectivamente, vectores singulares izquierdo y derecho correspondientes al mayor valor singular de la matriz* XY^H. *Sea* $E \in \mathbb{C}^{n\times n}$ *tal que* $Eu_1 = \beta v_1$ *con* $|\beta| = 1$. *Entonces* $\rho(EXY^H) \geq \|XY^H\|_2$.

Demostración:
Sea $XY^H = U\Sigma V^H$ una descomposición en valores singulares con $U = [u_1, \ldots, u_n]$, $V = [v_1, \ldots, v_n]$. Entonces

$$\begin{aligned}\rho(EXY^H) &= \rho(EU\Sigma V^H) = \rho(V^H EU\Sigma) = \rho(V^H[\beta v_1, Ev_2, \ldots, Ev_n]\Sigma) \\ &= \rho\left(\begin{pmatrix} \beta\|XY^H\|_2 & \star \\ 0 & \star \end{pmatrix}\right) \geq \|XY^H\|_2.\end{aligned}$$

□

También necesitamos para nuestro análisis emplear formas canónicas específicas para ciertas estructuras, ya sea de matrices o de haces de matrices. En general, las formas canónicas que emplearemos son la de Jordan para matrices (a la que se llega por semejanza) y la de Weierstrass para haces regulares de matrices (a la que se llega por equivalencia), aunque a veces recurriremos también a formas canónicas *por congruencia* para haces estructurados de matrices.

Como ya hemos visto, las matrices X e Y definidas en (2.13) y (2.14) juegan un papel crucial a la hora de definir los números de condición de Hölder. Una de nuestras herramientas fundamentales en las secciones 2.2 y 2.3 será el hecho de que algunas estructuras inducen relaciones especiales entre X e Y. Estas relaciones son consecuencia inmediata de las *formas canónicas estructuradas*, descritas por ejemplo en [89] y [71], para haces de matrices $A - \lambda B$, cuando A y/o B son matrices simétricas o antisimétricas [89] o, de forma más general, autoadjuntas o antiautoadjuntas con respecto a productos escalares arbitrarios [71]. Por ejemplo, la forma de Jordan de una matriz simétrica (resp. antisimétrica) se obtiene fácilmente de la forma canónica por congruencia del haz simétrico (resp. del haz antisimétrico/simétrico) $A - \lambda I$: si el haz $\mathcal{G} - \lambda\mathcal{H}$ con

$$\mathcal{G} = P^T AP, \quad \mathcal{H} = P^T P$$

es una forma canónica por congruencia de $A - \lambda I$ entonces puede comprobarse que

$$\mathcal{H}^{-1}\mathcal{G} = P^{-1}AP$$

es una semejanza que lleva A a una forma muy cercana a su forma de Jordan. Para llegar a la forma de Jordan basta llevar a cabo ciertas reordenaciones de filas y columnas, que suelen involucrar a las matrices

$$\Sigma_k = \begin{pmatrix} & & (-1)^0 \\ & \iddots & \\ (-1)^{k-1} & & \end{pmatrix}, \qquad F_k = \begin{pmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{pmatrix}, \tag{2.18}$$

ambas de $k \times k$. Las matrices F_k y Σ_k aparecerán con frecuencia a lo largo de todo el capítulo.

2.2. Perturbaciones estructuradas genéricas: el caso matricial

2.2.1. Número de condición de Hölder estructurado para autovalores múltiples

En esta sección λ denota un autovalor de la matriz $A \in \mathbb{C}^{n\times n}$ con número de condición de Hölder *no estructurado* $\kappa(A,\lambda) = (n_1, \alpha)$ dado por la Definición 2.1.3. Las matrices X e Y están definidas como en (2.13) y (2.14) respectivamente.

Si restringimos las perturbaciones admisibles E de $\mathbb{C}^{n\times n}$ a un subconjunto $\mathbb{S} \subset \mathbb{C}^{n\times n}$, tendremos el correspondiente número de condición estructurado $\kappa(A,\lambda;\mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$.

Definición 2.2.1 *Sea λ un autovalor de la matriz $A \in \mathbb{C}^{n\times n}$ y sea $\mathbb{S}$ un subconjunto de matrices de $\mathbb{C}^{n\times n}$. El* **número de condición estructurado de Hölder** *de λ con respecto a $\mathbb{S}$ se define como*

$$\kappa(A,\lambda;\mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}}),$$

donde $1/n_{\mathbb{S}}$ es el menor exponente posible γ_k de ϵ en el desarrollo de los autovalores perturbados (2.5) entre todas las perturbaciones $E \in \mathbb{S}$, y $\alpha_{\mathbb{S}} > 0$ es el mayor valor posible para α_k en (2.5) para toda $E \in \mathbb{S}$ con $\|E\| \leq 1$.

Como se dijo en la sección anterior existen casos en los que $n_{\mathbb{S}} < n_1$, pero en esta sección nos restringiremos al caso más frecuente, en el que $n_{\mathbb{S}} = n_1$. En tal caso, por el Teorema 2.1.1 y el Lema 2.1.2, se tiene que

$$\alpha_{\mathbb{S}} = \sup_{\substack{\|E\|\leq 1 \\ E\in\mathbb{S}}} \rho(Y^H E X). \tag{2.19}$$

La presencia del radio espectral en (2.19) complica en gran medida el problema de determinar el supremo y dificulta el conseguir fórmulas explícitas o cotas razonables para $\alpha_{\mathbb{S}}$. Veremos, sin embargo, que para ciertas estructuras es posible identificar cuándo $\alpha_{\mathbb{S}}$ y α son valores cercanos construyendo perturbaciones $E \in \mathbb{S}$ para las cuales $\rho(Y^H E X)$ es cercano a α. En las próximas subsecciones determinaremos tales estructuras.

2.2.2. Matrices reales

Como primer ejemplo, nos restringimos a perturbaciones reales. Veremos que, en el mejor de los casos, sólo se mejora ligeramente la sensibilidad del autovalor λ. Esto ya se demostró para autovalores simples en [16]. El siguiente lema es la generalización para autovalores múltiples.

Lema 2.2.2 *Sea una matriz $A \in \mathbb{C}^{n\times n}$ y sea λ un autovalor de A con número de condición de Hölder $\kappa(A,\lambda) = (n_1, \alpha)$. Entonces $\kappa(A,\lambda;\mathbb{R}^{n\times n}) = (n_1, \alpha_{\mathbb{R}})$ con*

(1). $\alpha/2 \leq \alpha_{\mathbb{R}} \leq \alpha$ para cualquier norma unitariamente invariante $\|\cdot\|$ y

(II). $\alpha_{\mathbb{R}} = \alpha$ para la norma espectral si la matriz A es una matriz normal.

Demostración:

Descomponemos la matriz compleja $XY^H = M_R + \imath M_I$ con $M_R, M_I \in \mathbb{R}^{n\times n}$ tales que $\|M_R\|_2 \geq \|XY^H\|_2/2$ ó $\|M_I\|_2 \geq \|XY^H\|_2/2$. Sin pérdida de generalidad suponemos que $\|M_R\|_2 \geq \|XY^H\|_2/2$. Sea la perturbación $E = v_1 u_1^T \in \mathbb{R}^{n\times n}$, donde u_1 y v_1 son vectores singulares izquierdo y derecho, respectivamente, correspondientes al mayor valor singular de la matriz M_R. Entonces $\|E\| = 1$ y

$$\begin{aligned} \rho(EXY^H) &= \rho(v_1 u_1^T (M_R + \imath M_I)) = |u_1^T (M_R + \imath M_I) v_1| \\ &\geq |u_1^T M_R v_1| = \|M_R\|_2 \geq \|XY^H\|_2/2, \end{aligned}$$

lo que prueba $\alpha_{\mathbb{R}} \geq \alpha/2$. Para demostrar la segunda parte utilizamos el hecho de que A es normal, y por tanto unitariamente diagonalizable. Por tanto, podemos elegir la matriz Y tal que $Y = X$. En tal caso, la perturbación $E = I \in \mathbb{R}^{n\times n}$ cumple que $\rho(EXX^H) = \alpha = \alpha_{\mathbb{R}}$. □

Observemos que en el caso $r_1 = 1$ (un único bloque de Jordan de tamaño n_1) se puede usar el mismo argumento que en [16] para obtener la cota inferior del Lema 2.2.2 (I). En cambio, no es claro que se puedan utilizar los mismos argumentos para el caso $r_1 > 1$.

Supongamos ahora que $\mathbb{S}$ es una estructura tal que si $E \in \mathbb{S}$, entonces tanto su parte real como su parte imaginaria están en $\mathbb{S} \cap \mathbb{R}^{n\times n}$. Por ejemplo, si $\mathbb{S}$ es el conjunto de las matrices simétricas complejas, al restringirnos a la parte real o a la imaginaria obtenemos las matrices *reales y simétricas*. Para un autovalor simple, Rump [82] extendió las cotas del Lema 2.2.2 a números de condición estructurados, en el sentido de restringir las perturbaciones de $\mathbb{S}$ a $\mathbb{S} \cap \mathbb{R}^{n\times n}$, mejorando así el número de condición a lo sumo un factor $1/\sqrt{2}$. Este resultado se puede extender facilmente al caso múltiple cuando $r_1 = 1$, pero no ocurre lo mismo para autovalores múltiples con varios bloques de Jordan del mayor tamaño. El siguiente lema es un primer paso para conseguir resultados en esta dirección.

Lema 2.2.3 *Sea $\mathbb{SR} \subset \mathbb{R}^{n\times n}$ y sea $\mathbb{S} = \mathbb{SR} + \imath\mathbb{SR}$ el subconjunto de todas las matrices con parte real e imaginaria en $\mathbb{SR}$. Si existe una matriz $E \in \mathbb{S}$ de rango uno con $\|E\| = 1$ tal que $\alpha_{\mathbb{S}} = \rho(Y^H E X)$ entonces*

$$\alpha_{\mathbb{S}}/4 \leq \alpha_{\mathbb{SR}} \leq \alpha_{\mathbb{S}}$$

en la norma Frobenius y la norma dos.

Demostración:

Consideremos la matriz $E = vu^H$ con $u, v \in \mathbb{C}^{n\times n}$ y $\|u\|_2 = \|v\|_2 = 1$. Si descomponemos los vectores como $u = u_R + \imath u_I$ y $v = v_R + \imath v_I$ con $u_R, u_I, v_R, v_I \in \mathbb{R}^n$, tenemos que

$$\alpha_{\mathbb{S}} = |u^H XY^H v| = |(u_R^T XY^H v_R + u_I^T XY^H v_I) - \imath(u_I^T XY^H v_R - u_R XY^H v_I^T)|.$$

Uno de los dos sumandos entre paréntesis del término de la derecha tiene que ser mayor o igual que $\alpha_{\mathbb{S}}/2$ en valor absoluto. Supongamos sin pérdida de generalidad que

$$|u_R^T XY^H v_R + u_I^T XY^H v_I| \geq \alpha_{\mathbb{S}}/2$$

y sea $U = [u_R, u_I]$, $V = [v_R, v_I]$. Entonces $|\mathrm{tr}(U^T X Y^H V)| \geq \alpha_{\mathbb{S}}/2$. Y, por tanto

$$\rho(U^T X Y^H V) \geq \alpha_{\mathbb{S}}/4.$$

Así, la perturbación real $E_R = VU^T$ (parte real de E) cumple $\alpha_{\mathbb{SR}} \geq \rho(EXY^H) \geq \alpha_{\mathbb{S}}/4$ con $\|E\|_2 \leq 1$ y $\|E\|_F \leq 1$, lo cual completa la demostración.

□

2.2.3. Estructuras lineales

Estudiemos ahora el caso en que $\mathbb{S}$ es un espacio lineal de matrices en el cuerpo $\mathbb{F}^{n\times n}$ con $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. Usando técnicas desarrolladas por Higham y Higham [41], dada una matriz $E \in \mathbb{S}$ existe un único vector $p = [p_1, \ldots, p_l]^T \in \mathbb{F}^l$ y una base $\{M_1, \ldots, M_l\}$ del subespacio $\mathbb{S}$ tal que $E = p_1 M_1 + \cdots + p_l M_l$ y $\|E\|_F = \|p\|_2$. Si $n_{\mathbb{S}} = n_1$, el número de condición estructurado $\kappa(A, \lambda; \mathbb{S}) = (n_1, \alpha_{\mathbb{S}})$ satisface

$$\alpha_{\mathbb{S}} = \sup_{\substack{\|p\|_2 \leq 1 \\ p \in \mathbb{F}^l}} \rho\left(p_1 Y^H M_1 X + \cdots + p_l Y^H M_l X\right) \tag{2.20}$$

para la norma Frobenius. Maximizar una función espectral no simétrica es un problema de optimización no trivial como se puede ver, por ejemplo, en [15]. Por tanto, conseguir en general una fórmula explícita de la expresión $\alpha_{\mathbb{S}}$ no parece fácil. Dos casos especiales para los que existe una expresión más manejable de $\alpha_{\mathbb{S}}$ son los siguientes.

1. El caso $r_1 = 1$ (es decir, X e Y son vectores) puede tratarse como en el caso de un autovalor simple [42, 91]. Definimos la *matriz de patrones*

$$\mathcal{M} = [\mathrm{vec}(M_1), \ldots, \mathrm{vec}(M_l)], \tag{2.21}$$

donde el comando *vec* convierte una matriz en un vector concatenando las columnas de dicha matriz. Entonces $vec(E) = \mathcal{M}p$ y por tanto

$$\alpha_{\mathbb{S}} = \sup_{\substack{\|p\|_2 \leq 1 \\ p \in \mathbb{F}^l}} |p_1 Y^H M_1 X + \cdots + p_l Y^H M_l X|$$

Por otra parte, si $\otimes$ denota el producto de Kronecker, entonces

$$|p_1 Y^H M_1 X + \cdots + p_l Y^H M_l X| = |vec(p_1 Y^H M_1 X + \cdots + p_l Y^H M_l X)| =$$

$$= |(X^T \otimes Y^H)\mathcal{M}p| \leq \|(X^T \otimes Y^H)\mathcal{M}\|_2 \|p\|_2,$$

alcánzándose la igualdad para un cierto vector p. De este modo

$$\alpha_{\mathbb{S}} = \|(X^T \otimes Y^H)\mathcal{M}\|_2$$

cuando $\mathbb{F} = \mathbb{C}$ o cuando $\mathbb{F} = \mathbb{R}$ y $X, Y \in \mathbb{R}^n$. Para $\mathbb{F} = \mathbb{R}$ y $X, Y \notin \mathbb{R}^n$ podemos demostrar como en [55, Sección 2] que $\|(X^T \otimes Y^H)\mathcal{M}\|_2/\sqrt{2} \leq \alpha_{\mathbb{S}} \leq \|(X^T \otimes Y^H)\mathcal{M}\|_2$.

2. Si $\mathbb{F} = \mathbb{C}$ y todas las matrices $N_j = Y^H M_j X$ son matrices hermíticas, entonces

$$\begin{aligned} \alpha_{\mathbb{S}} &= \sup_{\substack{\|p\|_2 \leq 1 \\ p \in \mathbb{C}^l}} \|p_1 N_1 + \cdots + p_l N_l\|_2 \\ &= \sup_{\substack{\|x\|_2 = 1 \\ x \in \mathbb{C}^n}} \|[x^H N_1 x, \ldots, x^H N_l x]\|_2. \end{aligned}$$

Se sigue por tanto

$$\max_i \|N_i\|_2 \leq \alpha_{\mathbb{S}} \leq \sqrt{l} \max_i \|N_i\|_2.$$

2.2.4. Matrices de Toeplitz y Hankel

Como dijimos en la introducción, el número de condición mide la sensibilidad de un autovalor frente a perturbaciones infinitesimales. El comportamiento frente a perturbaciones finitas de la matriz se caracteriza por medio del *pseudoespectro* (véase [92, 93]). El *pseudoespectro asociado a* $\epsilon > 0$ de una matriz se define como

$$\Lambda_\epsilon := \{\lambda \in \mathbb{C} : \exists\, E \in \mathbb{C}^{n\times n},\ ||E|| \leq \epsilon,\ \lambda \in \sigma(A+E)\}.$$

Si consideramos una estructura $\mathbb{S}$, podemos definir el *pseudoespectro estructurado* con respecto a $\mathbb{S}$ como

$$\Lambda_\epsilon^{\mathbb{S}} := \{\lambda \in \mathbb{C} : \exists\, E \in \mathbb{S},\ ||E|| \leq \epsilon,\ \lambda \in \sigma(A+E)\}.$$

Los números de condición estructurado y no estructurado pueden obtenerse a partir del pseudoespectro estructurado y no estructurado, respectivamente. En [82] se prueba que el pseudoespectro estructurado de una matriz $A \in \mathbb{S}$ coincide con el pseudoespectro no estructurado para las siguientes clases $\mathbb{S}$ de matrices *complejas* $\mathbb{S}$: simétricas, persimétricas, Toeplitz, Toeplitz simétricas, Hankel, Hankel persimétricas y circulantes; por tanto $\kappa(A, \lambda; \mathbb{S}) = \kappa(A, \lambda)$ para todas estas estructuras. Por tanto, alguno de los resultados que enunciamos a continuación son consecuencia de los resultados en [82]. No obstante, en las demostraciones construimos explícitamente perturbaciones que alcanzan el supremo $\kappa(A, \lambda)$, algo que no se obtiene del enfoque empleado en [82].

Una *matriz Toeplitz* tiene la forma

$$T = \begin{pmatrix} t_0 & t_{-1} & \ldots & t_{-n+1} \\ t_1 & t_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_{-1} \\ t_{n-1} & \ldots & t_1 & t_0 \end{pmatrix} \in \mathbb{C}^{n\times n}$$

y $H \in \mathbb{C}^{n\times n}$ es una *matriz Hankel* si $F_n H$ es una matriz Toeplitz, donde la matriz cuadrada F_n es la matriz con unos en la antidigonal y ceros en el resto de sus entradas definida en (2.18).

Teorema 2.2.4 *Sean $\mathbb{T}$ y $\mathbb{SY}$ los conjuntos de las matrices Toeplitz y de las matrices simétricas complejas, respectivamente. Entonces, para la norma espectral se tiene:*

*(*I*).* $\kappa(A,\lambda;\mathbb{T}) = \kappa(A,\lambda) = (n_1, \|XX^T\|_2)$ *para* $A \in \mathbb{T}$*;*

*(*II*).* $\kappa(A,\lambda;\mathbb{T}\cap\mathbb{SY}) = \kappa(A,\lambda) = (n_1, \|XX^T\|_2)$ *para* $A \in \mathbb{T}\cap\mathbb{SY}$*;*

*(*III*).* $\kappa(A,\lambda;\mathbb{T}\cap\mathbb{R}^{n\times n}) = \kappa(A,\lambda) = (n_1, \|XX^T\|_2)$ *para* $A \in \mathbb{T}\cap\mathbb{R}^{n\times n}$ *y* $\lambda \in \mathbb{R}$*;*

*(*IV*).* $\kappa(A,\lambda;\mathbb{T}\cap\mathbb{SY}\cap\mathbb{R}^{n\times n}) = \kappa(A,\lambda) = (1,1)$ *para* $A \in \mathbb{T}\cap\mathbb{SY}\cap\mathbb{R}^{n\times n}$*,*

donde X es la matriz de autovectores derechos asociados a λ definida en (2.13)*.*

Demostración:

Una matriz Toeplitz es una matriz *persimétrica*, es decir, F_nT es simétrica, con F_n dada por (2.18). Podemos, por tanto, aplicar resultados de Thompson [89] sobre formas canónicas de haces simétricos complejos al haz simétrico $F_nT - \lambda F_n$ (véase también [71, Theorem 7.2]). Según estos resultados, podemos elegir las matrices P y Q en la forma canónica de Jordan (2.10) de manera que

$$Q = \mathrm{diag}(F_{n_1}, \ldots, F_{n_1})P^TF_n$$

y así, de acuerdo con (2.14), se tiene $Y = F_n\overline{X}$. A continuación tomamaos una factorización de Takagi [48, §4.4] de la matriz simétrica compleja XX^T, esto es, una descomposición en valores singulares especial $XX^T = U\Sigma U^T$, donde $U \in \mathbb{C}^{n\times r_1}$ tiene columnas ortonormales y $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{r_1})$ con $\sigma_1 \geq \cdots \geq \sigma_{r_1} > 0$. Por [81, Lema 10.1], existe una matriz de Hankel H con $\|H\|_2 = 1$ y $Hu_1 = \overline{u}_1$, donde u_1 denota la primera columna de la matriz U. Basta escoger la matriz de perturbación $E = F_nH \in \mathbb{T}$, que cumple $\|E\|_2 = 1$ y $Eu_1 = F_n\overline{u}_1$, para demostrar (I) gracias al Lema 2.1.5.

Una matriz Toeplitz simétrica es a la vez persimétrica; esto permite, a través de una simple transformación ortogonal, diagonalizar por bloques la matriz A de la forma :

$$G^TAG = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix},$$

donde $A_{11} \in \mathbb{R}^{\lfloor n/2\rfloor\times\lfloor n/2\rfloor}$ y $A_{22} \in \mathbb{R}^{\lceil n/2\rceil\times\lceil n/2\rceil}$ son matrices simétricas complejas,

$$G = \frac{1}{\sqrt{2}}\begin{pmatrix} I & F_{n/2} \\ -F_{n/2} & I \end{pmatrix} \text{ si } n \text{ es par}$$

y

$$G = \frac{1}{\sqrt{2}}\begin{pmatrix} I & 0 & F_{(n-1)/2} \\ 0 & \sqrt{2} & 0 \\ -F_{(n-1)/2} & 0 & I \end{pmatrix} \text{ si } n \text{ es impar.}$$

Este resultado, que puede encontrarse en [98], muestra que $X = [X_1, X_2]$ con $X_1 = -FX_1$ y $X_2 = FX_2$. Los autovectores contenidos en las matrices X_1 y X_2 provienen de los bloques de Jordan de A_{11} y A_{22}, respectivamente. Además, $Y = F_n[\overline{X}_1, \overline{X}_2]$ y por tanto

$$\begin{aligned} \alpha_{\mathbb{T}\cap\mathbb{SY}} &= \sup_{\substack{\|E\|_2=1 \\ E\in\mathbb{T}\cap\mathbb{SY}}} \text{máx}\left(\rho(EX_1X_1^TF_n), \rho(EX_2X_2^TF_n)\right) \\ &= \sup_{\substack{\|E\|_2=1 \\ E\in\mathbb{T}\cap\mathbb{SY}}} \text{máx}\left(\rho(EX_1X_1^T), \rho(EX_2X_2^T)\right). \end{aligned}$$

De $X_2^HX_1 = X_2^HF_nF_nX_1 = -X_2^HX_1$ se sigue que $X_2^HX_1 = 0$ y por lo tanto

$$\|XX^T\|_2 = \|[X_1, X_2][X_1, X_2]^T\|_2 = \text{máx}(\|X_1X_1^T\|_2, \|X_2X_2^T\|_2).$$

Supongamos que $\|X_1X_1^T\|_2 \geq \|X_2X_2^T\|_2$ (el otro caso se trata análogamente) y sea $X_1X_1^T = U\Sigma U^T$ una factorización de Takagi. Entonces $U = -F_nU$ y, por [82, Lema 2.4], existe una matriz Toeplitz simétrica E tal que $\|E\|_2 = 1$ y $Eu_1 = \overline{u}_1$. La demostración de (II) se completa usando de nuevo el Lema 2.1.5.

Los apartados (III) y (IV) son obvios si observamos que $\lambda \in \mathbb{R}$ implica $X \in \mathbb{R}^{n\times r_1}$, y por tanto las perturbaciones construidas antes pueden elegirse reales [82].

□

El Teorema 2.2.4 puede extenderse fácilmente a matrices Hankel.

Corolario 2.2.5 *Sean $\mathbb{HA}$ y $\mathbb{PS}$ los conjuntos de matrices Hankel y persimétricas, respectivamente. Entonces para la norma espectral tenemos que:*

(I). $\kappa(A, \lambda; \mathbb{HA}) = \kappa(A, \lambda) = (n_1, \|XX^T\|_2)$ *si* $A \in \mathbb{HA}$;

(II). $\kappa(A, \lambda; \mathbb{HA} \cap \mathbb{PS}) = \kappa(A, \lambda) = (n_1, \|XX^T\|_2)$ *si* $A \in \mathbb{HA} \cap \mathbb{PS}$;

(III). $\kappa(A, \lambda; \mathbb{HA} \cap \mathbb{R}^{n\times n}) = \kappa(A, \lambda) = (1, 1)$ *si* $A \in \mathbb{HA} \cap \mathbb{R}^{n\times n}$;

(IV). $\kappa(A, \lambda; \mathbb{HA} \cap \mathbb{PS} \cap \mathbb{R}^{n\times n}) = \kappa(A, \lambda) = (1, 1)$ *si* $A \in \mathbb{HA} \cap \mathbb{PS} \cap \mathbb{R}^{n\times n}$,

siendo X la matriz de autovectores derechos asociados a λ definida en (2.13).

Demostración:

Una matriz de Hankel es simétrica luego, usando de nuevo las formas canónicas para haces simétricos [89, 71], podemos elegir las matrices P y Q en la forma de Jordan (2.10) de modo que

$$Q = \text{diag}(F_{n_1}, \ldots, F_{n_1})P^T,$$

y, en consecuencia, $Y = \overline{X}$. El resto de la demostración es análogo a la demostración del Teorema 2.2.4.

□

2.2.5. Matrices simétricas, antisimétricas y hermíticas

La demostración del Teorema 2.2.4 sólo hace uso de la *persimetría* de las matrices Toeplitz. Si denotamos el conjunto de las matrices persimétricas como $\mathbb{PS}$, el mismo argumento demuestra que $\kappa(A,\lambda;\mathbb{PS}) = \kappa(A,\lambda)$ para las matrices $A \in \mathbb{PS}$. En un marco más amplio, tenemos el siguiente resultado, que incluye, aparte de las matrices persimétricas ($M = F_n$), clases de matrices estructuradas como las simétricas($M = I_n$) y las pseudosimétricas ($M = I_p \oplus (-I_q)$ con $p+q=n$) (véanse las definiciones al comienzo de la sección 2.4.2).

Teorema 2.2.6 *Dada una matriz simétrica y ortogonal $M \in \mathbb{R}^{n\times n}$, sea $\mathbb{S} = \{A \in \mathbb{C}^{n\times n} : A^TM = MA\}$. Para cualquier matriz $A \in \mathbb{S}$ y cualquier norma unitariamente invariante se tiene*

(I). $\kappa(A,\lambda;\mathbb{S}) = \kappa(A,\lambda) = (n_1, \|XX^T\|_2)$;

(II). $\kappa(A,\lambda;\mathbb{S}\cap\mathbb{R}^{n\times n}) = (n_1, \alpha_{\mathbb{S}\cap\mathbb{R}^{n\times n}})$ *con* $\|XX^T\|_2/2 \le \alpha_{\mathbb{S}\cap\mathbb{R}^{n\times n}} \le \|XX^T\|_2$,

donde X es la matriz de autovectores derechos de A asociados a λ definida en (2.13).

Demostración:
Como en la demostración del Teorema 2.2.4, la forma canónica del haz simétrico $MA - \lambda M$ nos dice que $Y = M\overline{X}$. Sea $XX^T = U\Sigma U^T$ una factorización de Takagi. Llamamos $u_1 = Ue_1$ y $E = M\overline{u}_1 u_1^H$, de modo que $\|E\| = 1$ y

$$\alpha_{\mathbb{S}} \ge \rho(EXX^TM) = \rho(Mu_1u_1^TXX^TM) = \rho(u_1^TXX^Tu_1) = \|XX^T\|_2 = \alpha,$$

Esto completa la demostración de la primera parte. La segunda parte es análoga a la demostración del Lema 2.2.2 (i).

□

Usando la terminología de [55], el Teorema 2.2.6 se ocupa de las álgebras de Jordan asociadas a la forma bilineal simétrica $\langle x,y\rangle = x^TMy$. Para las correspondientes álgebras de Lie, dadas por $\mathbb{S} = \{A \in \mathbb{C}^{n\times n} : A^TM = -MA\}$, es sabido que los números de condición estructurado y no estructurado para autovalores simples pueden diferir de manera considerable [55, 82]. De hecho, ya hemos visto en el ejemplo (2.9) (con $M = I$) que puede darse $n_{\mathbb{S}} < n_1$, esto es, que los autovalores múltiples pueden tener un comportamiento *cualitativamente mejor* frente a perturbaciones estructuradas. El siguiente resultado identifica una de las situaciones en que esto sucede, y demuestra que esto sólo ocurre bajo condiciones muy específicas: para autovalores nulos que tienen asociado un único bloque de Jordan de tamaño máximo y de dimensión impar. Además, el Teorema 2.2.7 pone de manifiesto las diferencias que se pueden esperar entre $\alpha_{\mathbb{S}}$ y α cuando $n_{\mathbb{S}} = n_1$.

Teorema 2.2.7 *Dada una matriz simétrica y ortogonal $M \in \mathbb{R}^{n\times n}$ sea $\mathbb{S} = \{A \in \mathbb{C}^{n\times n} : A^TM = -MA\}$. Entonces el número de condición estructurado $\kappa(A,\lambda;\mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$ para $A \in \mathbb{S}$ y la norma espectral cumple que:*

(I). *Si $\lambda = 0$, n_1 es impar y $r_1 = 1$, entonces $n_{\mathbb{S}} < n_1$;*

(II). *Si $\lambda = 0$, n_1 es impar y $r_1 > 1$, entonces $n_{\mathbb{S}} = n_1$ y $\alpha_{\mathbb{S}} = \sqrt{\sigma_1\sigma_2}$, donde σ_1, σ_2 son los dos mayores valores singulares de la matriz XX^T (mientras que $\alpha = \sigma_1$);*

(III). *Si $\lambda = 0$ y n_1 es par, entonces r_1 es par, $n_{\mathbb{S}} = n_1$ y $\alpha_{\mathbb{S}} = \alpha = \left\| X \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^T \right\|_2$;*

(IV). *Si $\lambda \neq 0$ y $r_1 = 1$, entonces $n_{\mathbb{S}} = n_1$ y $\alpha_{\mathbb{S}} = \sqrt{\|X\|_2^2\|Y\|_2^2 - |Y^T M X|^2}$.*

(V). *Si $\lambda \neq 0$ y $r_1 > 1$, entonces $n_{\mathbb{S}} = n_1$,*

donde X e Y son las matrices de autovectores de A asociados a λ definidas en (2.13) *y* (2.14)*, respectivamente.*

Demostración:

La forma canónica estructurada de $A \in \mathbb{S}$ puede extraerse fácilmente de la forma canónica del haz antisimétrico/simétrico $MA - \lambda M$ (véase, por ejemplo, [71, Teorema 7.3]). En particular, si $\lambda = 0$ es un autovalor y n_1 es impar, entonces las matrices P, Q en (2.10) cumplen que

$$Q = \operatorname{diag}(\Sigma_{n_1}, \ldots, \Sigma_{n_1}) P^T M.$$

Por tanto, $Y = M\overline{X}$ y si $n_{\mathbb{S}}$ fuese igual a n_1 existiría alguna perturbación $E \in \mathbb{S}$ con $\rho(Y^H E X) = \rho(X^T M E X) > 0$. Esto es imposible si $r_1 = 1$, ya que X es un vector y ME es una matriz antisimétrica y por tanto $\rho(X^T M E X) = |X^T M E X| = 0$. Esto demuestra (I).

Para demostrar que $n_{\mathbb{S}} = n_1$ cuando $\lambda = 0$, n_1 impar y $r_1 > 1$, basta construir una perturbación $E \in \mathbb{S}$ tal que $\rho(X^T M E X) > 0$. Para ello, consideramos una factorización de Takagi $XX^T = U\Sigma U^T$, donde $U = [u_1, \ldots, u_{r_1}]$ tiene columnas ortonormales y $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{r_1})$ con $\sigma_1 \geq \cdots \geq \sigma_{r_1} > 0$. Tomando $E = M[\overline{u}_1, \overline{u}_2] \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right] [u_1, u_2]^H$ tenemos que $E \in \mathbb{S}$, $\|E\|_2 = 1$ y

$$\alpha_{\mathbb{S}} \geq \rho(X^T M E X) = \rho\left(\begin{pmatrix} 0 & \sigma_2 \\ -\sigma_1 & 0 \end{pmatrix} \right) = \sqrt{\sigma_1\sigma_2} > 0.$$

Por otro lado, denotando por $\mathbb{SK}$ al conjunto de matrices complejas antisimétricas, tenemos que

$$\begin{aligned} \alpha_{\mathbb{S}} &= \sup_{\substack{\|\widetilde{E}\|_2 \leq 1 \\ \widetilde{E} \in \mathbb{SK}}} \rho(\widetilde{E} X X^T) = \sup_{\substack{\|G\|_2 \leq 1 \\ G \in \mathbb{SK}}} \rho(G\Sigma) = \sup_{\substack{\|G\|_2 \leq 1 \\ G \in \mathbb{SK}}} \rho(\Sigma^{1/2} G \Sigma^{1/2}) \\ &\leq \sup_{\substack{\|G\|_2 \leq 1 \\ G \in \mathbb{SK}}} \|\Sigma^{1/2} G \Sigma^{1/2}\|_2 = \sup_{\substack{\|G\|_2 \leq 1 \\ G \in \mathbb{SK}}} \|\widetilde{\Sigma} \circ G\|_2, \end{aligned}$$

donde $\widetilde{\Sigma} = [\sqrt{\sigma_i \sigma_j}]_{i,j=1}^{r_1}$ y $\circ$ denota el producto de Hadamard, esto es elemento a elemento, entre matrices. Un resultado de Mathias [70, Corolario 2.6] implica que $\|\widetilde{\Sigma} \circ G\|_2 \leq \sqrt{\sigma_1\sigma_2}\|G\|_2$, lo cual concluye la demostración de (II).

Si $\lambda = 0$ y n_1 es par, puede verse usando [71, Theorem 7.3] que r_1 es también par e $Y = M\overline{X}\begin{bmatrix} 0 & I_{r_1/2} \\ -I_{r_1/2} & 0 \end{bmatrix}$. Para alcanzar $\rho(Y^H E X) = \alpha = \left\| X \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^T \right\|_2$ se puede usar una perturbación de Lidskii como en (2.17), en este caso $E = \frac{1}{\alpha} M\overline{X}\begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^H \in \mathbb{S}$. Esto prueba (III).

Para el caso de un autovalor no nulo la forma canónica estructurada de Jordan de una matriz $A \in \mathbb{S}$ parece no revelar ninguna estructura particular en las matrices X, Y. Por otro lado, $-\lambda$ es también un autovalor de la matriz A con la misma forma de Jordan que el autovalor λ. De hecho, si denotamos por $\widetilde{X}$, $\widetilde{Y}$ las matrices de autovectores derechos e izquierdos asociadas a los bloques de Jordan de tamaño n_1 correspondientes al autovalor $-\lambda$, entonces $\widetilde{X} = M\overline{Y}$ e $\widetilde{Y} = -M\overline{X}$. Esto no sólo implica $\kappa(A, -\lambda) = \kappa(A, \lambda)$ y que $\kappa(A, -\lambda; \mathbb{S}) = \kappa(A, \lambda; \mathbb{S})$, sino que además $[X, M\overline{Y}]$ es una matriz de rango completo. Para el caso particular en que $r_1 = 1$ se tiene que

$$|Y^H E X| = \rho\left([M\overline{Y}, X]^T M E [M\overline{Y}, X]\right) = \rho\left(M E [M\overline{Y}, X][M\overline{Y}, X]^T\right)$$

para cualquier perturbación $E \in \mathbb{S}$. Usando los argumentos de la demostración de (II), tenemos que $\alpha_{\mathbb{S}} = \sqrt{\sigma_1 \sigma_2}$, donde σ_1 y σ_2 son los dos mayores valores singulares de la matriz simétrica $[M\overline{Y}, X][M\overline{Y}, X]^T$. Cálculos laboriosos, aunque elementales, revelan que $\sigma_1\sigma_2 = \|X\|_2^2\|Y\|_2^2 - |Y^T M X|^2$, lo que demuestra (IV). Esta técnica, sin embargo, no es válida para el caso $r_1 > 1$. Sí que podemos demostrar en ese caso que $\alpha_{\mathbb{S}} > 0$, aunque no parece fácil conseguir una cota superior o inferior de $\alpha_{\mathbb{S}}$. El rango completo de la matriz $[X, M\overline{Y}]$ implica la existencia de una matriz invertible L tal que

$$L^{-1}[X, M\overline{Y}] = \begin{pmatrix} I_{r_1} & \star \\ 0 & I_{r_1} \\ 0 & 0 \end{pmatrix}$$

Tomando $E = ML^{-T}([0, I_{r_1}, 0][I_{r_1}, 0, 0]^T - [I_{r_1}, 0, 0][0, I_{r_1}, 0]^T)L^{-1} \in \mathbb{S}$ tenemos que

$$\rho(Y^H E X) = \rho(I_r) = 1$$

y por lo tanto $\alpha_{\mathbb{S}} > 0$, completando la demostración de (V).

□

Observemos que el Teorema 2.2.7 (IV) no sólo corrobora los resultados de [55, Teorema 4.3] y [82, Teorema 3.2], que se limitan a acotar los números de condición, sino que además ofrece fórmulas explícitas para el número de condición estructurado de autovalores simples no nulos.

Afortunadamente, el estudio de los números de condición estructurados es más sencillo para el caso de álgebras de Jordan y de Lie asociadas con formas sesquilineales $\langle x, y\rangle = x^H M y$.

Lema 2.2.8 *Dada una matriz ortogonal simétrica o antisimétrica $M \in \mathbb{R}^{n\times n}$, sea $\mathbb{S} = \{A \in \mathbb{C}^{n\times n} : A^H M = \gamma M A\}$ para $\gamma \in \{1, -1\}$ fijo. Entonces para cualquier matriz $A \in \mathbb{C}^{n\times n}$ se tiene que $\kappa(A, \lambda; \mathbb{S}) = \kappa(A, \lambda)$ en la norma espectral.*

Demostración:
Sea $XY^H = U\Sigma V^H$ una descomposición en valores singulares y conideramos los vectores $u_1 = Ue_1$, $v_1 = Ve_1$, siendo e_1 la primera columna de la matriz identidad. Entonces $\|u_1\|_2 = \|v_1\|_2 = 1$ y por [68, Theorem 8.6] podemos encontrar una matriz H tal que $\|H\|_2 = 1$ y $Hu_1 = \mu v_1$ para algún $\mu \in \mathbb{C}$ con $|\mu| = 1$. Tomando la perturbación $E = \sqrt{\gamma}M^T H$ se tiene que $E \in \mathbb{S}$, $\|E\|_2 = 1$ y $\alpha_{\mathbb{S}} \geq \rho(EXY^H) = \rho(HXY^H M) = \rho(V^H HU\Sigma) \geq \alpha$ por el Lema 2.1.5.

□

2.2.6. Matrices J–simétricas y J–antisimétricas

Si la matriz M del producto escalar es

$$J_{2n} = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}, \tag{2.22}$$

la estructura $\mathbb{S} = \{A \in \mathbb{C}^{2n\times 2n} : \ A^H M = \gamma MA\}$ considerada en el Lema 2.2.8 coincide con el conjunto de matrices antihamiltonianas *complejas* si $\gamma = 1$, y con el conjunto de matrices hamiltonianas *complejas* si $\gamma = -1$. Los siguientes teoremas se refieren a estructuras similares de la forma $\mathbb{S} = \{A \in \mathbb{C}^{2n\times 2n} : \ A^T J_{2n} = \gamma J_{2n} A\}$. En particular, se obtienen cotas para los números de condición estructurados de matrices antihamiltonianas y hamiltonianas *reales*.

Teorema 2.2.9 *Dada una matriz ortogonal y antisimétrica $M \in \mathbb{R}^{2n\times 2n}$, sea $\mathbb{S} = \{A \in \mathbb{C}^{2n\times 2n} : \ A^T M = MA\}$. Para cada $A \in \mathbb{S}$ y cada autovalor λ de A se tienen los siguientes resultados en la norma espectral:*

(I). $\kappa(A, \lambda; \mathbb{S}) = \kappa(A, \lambda) = (n_1, \|XJ_{r_1}X^T\|_2)$;

(II). $\kappa(A, \lambda; \mathbb{S} \cap \mathbb{R}^{2n\times 2n}) = (n_1, \alpha_{\mathbb{S}\cap\mathbb{R}^{2n\times 2n}})$ *con* $\|XJ_{r_1}X^T\|_2/4 \leq \alpha_{\mathbb{S}\cap\mathbb{R}^{2n\times 2n}} \leq \|XJ_{r_1}X^T\|_2$,

donde X es la matriz de autovectores derechos asociados a λ definida en (2.13) *y J_{r_1} viene dada por* (2.22).

Demostración: La forma canónica estructurada de A se sigue del hecho de que cualquier matriz antihamiltoniana tiene una forma canónica diagonal por bloques [34, 51]. Más concretamente, de [71, Theorem 8.2] se sigue que r_1 es par y que las matrices P y Q en la forma de Jordan (2.10) pueden elegirse de modo que

$$Q = GJ_{r_1 n_1}^T GP^T M,$$

donde $G = \mathrm{diag}(I_{r_1 n_1/2}, F_{n_1}, \ldots, F_{n_1})$, la matriz $J_{r_1 n_1}$ viene dada por (2.22) y la matriz F_{n_1} por (2.18). Por tanto, $Y = M^T \overline{X} J_{r_1}$ y podemos usar una perturbación del tipo (2.17), en este caso

$$E = (M^T \overline{X} J_{r_1} X^H)/\|XJ_{r_1}X^T\|_2,$$

que cumple $E \in \mathbb{S}$, $\|E\|_2 = 1$ y es tal que

$$\alpha_{\mathbb{S}} \geq \rho(J_{r_1}^T X^T M E X)/\|X J_{r_1} X^T\|_2 = \|X J_{r_1} X^T\|_2 = \alpha.$$

Para probar la segunda parte, sean $u = u_R + \imath u_I$ y $v = v_R + \imath v_I$, con $u_R, u_I, v_R, v_I \in \mathbb{R}^{2n}$, siendo u y v vectores singulares izquierdo y derecho asociados al mayor valor singular de la matriz $K = X J_{r_1} X^T$. Entonces

$$\alpha_{\mathbb{S}} = \|K\|_2 = u^H K v = (u_R^T K v_R + u_I^T K v_I) + \imath(u_R^T K v_I - u_I^T K v_R).$$

Al menos uno de los cuatro términos en esta suma será mayor o igual en valor absoluto que $\alpha_{\mathbb{S}}/4$. Elegimos este término y definimos la matriz $W = [w_1, w_2] \in \mathbb{R}^{2n\times 2}$ formada por los dos vectores que definen este término. Por ejemplo si $|u_I^T K v_R| \geq \alpha_{\mathbb{S}}/4$, entonces $W = [u_I, v_R]$. Como la matriz K es antisimétrica, podemos suponer que los vectores u y v cumplen $v^T u = 0$, lo que implica $\|W\|_2 \leq 1$. Tomando la perturbación $E = M^T W J_2 W^T \in \mathbb{S} \cap \mathbb{R}^{2n\times 2n}$, se tiene $\|E\|_2 \leq 1$ y

$$\begin{aligned} \alpha_{\mathbb{S}\cap\mathbb{R}^{2n\times 2n}} &\geq \rho(J_{r_1/2}^T X^T M E X) = \rho(J_2 W^T K W) \\ &= \rho(\mathrm{diag}(w_2^T K w_1, -w_1^T K w_2)) \geq \alpha_{\mathbb{S}}/4, \end{aligned}$$

lo que completa la demostración.

□

Teorema 2.2.10 *Dada una matriz ortogonal antisimétrica $M \in \mathbb{R}^{2n\times 2n}$, sea $\mathbb{S} = \{A \in \mathbb{C}^{2n\times 2n} : A^T M = -MA\}$. Entonces para cada $A \in \mathbb{C}^{n\times n}$ y cada autovalor λ de A se tiene que*

(I). $\kappa(A, \lambda; \mathbb{S}) = (n_1, \alpha_{\mathbb{S}})$ con $\alpha/\sqrt{2} \leq \alpha_{\mathbb{S}} \leq \alpha$ para la norma de Frobenius y $\alpha_{\mathbb{S}} = \alpha$ para la norma $||\cdot||_2$;

(II). $\kappa(A, \lambda; \mathbb{S} \cap \mathbb{R}^{2n\times 2n}) = (n_1, \alpha_{\mathbb{S}\cap\mathbb{R}^{2n\times 2n}})$ con $\alpha/8 \leq \alpha_{\mathbb{S}\cap\mathbb{R}^{2n\times 2n}} \leq \alpha_{\mathbb{S}}$ para la norma $||\cdot||_2$.

Demostración:

Sean u_1, v_1 los vectores singulares izquierdo y derecho, respectivamente, asociados al mayor valor singular de la matriz $XY^H M$. Definimos

$$\widetilde{E} = [v_1, \overline{u}_1] \begin{bmatrix} 0 & 1 \\ 1 & -v_1^T u_1 \end{bmatrix} [v_1, \overline{u}_1]^T.$$

Entonces $\widetilde{E}$ es una matriz simétrica y puede demostrarse que $\|\widetilde{E}\|_F = \sqrt{2 - |u_1^T v_1|^2} \leq \sqrt{2}$ (véase [55, Theorem 4.3]). Eligiendo la perturbación $E = M\widetilde{E}/\sqrt{2}$ tenemos que $E \in \mathbb{S}$, $\|E\|_F \leq 1$ y

$$\alpha_{\mathbb{S}} \geq \rho(Y^H E X) = \rho(\widetilde{E} X Y^H M)/\sqrt{2} \geq \|XY^H\|_2/\sqrt{2},$$

donde hemos aplicado el Lema 2.1.5. Para la norma espectral, el Teorema 5.8 de [69] nos dice que existe una matriz simétrica $\widetilde{E}$ que transforma el vector u_1 en v_1 con $\|\widetilde{E}\|_2 = 1$. Así, tomando como perturbación $E = M\widetilde{E}$ queda demostrada la segunda parte de (I).

Para demostrar (II) descomponemos la matriz $XY^HM = S + W$, donde

$$S = (XY^H + \overline{Y}X^T)/2$$

y

$$W = (XY^H - \overline{Y}X^T)/2.$$

Entonces $\alpha = \|XY^H\|_2 \leq \|S\|_2 + \|W\|_2$. Distinguimos dos casos, dependiendo de si la parte antisimétrica W domina o no a la parte simétrica S.

1. Si $\|S\|_2 \geq \|W\|_2/3$, descomponemos $S = S_R + \imath S_I$ con S_R, S_I matrices simétricas reales. Entonces $\|S_R\|_2 \geq \|S\|_2/2$ ó $\|S_I\|_2 \geq \|S\|_2/2$. En el primer caso, sea u_1 un autovector ortonormalizado asociado al autovalor de la matriz S_R cuyo valor absoluto coincide con $\|S_R\|_2$. Entonces $E = Mu_1u_1^T \in \mathbb{S} \cap \mathbb{R}^{2n\times 2n}$ con $\|E\|_2 = 1$ cumple que

$$\begin{aligned} \alpha_{\mathbb{S}\cap\mathbb{R}^{n\times n}} &\geq \rho(EXY^H) = |u_1^TXY^HMu_1| = |u_1^TSu_1| \geq |u_1^TS_Ru_1| \\ &\geq \frac{\|S\|_2}{2} = \frac{\|S\|_2 + 3\|S\|_2}{8} \geq \frac{\|S\|_2 + \|W\|_2}{8} \geq \frac{\alpha}{8}. \end{aligned}$$

   El caso $\|S_I\|_2 \geq \|S\|_2/2$ puede demostrarse análogamente.

2. Si $\|S\|_2 \leq \|W\|_2/3$, descomponemos $W = W_R + \imath W_I$ con W_R, W_I matrices antisimétricas reales. Supongamos que $\|W_R\|_2 \geq \|W\|_2/2$ (de nuevo, el caso $\|W_I\|_2 \geq \|W\|_2/2$ es análogo). Sean u_1, v_1 los vectores singulares izquierdo y derecho, respectivamente, asociados al mayor valor singular de la matriz W_R. Como W_R es antisimétrica se tiene que $v_1^Tu_1 = 0$. Eligiendo la perturbación

$$E = M[u_1, v_1] \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} [u_1, v_1]^T \in \mathbb{S} \cap \mathbb{R}^{2n\times 2n},$$

   que cumple $\|E\|_2 = 1$, se tiene

$$\alpha_{\mathbb{S}\cap\mathbb{R}^{2n\times 2n}} \geq \rho(EXY^H) = \rho\left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} [u_1, v_1]^T(S + W)[u_1, v_1]\right) = \rho(\Phi),$$

   donde

$$\Phi = \begin{pmatrix} -\beta & 0 \\ 0 & \beta \end{pmatrix} + \begin{pmatrix} u_1^TSv_1 & v_1^TSv_1 \\ u_1^TSu_1 & u_1^TSv_1 \end{pmatrix}$$

   con $\beta = \|W_R\|_2 + \imath u_1^TW_Iv_1$. Además, $\det(\Phi) = -(\beta + \gamma)(\beta - \gamma)$ con

$$\gamma = \sqrt{(u_1^TSu_1)(v_1^TSv_1) - (u_1^TSv_1)^2}$$

cumpliendo $|\gamma| \leq \|S\|_2$. Esto demuestra que

$$\begin{aligned}\rho(\Phi) &\geq |\beta| - |\gamma| \geq \|W_R\|_2 - \|S\|_2 \geq \frac{\|W\|_2}{2} - \|S\|_2 \\ &= \frac{\|W\|_2}{2} - \frac{9}{8}\|S\|_2 + \frac{1}{8}\|S\|_2 \geq \frac{\|S\|_2 + \|W\|_2}{8} \geq \frac{\alpha}{8},\end{aligned}$$

y concluye la demostración.

□

El Teorema 2.2.10 (II) nos dice que el restringir las perturbaciones al conjunto de matrices hamiltonianas reales puede tener un efecto positivo, aunque limitado, sobre la sensibilidad de autovalores múltiples. Por otro lado, debemos señalar que los números de condición no proporcionan información alguna sobre la dirección en la cual se mueven los autovalores perturbados. Esta es una cuestión crucial a la hora de decidir si un autovalor imaginario puro de una matriz hamiltoniana permanece o no sobre el eje imaginario cuando se le somete a perturbaciones estructuradas, algo que es importante en numerosas aplicaciones. Para resultados relacionados véase [8] y las referencias que contiene.

2.3. Perturbaciones estructuradas genéricas: problemas generalizados

Consideramos en esta sección dos extensiones del problema usual de autovalores: los haces de matrices y los polinomios matriciales.

2.3.1. Haces de matrices

Un *haz de matrices* (en inglés, *matrix pencil*) es una familia de matrices de la forma $\{A - \lambda B : \lambda \in \mathbb{C}\}$ con $A, B \in \mathbb{C}^{n\times n}$. Se dice que un haz es *regular* si existe un valor de λ tal que $\det(A - \lambda B) \neq 0$. En caso contrario, se dice que el haz es *singular*. Nosotros sólo trataremos el caso de haces regulares.

Se suele identificar el haz $A - \lambda B$ con el par de matrices (A, B) y se dice que $\lambda \in \mathbb{C}$ es un *autovalor finito* del haz (A, B) si $\det(A - \lambda B) = 0$. Cuando la matriz B es singular, el haz (A, B) tiene uno o varios *autovalores infinitos*, que se identifican con los autovalores nulos del *haz dual* (B, A). Si λ_0 es un autovalor finito del haz $A - \lambda B$, decimos que $x \in \mathbb{C}^n$ es un autovector derecho asociado a λ_0 si $(A - \lambda_0 B)x = 0$, y que $y \in \mathbb{C}^n$ es un autovector izquierdo de (A, B) asociado a λ_0 si $y^H(A - \lambda_0 B) = 0$. Los autovectores asociados a un autovalor infinito de (A, B) son los asociados al autovalor nulo de (B, A). Nótese que el problema de hallar los autovalores y los autovectores de un haz de matrices (el llamado *problema generalizado de autovalores*) incluye como caso particular al problema usual de autovalores (basta tomar $B = I_n$). Dado un autovalor finito del haz (A, B) diremos que es *simple* si el rango de $A - \lambda B$ es $n - 1$. En caso contrario, diremos que el autovalor es *múltiple*.

De forma análoga al problema usual de autovalores, consideramos perturbaciones de la forma $(A+\epsilon E, B+\epsilon F)$ a un haz (A,B), donde $E,F\in\mathbb{C}^{n\times n}$, y estudiamos la sensibilidad de los autovalores de (A,B) en el límite cuando ϵ tiende a cero.

En un contexto más general, Langer y Najman emplean en [61, 62, 63] la forma local de Smith para extender la teoría de perturbaciones de Lidskii, obteniendo desarrollos asintóticos de los autovalores perturbados de funciones matriciales analíticas $L(\lambda)$. Recientemente, de Terán, Dopico y Moro [88] han investigado el caso especial en que $L(\lambda)$ es un haz $A-\lambda B$, reemplazando la forma local de Smith por la *forma canónica de Weierstrass*, mucho más natural para haces de matrices, y que pasamos a describir a continuación: para todo haz regular $A-\lambda B$ existen matrices cuadradas no singulares $\mathcal{Q}$ y $\mathcal{P}$ tales que

$$\mathcal{Q}(A-\lambda B)\mathcal{P}=\operatorname{diag}(\mathcal{J}-\lambda \mathrm{I}\ ,\ \lambda\mathcal{J}_\infty-\mathrm{I}), \tag{2.23}$$

donde $\mathcal{J}$ es suma directa de bloques de Jordan (2.12) asociados a los autovalores finitos del haz, y $\mathcal{J}_\infty$ es suma directa de bloques de Jordan nilpotentes

$$\begin{pmatrix} 0 & 1 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 1 \\ & & & & 0 \end{pmatrix} \tag{2.24}$$

asociados al autovalor infinito. Al mayor tamaño de estos bloques nilpotentes se le llama *índice de nilpotencia del haz* (A,B). La descomposición (2.23) es conocida como *forma canónica de Weierstrass* del haz (A,B) (véase, por ejemplo, [87, §VI.1.2]).

Dado un autovalor finito λ de un haz regular (A,B), la forma canónica de Weierstrass puede escribirse de la forma

$$\left(\begin{array}{c|c} J & 0 \\ \hline 0 & \widetilde{J}_A \end{array}\right)=\begin{pmatrix} Q \\ \hline \widetilde{Q} \end{pmatrix} A \left(\begin{array}{c|c} P & \widetilde{P} \end{array}\right), \quad \left(\begin{array}{c|c} I & 0 \\ \hline 0 & \widetilde{J}_B \end{array}\right)=\begin{pmatrix} Q \\ \hline \widetilde{Q} \end{pmatrix} B \left(\begin{array}{c|c} P & \widetilde{P} \end{array}\right), \tag{2.25}$$

donde tanto $(P,\widetilde{P})$ como $\left(\begin{smallmatrix} Q \\ \widetilde{Q} \end{smallmatrix}\right)$ son matrices invertibles, y J contiene los r_1 bloques de Jordan asociados a λ de tamaño máximo n_1 . Análogamente, para un autovalor infinito del haz (A,B), tenemos

$$\left(\begin{array}{c|c} I & 0 \\ \hline 0 & \widetilde{J}_A \end{array}\right)=\begin{pmatrix} Q \\ \hline \widetilde{Q} \end{pmatrix} A \left(\begin{array}{c|c} P & \widetilde{P} \end{array}\right), \quad \left(\begin{array}{c|c} N & 0 \\ \hline 0 & \widetilde{J}_B \end{array}\right)=\begin{pmatrix} Q \\ \hline \widetilde{Q} \end{pmatrix} B \left(\begin{array}{c|c} P & \widetilde{P} \end{array}\right), \tag{2.26}$$

donde N contiene r_1 bloques de Jordan nilpotentes N_{n_1} de la forma (2.24). Al igual que en el problema usual de autovalores, reunimos en las matrices X e Y autovectores derechos e izquierdos contenidos en las matrices P y Q:

$$\begin{aligned} X &= \left(\ Pe_1,\ Pe_{n_1+1},\ \ldots,\ Pe_{(r_1-1)n_1+1}\ \right), \\ Y &= \left(\ Q^He_{n_1},\ Q^He_{2n_1},\ \ldots,\ Q^He_{r_1n_1}\ \right). \end{aligned} \tag{2.27}$$

Obviamente, la relación entre X, Y y P, Q impone una normalización sobre las matrices X, Y. Para $n_1 = 1$ tenemos que $Y^H BX = I$ si λ es finito e $Y^H AX = I$ si λ es infinito. Para $n_1 > 1$ tenemos la igualdad $Y^H AX = Y^H BX = 0$ para cualquier autovalor λ.

Los siguientes teoremas resumen los principales resultados de [88] en cuanto a desarrollos para autovalores perturbados de haces regulares.

Teorema 2.3.1 *Sea λ un autovalor finito de un haz regular (A, B) y sean*

$$(E, F) \in \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n}$$

matrices tales que $Y^H(E - \lambda F)X$ es invertible, con X e Y definidas como en (2.27). Entonces existen $n_1 r_1$ autovalores $\widehat{\lambda}_k$ del haz perturbado $(A + \epsilon E, B + \epsilon F)$ que admiten un desarrollo

$$\widehat{\lambda}_k = \lambda + (\xi_k)^{1/n_1}\epsilon^{1/n_1} + o(\epsilon^{1/n_1}), \quad k = 1, \ldots, r_1 \tag{2.28}$$

en potencias fraccionarias, donde $\xi_1, \ldots, \xi_{r_1}$ son los autovalores de la matriz $Y^H(E - \lambda F)X$. Para autovalores infinitos del haz (A, B), sea $F \in \mathbb{C}^{n\times n}$ tal que la matriz $Y^H FX$ es invertible. Entonces existen $n_1 r_1$ autovalores $\widehat{\lambda}_k$ del haz perturbado $(A+\epsilon E, B+\epsilon F)$ que admiten el desarrollo en potencias fraccionarias

$$\frac{1}{\widehat{\lambda}_k} = (\xi_k)^{1/n_1}\epsilon^{1/n_1} + o(\epsilon^{1/n_1}), \quad k = 1, \ldots, r_1, \tag{2.29}$$

donde ξ_k son los autovalores de $Y^H FX$ para $k \in \{1, \ldots, r_1\}$.

2.3.1.a. Números de condición no estructurados para haces de matrices

Basándonos en el Teorema 2.3.1, podemos definir un número de condición para autovalores múltiples de un haz regular de matrices como sigue.

Definición 2.3.2 *Sea λ un autovalor finito del haz regular (A, B). El* **número de condición (absoluto)de Hölder** *de λ viene dado por*

$$\kappa(A, B, \lambda) = (n_1, \alpha),$$

donde n_1 es la dimensión del mayor bloque de Jordan asociado a λ en la forma canónica de Weierstrass de (A, B) y

$$\alpha = \sup_{\substack{||E||\leq w_A, ||F||\leq w_B \\ E,F\in\mathbb{C}^{n\times n}}} \rho(Y^H(E - \lambda F)X),$$

donde X e Y vienen dadas por (2.27). El número de condición (absoluto) de Hölder *para un autovalor $\lambda = \infty$ de (A, B) viene dado por $\kappa(A, B, \infty) = (n_1, \alpha)$, donde n_1 es el índice de nilpotencia del haz (A, B) y*

$$\alpha = \sup_{\substack{||F||\leq w_B \\ F\in\mathbb{C}^{n\times n}}} \rho(Y^H FX).$$

Nótese que la Definición 2.3.2 no sólo depende de la norma matricial, sino que también depende de la elección de los pesos no negativos w_A y w_B. Estamos suponiendo implícitamente que estas cantidades son estrictamente positivas cero, ya que, en caso contrario, sería $\kappa(A, B, \lambda) = (0, 0)$. Más concretamente, supondremos que $w_A > 0$ si $\lambda = 0$, $w_B > 0$ si $\lambda = \infty$ y máx$\{w_A, w_B\} > 0$ para cualquier otro autovalor. Los pesos w_A y w_B se introducen en la Definición 2.3.2 para equilibrar la influencia de las perturbaciones de las matrices A y B. Por ejemplo, si tanto E y F tienen norma pequeña comparada, respectivamente, con las normas de las matrices A y B, parece razonable escoger los pesos de la forma $w_A = \|A\|/\sqrt{\|A\|^2 + \|B\|^2}$ y $w_B = \|B\|/\sqrt{\|A\|^2 + \|B\|^2}$.

El lema siguiente representa una extensión directa de [72, Theorem 4.2].

Lema 2.3.3 *Para cualquier norma unitariamente invariante, se tiene que*

$$\kappa(A, B, \lambda) = (n_1, (w_A + w_B|\lambda|)\|XY^H\|_2)$$

para autovalores finitos y $\kappa(A, B, \lambda) = (n_1, w_B\|XY^H\|)$ *si* $\lambda = \infty$.

Demostración:
Por un lado

$$\begin{aligned} \rho(Y^H(E - \lambda F)X) &= \rho((E - \lambda F)XY^H) \leq \|(E - \lambda F)XY^H\|_2 \\ &\leq \|(E - \lambda F)XY^H\| \leq (w_A + w_B|\lambda|)\|XY^H\|_2 \end{aligned}$$

para perturbaciones E, F tales que $\|E\| \leq w_A, \|F\| \leq w_B$. Por tanto

$$\alpha \leq (w_A + w_B|\lambda|)\|XY^H\|_2.$$

Por otro lado, sean u_1, v_1 los vectores singulares izquierdo y derecho correspondiente al mayor valor singular de XY^H. Escogiendo las perturbaciones $E = w_A v_1 u_1^H$ y $F = -\frac{\bar{\lambda}}{|\lambda|} w_B v_1 u_1^H$ ($F = 0$ si $\lambda = 0$) tenemos que $\|E\| \leq w_A, \|F\| \leq w_B$ y

$$\alpha \geq \rho((w_A + w_B|\lambda|)v_1 u_1^H XY^H) =$$

$$= (w_A + w_B|\lambda|)\rho(u_1^H XY^H v_1) = (w_A + w_B|\lambda|)\|XY^H\|_2.$$

La demostración para $\kappa(A, B, \infty)$ es análoga.

□

La Definición 2.3.2 está basada en la distancia usual $|\widehat{\lambda} - \lambda|$ en $\mathbb{C}$. Para autovalores infinitos la distancia usual no es invariante bajo intercambios de las matrices A y B, es decir, que $|\widehat{\lambda} - \lambda| \neq |\frac{1}{\widehat{\lambda}} - \frac{1}{\lambda}|$. Un concepto más elegante de distancia es el que ofrece la *distancia cordal*

$$\chi(\widehat{\lambda}, \lambda) = \frac{|\widehat{\lambda} - \lambda|}{\sqrt{|\widehat{\lambda}|^2 + 1}\sqrt{|\lambda|^2 + 1}},$$

que incluye de forma natural autovalores infinitos

$$\chi(\widehat{\lambda}, \infty) = \lim_{|\mu| \to \infty} \chi(\widehat{\lambda}, \mu) = \frac{1}{\sqrt{|\widehat{\lambda}|^2 + 1}},$$

(véase [87] para más detalles). Sustituyendo los desarrollos (2.28) y (2.29) se tiene que

$$\chi(\widehat{\lambda}_k, \lambda) = \frac{(\xi_k)^{1/n_1}\epsilon^{1/n_1}}{\sqrt{|\lambda|^2+1}} + o(\epsilon^{1/n_1})$$

y

$$\chi(\widehat{\lambda}_k, \infty) = (\xi_k)^{1/n_1}\epsilon^{1/n_1} + o(\epsilon^{1/n_1}).$$

Esto demuestra que cuando se trabaja con la métrica cordal χ la cantidad α en la definición del número de condición de Hölder para un autovalor finito debe dividirse por $\sqrt{|\lambda|^2+1}$, mientrás que para un autovalor infinito es igual. Es fácil ver que esta modificación de número de condición tiene la propiedad de ser continua para $|\lambda| = \infty$.

El empleo de la métrica usual o de χ dependerá de la aplicación en la que estemos interesados. Si el objetivo es calcular un autovalor finito λ parece ser más relevante utilizar la métrica usual.

Todos los resultados que veremos a continuación están basados en la métrica usual de $\mathbb{C}$ pero como hemos visto por los comentarios anteriores, escribirlos en términos de la distancia cordal es inmediato.

2.3.1.b. Números de condición estructurados para haces de matrices

El número de condición de Hölder estructurado $\kappa(A, B, \lambda; \mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$ para un subconjunto $\mathbb{S} \subset \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n}$ puede definirse análogamente al del problema usual de autovalores. Como antes, suponemos que las cantidades w_A, w_B son positivas:

Definición 2.3.4 *Sea λ un autovalor del haz regular (A, B) con $A, B \in \mathbb{C}^{n\times n}$, y sea $\mathbb{S}$ un subconjunto de $\mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n}$. El* **número de condición estructurado de Hölder** *de λ con respecto a $\mathbb{S}$ se define como*

$$\kappa(A, B, \lambda; \mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}}),$$

donde $1/n_{\mathbb{S}}$ es el menor exponente posible γ_k de ϵ en los desarrollos

$$\widehat{\lambda}_k = \lambda + \alpha_k^{\gamma_k}\epsilon^{\gamma_k} + o(\epsilon), \qquad k = 1, \ldots, m,$$

de los autovalores del haz perturbado $(A + \epsilon E, B + \epsilon F)$ entre todas las perturbaciones $(E, F) \in \mathbb{S}$, y $\alpha_{\mathbb{S}} > 0$ es el mayor valor posible de α_k para toda $(E, F) \in \mathbb{S}$ con $\|E\| \leq w_A$, $\|F\| \leq w_B$.

En particular, si $n_{\mathbb{S}} = n_1$ entonces

$$\alpha_{\mathbb{S}} = \sup_{\substack{||E||\leq w_A, ||F||\leq w_B \\ (E,F)\in\mathbb{S}}} \rho(Y^H(E - \lambda F)X).$$

Algunas demostraciones de la Seccion 2.2.1 pueden extenderse al caso de autovalores de haces de matrices si la estructura es *separable*, es decir si $\mathbb{S} = \mathbb{S}_1 \times \mathbb{S}_2$ con $\mathbb{S}_1, \mathbb{S}_2 \subset \mathbb{C}^{n\times n}$. Esto se refleja en el siguiente teorema:

Teorema 2.3.5 *Sean $\mathbb{S}_1, \mathbb{S}_2$ dos subconjuntos de $\mathbb{C}^{n\times n}$ y sea λ un autovalor del haz regular $(A, B) \in \mathbb{C}^{n\times n}$. Sean*

$$\kappa(A, B, \lambda) = (n_1, \alpha), \qquad \kappa(A, B, \lambda; \mathbb{S}_1 \times \mathbb{S}_2) = (n_{\mathbb{S}_1\times\mathbb{S}_2}, \alpha_{\mathbb{S}_1\times\mathbb{S}_2}).$$

Entonces,

*(*I*).* Haces de matrices reales*: Si $\mathbb{S}_1 = \mathbb{S}_2 = \mathbb{R}^{n\times n}$, entonces $n_{\mathbb{S}_1\times\mathbb{S}_2} = n_1$ y*

$$\alpha/4 \leq \alpha_{\mathbb{S}_1\times\mathbb{S}_2} \leq \alpha$$

para matrices $A, B \in \mathbb{C}^{n\times n}$ y cualquier norma unitariamente invariante.

*(*II*).* Haces de matrices simétricas *: Si $\mathbb{S}_1 = \mathbb{S}_2 = \{A \in \mathbb{C}^{n\times n} : \ A^T = A\}$, entonces $n_{\mathbb{S}_1\times\mathbb{S}_2} = n_1$ y $\alpha_{\mathbb{S}_1\times\mathbb{S}_2} = \alpha$ para matrices $A, B \in \mathbb{S}_1$ y cualquier norma unitariamente invariante.*

*(*III*).* Haces de matrices simétricas reales*: Si $\mathbb{S}_1 = \mathbb{S}_2 = \{A \in \mathbb{R}^{n\times n} : \ A^T = A\}$ entonces $n_{\mathbb{S}_1\times\mathbb{S}_2} = n_1$ y $\alpha/2 \leq \alpha_{\mathbb{S}_1\times\mathbb{S}_2} \leq \alpha$ para matrices $A, B \in \mathbb{C}^{n\times n}$ con $A^T = A$, $B^T = B$ y cualquier norma unitariamente invariante.*

*(*IV*).* Haces de matrices simétricas/antisimétricas *: Si $\mathbb{S}_1 = \{A \in \mathbb{C}^{n\times n} : \ A^T = A\}$ y $\mathbb{S}_2 = \{B \in \mathbb{C}^{n\times n} : \ B^T = -B\}$ entonces los siguientes resultados son ciertos para haces de matrices $(A, B) \in \mathbb{S}_1 \times \mathbb{S}_2$ en la norma dos:*

a) *Si $\lambda = \infty$, n_1 es impar y $r_1 = 1$, entonces $n_{\mathbb{S}} < n_1$.*

b) *Si $\lambda = \infty$, n_1 es impar y $r_1 > 1$, entonces $n_{\mathbb{S}} = n_1$ y $\alpha_{\mathbb{S}} = w_B\sqrt{\sigma_1\sigma_2}$, donde σ_1, σ_2 son los dos mayores valores singulares de la matriz XX^T (mientras que $\alpha = w_B\sigma_1$).*

c) *Si $\lambda = \infty$ y n_1 es par, entonces r_1 es par, $n_{\mathbb{S}} = n_1$ y*

$$\alpha_{\mathbb{S}} = \alpha = w_B \left\| X \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^T \right\|_2 .$$

d) *Si $\lambda = 0$ y n_1 es par, entonces $n_{\mathbb{S}} = n_1$ y $\alpha_{\mathbb{S}} = \alpha = w_A\|XX^T\|_2$.*

e) *Si $\lambda = 0$ y n_1 es impar, entonces r_1 es par, $n_{\mathbb{S}} = n_1$ y*

$$\alpha_{\mathbb{S}} = \alpha = w_A \left\| X \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^T \right\|_2 .$$

f) *Si $\lambda \neq \infty$, $\lambda \neq 0$ y $r_1 = 1$, entonces $n_{\mathbb{S}} = n_1$ y $\alpha_{\mathbb{S}} = w_A\alpha_1 + w_B|\lambda|\alpha_2$, donde $\alpha_1 = \|X\|_2\|Y\|_2$ y $\alpha_2 = \sqrt{\|X\|_2^2\|Y\|_2^2 - |Y^TX|^2}$.*

g) *Si $\lambda \neq \infty$, $\lambda \neq 0$ y $r_1 > 1$, entonces $n_{\mathbb{S}} = n_1$ y $\alpha_{\mathbb{S}} \geq w_A\|XY^H\|_2$.*

(v). Haces de matrices antisimétricas:*Si* $\mathbb{S}_1 = \mathbb{S}_2 = \{A \in \mathbb{C}^{n\times n} : A^T = -A\}$ *entonces* $n_{\mathbb{S}_1\times\mathbb{S}_2} = n_1$, r_1 *es par y* $\alpha_{\mathbb{S}_1\times\mathbb{S}_2} = \alpha$ *para cualesquiera* $A, B \in \mathbb{S}_1$ *en norma* $||\cdot||_2$.

(vi). Haces de matrices hermíticas *: Sea* $\mathbb{S}_j = \{A \in \mathbb{C}^{n\times n} : A^H = \gamma_j A\}$ *para* $j \in \{1, 2\}$ *con* $\gamma_1, \gamma_2 \in \{1, -1\}$ *fijos. Entonces* $n_{\mathbb{S}_1\times\mathbb{S}_2} = n_1$ *y se cumplen las desigualdades*

$$\alpha/\sqrt{2} \leq \alpha_{\mathbb{S}_1\times\mathbb{S}_2} \leq \alpha$$

para cualesquiera matrices $A, B \in \mathbb{C}^{n\times n}$ *en norma dos. Si, además,* $\gamma_1 = \gamma_2$ *y* $\lambda \in \mathbb{R}$ *entonces* $\alpha_{\mathbb{S}_1\times\mathbb{S}_2} = \alpha$.

Demostración:

Mientras no se indique lo contrario, suponemos que el autovalor λ es finito (para el caso $\lambda = \infty$ las demostraciones son análogas).

(i) En vista de la demostración del Lema 2.2.2, sabemos que existe una matriz real $\widetilde{E}$ con $\|\widetilde{E}\| \leq 1$ tal que $\rho(Y^H\widetilde{E}X) \geq \|XY^H\|_2/2$. Escogemos $E = w_A\widetilde{E}, F = 0$ si $w_A \geq w_B|\lambda|$ y $E = 0, F = w_B\widetilde{E}$ en otro caso. Entonces

$$\rho(Y^H(E - \lambda F)X) \geq \frac{w_A + w_B|\lambda|}{2}\rho(Y^H\widetilde{E}X) \geq \frac{w_A + w_B|\lambda|}{4}\|XY^H\|_2,$$

lo que demuestra (i).

(ii) y *(iii)* Usando el hecho de que la forma canónica estructurada de un haz simétrico [89, Theorem 1] implica que $Y = \overline{X}$, los apartados (ii) y (iii) pueden demostrarse de igual manera que el Teorema 2.2.6.

(iv) Para $\lambda = \infty$, la forma canónica estructurada de un haz simétrico/antisimétrico [89, Theorem 1] impone la misma estructura en las matrices X e Y que para los autovalores nulos de una matriz antisimétrica. Esto implica que α/w_B coincide con el número de condición no estructurado para autovalores nulos de B y los apartados (a)–(c) se siguen del Teorema 2.2.7 (i)–(iii).

Para $\lambda = 0$ y n_1 par, la forma canónica del haz (A, B) [71, Theorem 8.3] conduce a que $Y = \overline{X}$, luego tomando $E = w_A\overline{X}X^H/\|XX^T\|_2$, $F = 0$ se demuestra (d). Si $\lambda = 0$ y n_1 es impar, entonces r_1 es par e $Y = \overline{X}\begin{bmatrix} 0 & I_{r_1/2} \\ -I_{r_1/2} & 0 \end{bmatrix}$. Sean u_1, v_1 vectores singulares izquierdo y derecho respectivamente, correspondientes al mayor valor singular σ_1 de la matriz antisimétrica $XY^H = X\begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix}X^T$. Entonces, el haz $E - \lambda F$ con $E = w_A[\overline{u_1}, v_1]\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}[\overline{u_1}, v_1]^T$ y $F = 0$ son tales que $E \in \mathbb{S}_1, F \in \mathbb{S}_2$, $\|E\|_2 = w_A$ y

$$\alpha_{\mathbb{S}} \geq \rho\left(EX\begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix}X^T\right) = \rho(V^HEU\Sigma) = \rho\left(\begin{bmatrix} w_A\sigma_1 & 0 \\ 0 & * \end{bmatrix}\right) \geq w_A\sigma_1 = \alpha,$$

donde $U\Sigma V^H$ es una descomposición en valores singulares de la matriz

$X\begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix}X^T$, $*$ denota una matriz no nula de tamaño $(r_1 - 1)\times(r_1 - 1)$ y hemos usado que $v_1^Tu_1 = 0$ ya que $X\begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix}X^T$ es una matriz antisimétrica. Esto prueba (e).

Para autovalores no nulos finitos λ y $r_1 = 1$, se tiene que

$$\rho(Y^H(E - \lambda F)X) = |Y^H(E - \lambda F)X|,$$

luego

$$\alpha_{\mathbb{S}} = \sup_{\substack{||E||\leq w_A, ||F||\leq w_B \\ (E,F)\in\mathbb{S}}} \rho(Y^H(E - \lambda F)X) \leq$$

$$\leq w_A \sup_{\substack{||E||\leq 1 \\ E\in\mathbb{S}_1}} \rho(Y^H EX) + w_B|\lambda| \sup_{\substack{||F||\leq 1 \\ F\in\mathbb{S}_2}} \rho(Y^H FX).$$

El supremo en $\mathbb{S}_1$ está claramente acotado por α_1, mientras que el supremo en $\mathbb{S}_2$ es igual a α_2 por el Teorema 2.2.7 *(iv)*. Por tanto, $\alpha_{\mathbb{S}} \leq w_A\alpha_1 + w_B|\lambda|\alpha_2$. Por [69, Teorema 5.8] tenemos que existe una matriz $\widetilde{E} \in \mathbb{S}_1$ con $\|\widetilde{E}\|_2 = \|Y\|_2/\|X\|_2$ tal que $\widetilde{E}X = Y$. Por tanto, la matriz simétrica $E_1 = \frac{\|X\|_2}{\|Y\|_2}\widetilde{E}$, cuya norma espectral es igual a uno, alcanza la cota superior α_1. Sea $F_2 \in \mathbb{S}_2$, con norma espectral igual a uno que alcanza el valor máximo α_2. Entonces existen $\gamma_1, \gamma_2 \in \mathbb{C}$, $|\gamma_1| = |\gamma_2| = 1$ de manera que el par $(E, F) = (\gamma_1 w_A E_1, \gamma_2 w_B F_2) \in \mathbb{S}_1 \times \mathbb{S}_2$ cumple $\rho(Y^H(E - \lambda F)X) = w_A\alpha_1 + w_B|\lambda|\alpha_2$. Esto prueba (f).

Para autovalores no nulos finitos λ y $r_1 > 1$ tenemos que, como se vio en la demostración del Teorema 2.2.10, existe una matriz simétrica E_1 con $\|E_1\|_2 = 1$ y $\rho(Y^H E_1 X) \geq \|XY^H\|_2$. Así, tomando $E = w_A E_1$ y $F = 0$ probamos (g).

(v) Para haces antisimétricos/antisimétricos, todo autovalor tiene un número par r_1 de bloques de Jordan. Además, la forma canónica estructurada [89] (véase también [71, Theorem 8.2]), induce la relación $Y = \overline{X}\begin{bmatrix} 0 & I_{r_1/2} \\ -I_{r_1/2} & 0 \end{bmatrix}$ entre X e Y. Si elegimos la matriz $\widetilde{E} = \overline{X}\begin{bmatrix} 0 & I_{r_1/2} \\ -I_{r_1/2} & 0 \end{bmatrix} X^H$, y las perturbaciones $E = \frac{w_A}{\|\widetilde{E}\|_2}\widetilde{E}$, $F = -\frac{w_B}{\|\widetilde{E}\|_2}\frac{\overline{\lambda}}{|\lambda|}\widetilde{E}$, se tiene que $\|E\|_2 = w_A$, $\|F\|_2 = w_B$ y

$$\rho(Y^H(E - \lambda F)X) = (w_A + |\lambda|w_B)\left\|X^T\begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix}X\right\|_2 = \alpha.$$

(vi) Como en la demostración del Lema 2.2.8 , podemos construir una matriz hermítica $\widetilde{E}$ tal que $\|\widetilde{E}\|_2 = 1$ y $\rho(\widetilde{E}XY^H) = \|XY^H\|_2$. Elijamos $\delta \in \{1, -1\}$ de modo que δ es igual al signo de λ_R si $\gamma_1 = \gamma_2$, y al signo de $-\lambda_I$ en caso contrario. Entonces $E = w_A\sqrt{\gamma_1}\widetilde{E} \in \mathbb{S}_1$ y $F = -\delta w_B\sqrt{\gamma_2}\widetilde{E} \in \mathbb{S}_2$ cumplen que

$$\alpha_{\mathbb{S}_1\times\mathbb{S}_2} \geq \rho((E - \lambda F)XY^H) = |w_A\sqrt{\gamma_1} + \delta w_B\sqrt{\gamma_2}\lambda|\,\|XY^H\|_2 \geq$$

$$\geq \frac{w_A + w_B|\lambda|}{\sqrt{2}}\|XY^H\|_2.$$

La última desigualdad se sigue del hecho de que

$$\begin{aligned} 2|w_A\sqrt{\gamma_1} + \delta w_B\sqrt{\gamma_2}\lambda|^2 - (w_A + w_B|\lambda|)^2 &\geq w_A^2 - 2w_A w_B|\lambda| + w_B^2|\lambda|^2 \\ &= (w_A - w_B|\lambda|)^2 \geq 0. \end{aligned}$$

Si $\gamma_1 = \gamma_2 = 1$ y $\lambda \in \mathbb{R}$ entonces $|w_A + \delta w_B \lambda| = w_A + w_B|\lambda|$, y el factor $1/\sqrt{2}$ puede eliminarse.

□

Una estructura diferente a las anteriores, en el sentido que no es separable, y que aparece en muchas aplicaciones, es la *estructura palindrómica*. Un *haz palindrómico* es un haz de la forma (A, A^T) (para más detalles véase [46, 66]). Para esta clase estructurada de haces se tiene el siguiente resultado:

Teorema 2.3.6 *Sea $\mathbb{S} = \{(A, A^T) : A \in \mathbb{C}^{n\times n}\}$ el conjunto de haces palíndrómicos de matrices y tomemos $w_A = w_B = 1$. Entonces las siguientes afirmaciones sobre $\kappa(A, A^T, \lambda; \mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$ son ciertas para cualquier $A \in \mathbb{C}^{n\times n}$ y para la norma espectral.*

(I). Si $\lambda = 1$, n_1 es impar y $r_1 = 1$, entonces $n_{\mathbb{S}} < n_1$;

(II). Si $\lambda = 1$, n_1 is impar y $r_1 > 1$, entonces $n_{\mathbb{S}} = n_1$ y $\alpha_{\mathbb{S}} = 2\sqrt{\sigma_1\sigma_2}$, donde σ_1, σ_2 son los dos mayores valores singulares de la matriz XX^T (mientras que $\alpha = 2\sigma_1$);

(III). Si $\lambda = 1$ y n_1 es par, entonces r_1 es par, $n_{\mathbb{S}} = n_1$ y

$$\alpha_{\mathbb{S}} = \alpha = 2\left\| X \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^T \right\|_2;$$

(IV). Si $\lambda = -1$ y n_1 es impar entonces r_1 es par, $n_{\mathbb{S}} = n_1$ y

$$\alpha_{\mathbb{S}} = \alpha = 2\left\| X \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^T \right\|_2;$$

(V). Si $\lambda = -1$ y n_1 es par, entonces $n_{\mathbb{S}} = n_1$ y $\alpha_{\mathbb{S}} = \alpha = 2\|XX^T\|_2$;

(VI). Si $\lambda \neq \pm 1$ es finito y $r_1 = 1$, entonces $n_{\mathbb{S}} = n_1$ y $\alpha_{\mathbb{S}} = |1-\lambda|\,\alpha_1 + |1+\lambda|\,\alpha_2$ donde $\alpha_1 = \|X\|_2\|Y\|_2$ y $\alpha_2 = \sqrt{\|X\|_2^2\|Y\|_2^2 - |Y^TX|^2}$.

(VII). Si $\lambda \neq \pm 1$ es finito y $r_1 > 1$, entonces $n_{\mathbb{S}} = n_1$ y $\frac{|1-\lambda|}{1+|\lambda|}\alpha \leq \alpha_{\mathbb{S}} \leq \alpha$;

(VIII). Si $\lambda = \infty$, entonces $n_{\mathbb{S}} = n_1$ y $\alpha_{\mathbb{S}} = \alpha$.

Demostración:
Si λ es un autovalor finito y $n_{\mathbb{S}} = n_1$ entonces

$$\alpha_{\mathbb{S}} = \sup_{\substack{\|E\|_2\leq 1 \\ E\in\mathbb{C}^{n\times n}}} \rho(Y^H(E - \lambda E^T)X) =$$

$$= \frac{1}{2} \sup_{\substack{\|E\|_2\leq 1 \\ E\in\mathbb{C}^{n\times n}}} \rho\left((1-\lambda)Y^H(E+E^T)X + (1+\lambda)Y^H(E-E^T)X\right). \tag{2.30}$$

Esta relación indica que el análisis de los haces palindrómicos está relacionado con el análisis de haces de matrices simétrica/antisimétrica. De hecho, en [49, 80, 83] se demuestra que la forma canónica de un haz palindrómico (A, A^T) puede extraerse la forma canónica [89] del haz de matrices simétrico/antisimétrico $(A + A^T, A - A^T)$.

Si el autovalor es $\lambda = 1$ y n_1 es impar entonces la forma canónica estructurada de (A, A^T) implica $Y = \overline{X}$. Si $\lambda = 1$ y n_1 es par, entonces r_1 es par e $Y = \overline{X} \begin{bmatrix} 0 & I_{r_1/2} \\ -I_{r_1/2} & 0 \end{bmatrix}$. Por otro lado, se sigue de (2.30) que

$$\alpha_{\mathbb{S}} = \sup_{\substack{\|E\|_2 \leq 1 \\ E \in \mathbb{C}^{n \times n}}} \rho(Y^H(E - \lambda E^T)X) = \sup_{\substack{\|G\|_2 \leq 2 \\ G \text{antisimétrica}}} \rho(Y^H G X) =$$

$$= 2 \sup_{\substack{\|G\|_2 \leq 1 \\ G \text{antisimétrica}}} \rho(Y^H G X).$$

Por tanto, el número de condición de Hölder estructurado para el autovalor $\lambda = 1$ de (A, A^T) coincide esencialmente con el número de condición de Hölder estructurado para el autovalor $\lambda = 0$ de la matriz antisimétrica $A - A^T$. En particular, los apartados (I), (II), y (III) del Teorema 2.2.7 implican los apartados (I), (II) y (III) de este teorema.

Si $\lambda = -1$ y n_1 es impar, entonces la forma canónica estructurada de (A, A^T) implica que r_1 es par e $Y = \overline{X} \begin{bmatrix} 0 & I_{r_1/2} \\ -I_{r_1/2} & 0 \end{bmatrix}$. Si $\lambda = -1$ y n_1 es par, entonces $Y = \overline{X}$. De nuevo se sigue de (2.30) que el supremos de $\rho(Y^H(E - \lambda E^T)X)$ es alcanzado por una matriz simétrica E si $\lambda = -1$. Por tanto, la situación en (iv) y (v) es completamente paralela a los apartados (e) y (d) respectivamente del Teorema 2.3.5 (iv). Luego una elección análoga de la matriz simétrica E demuestra (iv) y (v).

Para un autovalor finito λ con $r_1 = 1$ (es decir, X e Y son vectores), la relación (2.30) implica

$$\alpha_{\mathbb{S}} \leq |1 - \lambda| \sup_{\substack{\|E_1\|_2 \leq 1 \\ E_1 \text{ es simétrica}}} |Y^H E_1 X| + |1 + \lambda| \sup_{\substack{\|E_2\|_2 \leq 1 \\ E_2 \text{ es antisimétrica}}} |Y^H E_2 X|.$$

Como se demuestra en el Teorema 2.3.5 (iv) (f), el primer supremo, que se toma sobre matrices simétricas, es igual a α_1. Según el Teorema 2.2.7 (IV), el supremo en E_2 es igual a α_2. Esto demuestra que $\alpha_{\mathbb{S}} \leq |1 - \lambda|\, \alpha_1 + |1 + \lambda|\, \alpha_2$. Ahora, sea E_1 una matriz simétrica con $\|E_1\|_2 \leq 1$ y $|Y^H E_1 X| = \alpha_1$, y sea E_2 una matriz antisimétrica con $\|E_2\|_2 \leq 1$ y $|Y^H E_2 X| = \alpha_2$. Entonces, la matriz $E = \gamma_1 E_1 + \gamma_2 E_2$ con γ_1, γ_2 convenientemente elegidas cumpliendo $|\gamma_1| = |\gamma_2| = 1$, satisface $\|E\|_2 \leq 2$ y además $|Y^H E X| = |1 - \lambda|\, \alpha_1 + |1 + \lambda|\, \alpha_2$ lo que concluye la demostración de (VI).

El apartado (VII) se sigue directamente del hecho (véase la demostración del Teorema 2.2.10) de que existe una matriz simétrica E_1 tal que $\|E_1\|_2 = 1$ y $\rho(Y^H E X) \geq \|X Y^H\|_2$.

Finalmente, como la definición de α no cambia para autovalores infinitos, se verifica (VIII).

□

Resumiendo los resultados del Teorema 2.3.6 podemos concluir que los números de condición de Hölder estructurados y no estructurados de un haz de matrices palindrómico pueden diferir significativamente sólo para autovalores cercanos a 1.

2.3.2. Polinomios matriciales

Otras variantes en principio más generales del problema generalizado de autovalores, como son los polinomios matriciales, pueden reformularse de tal forma que se puedan tratar con las ideas de la sección 2.3.1. En esta sección veremos como obtener números de condición de Hölder para autovalores de un polinomio matricial

$$P(\lambda) = \lambda^m A_m + \lambda^{m-1} A_{m-1} + \ldots + \lambda A_1 + A_0, \quad A_i \in \mathbb{C}^{n\times n}. \tag{2.31}$$

Decimos que λ es autovalor de $P(\lambda)$ si $\det P(\lambda) = 0$, y se dice que los vectores $x, y \in \mathbb{C}^{n\times n}$ son, respectivamente, autovector derecho e izquierdo asociados al autovalor λ si $P(\lambda)x = 0$ e $y^H P(\lambda) = 0$, respectivamente.

En lo que sigue supondremos que el polinomio matricial P es *regular*, esto es, que el polinomio $\det P(\lambda)$ en la variable λ no es idénticamente nulo. El haz de matrices

$$A - \lambda B = \begin{pmatrix} -A_{m-1} & -A_{m-2} & \cdots & -A_1 & -A_0 \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{pmatrix} - \lambda \begin{pmatrix} A_m & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & I \end{pmatrix} \tag{2.32}$$

de tamaño $mn \times mn$ se llama *forma compañera* del polinomio matricial (2.31), y representa una de las *linealizaciones* más usuales: una linealización de (2.31) es un haz de matrices cuyos autovalores coinciden con los de $P(\lambda)$[1]. Además, hay también una correspondencia, que veremos a continuación, entre los autovectores de $P(\lambda)$ y los autovectores de su linealización [38, §1.1 & 7.2] (véanse también [67, 66]).

Consideramos el polinonio matricial perturbado $P + \epsilon\triangle P$ con

$$\triangle P(\lambda) = \lambda^m E_m + \lambda^{m-1} E_{m-1} + \cdots + \lambda E_1 + E_0, \quad E_i \in \mathbb{C}^{n\times n}.$$

Para investigar el comportamiento de los autovalores perturbados correspondientes, observamos que coinciden con los autovalores del haz de matrices perturbado

$$(A + \epsilon E, B + \epsilon F),$$

donde

$$E = -V \begin{pmatrix} E_{m-1} & E_{m-2} & \cdots & E_1 & E_0 \end{pmatrix}, \quad F = V E_m V^T, \tag{2.33}$$

[1]De forma más rigurosa, un haz $L(\lambda)$ de $mn \times mn$ es una linealización del polinomio $P(\lambda)$ en (2.31) si existen matrices unimodulares (esto es, con determinante constante, independiente de λ) $E(\lambda)$ y $F(\lambda)$ tales que

$$E(\lambda)L(\lambda)F(\lambda) = \begin{pmatrix} P(\lambda) & 0 \\ 0 & I_{(k-1)n} \end{pmatrix}.$$

y $V = [I_n, 0, \ldots, 0]^T$.

Esta relación nos permite extender fácilmente las Definiciones 2.3.2 y 2.3.4 de número de condición de Hölder, tanto estructurado como no estructurado, del marco de los haces de matrices al de los polinomios matriciales. Dedicaremos lo que queda de esta sección al condicionamiento no estructurado. El estructurado se define de manera análoga.

La siguiente definición es ligeramente más general que la Definición 2.3.2 de la que proviene, en el sentido de que permite flexibilidad en los coeficientes al permitir la elección de $m+1$ pesos no negativos $w_0, \ldots, w_m$. Para evitar situaciones degeneradas impondremos $w_0 > 0$ para $\lambda = 0$, $w_m > 0$ para $\lambda = \infty$, y $\max\{w_0, \ldots, w_m\} > 0$ para cualquier otro autovalor.

Definición 2.3.7 *Sea λ un autovalor finito del polinomio matricial P con forma compañera (A, B) como en (2.32), sean X e Y las correspondientes matrices de autovectores (2.27) de (A, B) y sean w_i, $i = 0, \ldots, m$ constantes no negativas. Consideramos las matrices de perturbación (E, F) de la forma (2.33), que preservan la forma* compañera. *Entonces el* **número de condición (absoluto) de Hölder** *para el autovalor λ viene dado por $\kappa(P, \lambda) = (n_1, \alpha)$, donde n_1 es el tamaño del mayor bloque de Jordan en la forma de Weierstrass del haz (A, B) asociado al autovalor λ y*

$$\alpha = \sup_{\substack{\|E_i\| \le w_i \\ E_i \in \mathbb{C}^{n\times n}}} \rho(Y^H(E - \lambda F)X).$$

El número de condición (absoluto) de Hölder *para $\lambda = \infty$ viene dado por $\kappa(P, \infty) = (n_1, \alpha)$, donde n_1 es el índice de nilpotencia del haz (A, B) y*

$$\alpha = \sup_{\substack{\|E_i\| \le w_i \\ E_i \in \mathbb{C}^{n\times n}}} \rho(Y^H F X).$$

Nótese que no perdemos generalidad por usar la linealización en forma compañera en la Definición 2.3.7, ya que la relación entre los autovalores de un polinomio matricial y los autovalores de cualquier linealización es biyectiva. Por tanto, cualquier otra *linealización fuerte*[2] en el sentido de [37, 60, 67] conducirá al mismo número de condición de Hölder. Una de las ventajas de haber elegido la forma *compañera* es la estructura especialmente sencilla de los autovectores del haz (A, B). Más concretamente, los resultados en [45, Lemma 7.2] y [44, Lema 3.7] demuestran que x_1 e y_1 son autovectores derecho e izquierdo asociados al autovalor λ de P si y sólo si

$$x = \begin{pmatrix} \lambda^{m-1}x_1 \\ \vdots \\ \lambda x_1 \\ x_1 \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ (\lambda A_m + A_{m-1})^H y_1 \\ \vdots \\ (\lambda^{m-1}A_m + \lambda^{m-2}A_{m-1} + \cdots + A_1)^H y_1 \end{pmatrix} \tag{2.34}$$

[2]El concepto de *linealización fuerte* fue introducido por Gohberg, Kaashoek y Lancaster en [37], aunque fueron Lancaster y Psarrakos quien le dieron ese nombre en [60]: un linealización $L(\lambda) = A - \lambda B$ de un polinomio regular $P(\lambda)$ es fuerte si, además, el haz dual $B - \lambda A$ de $L(\lambda)$ es una linealización del polinomio dual de $P(\lambda)$.

son autovectores derecho e izquierdo del haz (A, B), respectivamente. Para el autovalor $\lambda = \infty$, los autovectores de (A, B) vienen dados por $x = [x_1^H, 0, \ldots, 0]^H$ e $y = [y_1^H, 0, \ldots, 0]^H$. Esto demuestra que las matrices X e Y definidas en (2.27), que contienen autovectores derechos e izquierdos asociados a un autovalor finito (múltiple) λ del haz (A, B), son de la forma

$$X = \begin{pmatrix} \lambda^{m-1} X_1 \\ \vdots \\ \lambda X_1 \\ X_1 \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ (\lambda A_m + A_{m-1})^H Y_1 \\ \vdots \\ (\lambda^{m-1} A_m + \lambda^{m-2} A_{m-1} + \cdots + A_1)^H Y_1 \end{pmatrix}, \tag{2.35}$$

donde X_1 e Y_1 son matrices de autovectores derechos e izquierdos del polinomio matricial P. Para un autovalor infinito, sólo los primeros bloques de las matrices X e Y son no nulos e iguales a X_1 e Y_1, respectivamente.

El siguiente lema nos proporciona una fórmula explícita del número de condición de Hölder y además demuestra que $\alpha > 0$ (bajo las condiciones anteriores sobre los pesos), algo que – estrictamente hablando – es necesario para justificar la Definición 2.3.7.

Lema 2.3.8 *Sea $P(\lambda)$ un polinomio matricial regular de la forma (2.31) y sea λ un autovalor finito de P. Entonces*

$$\kappa(P, \lambda) = (n_1, (w_m|\lambda|^m + w_{m-1}|\lambda|^{m-1} + \cdots + w_0)\|X_1 Y_1^H\|_2)$$

para cualquier norma unitariamente invariante $\|\cdot\|$, donde X_1 e Y_1 son las matrices de autovectores de P relativas a las matrices de autovectores X e Y del haz (A, B) como se muestra en (2.35). Para un autovalor infinito, $\kappa(P, \lambda) = (n_1, w_m\|X_1 Y_1^H\|_2)$.

Demostración:
La estructura de las matrices E, F, X e Y dadas en (2.33) y (2.35) implica que

$$Y^H(E - \lambda F)X = -Y_1^H(\lambda^m E_m + \lambda^{m-1} E_{m-1} + \cdots + E_0)X_1$$

Como en la demostración del Lema 2.3.3, se prueba que

$$\rho(Y^H(E - \lambda F)X) \leq (w_m|\lambda|^m + w_{m-1}|\lambda|^{m-1} + \cdots + w_0)\|X_1 Y_1^H\|_2 \tag{2.36}$$

Sean u, v los vectores singulares izquierdo y derecho, respectivamente, asociados al mayor valor singular de la matriz $X_1 Y_1^H$. Entonces, la igualdad en (2.36) se alcanza con la elección de las siguientes perturbaciones

$$E_0 = w_0 v u^H, \; E_1 = w_1 \frac{\bar{\lambda}}{|\lambda|} v u^H, \; \ldots, \; E_m = w_m \frac{\bar{\lambda}^m}{|\lambda|^m} v u^H,$$

con $E_1 = \cdots = E_m = 0$ para $\lambda = 0$. Esto demuestra el resultado para λ finito. Para autovalores infinitos la demostración es análoga, teniendo en cuenta que $Y^H F X = E_m$. □

Queremos hacer notar que las matrices X_1 e Y_1 no pueden elegirse arbitrariamente en el Lema 2.3.8; el resultado depende de la normalización de las matrices X e Y impuesta por (2.27). Para ilustrar el efecto de tal normalización, sea, por ejemplo, λ un autovalor finito *semisimple* de P y supongamos que $\widetilde{X}_1$ y $\widetilde{Y}_1$ contienen bases *arbitrarias* de autovectores derechos e izquierdos asociados a λ. Si denotamos por $\widetilde{X}$ e $\widetilde{Y}$ las correspondientes bases de autovectores de $A - \lambda B$ entonces (2.35) implica

$$\widetilde{Y}^H B \widetilde{X} = \widetilde{Y}_1^H P'(\lambda) \widetilde{X}_1.$$

Como λ es semisimple y finito, la matriz $\widetilde{Y}_1^H P'(\lambda)\widetilde{X}_1$ es invertible y

$$X_1 = \widetilde{X}_1 (\widetilde{Y}_1^H P'(\lambda) \widetilde{X}_1)^{-1}, \quad Y_1 = \widetilde{Y}_1$$

cumple que $Y_1^H P'(\lambda) X_1 = I$. Esto implica la condición impuesta por (2.27) para un autovalor semisimple. Por el Lema 2.3.8,

$$\kappa(P, \lambda) = \left(1, (w_m|\lambda|^m + w_{m-1}|\lambda|^{m-1} + \cdots + w_0) \left\| \widetilde{X}_1 (\widetilde{Y}_1^H P'(\lambda) \widetilde{X}_1)^{-1} \widetilde{Y}_1^H \right\|_2 \right).$$

Para $r_1 = 1$, esta fórmula coincide con el resultado de Tisseur [90, Theorem 5] sobre números de condición para *autovalores simples* de un polinomio matricial.

2.4. Perturbaciones estructuradas completamente no genéricas: el caso matricial

Dada una matriz cuadrada $A \in \mathbb{C}^{n\times n}$, un autovalor λ de A y una perturbación aditiva de la forma $A + \epsilon E$ para una cierta $E \in \mathbb{C}^{n\times n}$, ya hemos visto en la sección 2.1 que los autovalores perturbados $\widehat{\lambda}_k$ pueden escribirse como

$$\widehat{\lambda}_k = \lambda + \alpha_k^{\gamma_k} \epsilon^{\gamma_k} + o(\epsilon^{\gamma_k}), \quad \epsilon \to 0, \quad k = 1, \ldots, m. \tag{2.37}$$

Además, el exponente γ_k coincide genéricamente con $1/n_1$, siendo n_1 el tamaño del mayor bloque de Jordan asociado al autovalor λ.

Sin embargo, como se observó en el Ejemplo 2.9, hay casos en los que γ_k es estrictamente mayor que $1/n_1$. Esto puede ocurrir para una perturbación E en concreto, o para todas las perturbaciones E que pertenezcan a una cierta clase de matrices estructuradas $\mathbb{S}$. Este último caso es el que más nos interesa, y a él dedicaremos la presente sección. Si, como antes, denotamos por $\kappa(A, \lambda; \mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$ y $\kappa(A, \lambda) = (n_1, \alpha)$ los números de condición estructurado y no estructurado, nuestro primer objetivo en esta sección es hallar condiciones necesarias y suficientes sobre la clase $\mathbb{S}$ bajo las cuales $n_{\mathbb{S}} < n_1$. Si la estructura cumple esta propiedad, diremos que es *completamente no genérica* para el autovalor λ y la matriz A en cuestión. Para estas estructuras, por tanto, el comportamiento de λ frente a perturbaciones en $\mathbb{S}$ es cualitativamente mejor que frente a perturbaciones arbitrarias. Una vez identificadas tales estructuras, nuestro objetivo será identificar $n_{\mathbb{S}}$ y, si es posible,

obtener fórmula explícitas o, al menos, estimaciones para $\alpha_{\mathbb{S}}$. Para ello emplearemos en la sección 2.4.3 el diagrama de Newton.

Nótese en primer lugar que, puesto que el número de condición refleja el mayor efecto posible sobre los autovalores, dada una clase $\mathbb{S}$ de matrices estructuradas, basta que haya una sola matriz $E \in \mathbb{S}$ tal que el exponente γ_k en (2.37) sea igual a $1/n_1$ para concluir que $n_{\mathbb{S}} = n_1$. Por tanto, si X e Y son las matrices de autovectores definidas en (2.13) y (2.14), una estructura $\mathbb{S}$ es completamente no genérica si y sólo si para toda matriz $E \in \mathbb{S}$ se cumple la ecuación

$$Y^H E X = 0_{r_1 \times r_1}. \tag{2.38}$$

Veamos un par de ejemplos de estructuras completamente no genéricas.

Ejemplo 2.4.1 *Sea* $\mathbb{S} = \{A \in \mathbb{C}^{3\times 3} : A = -A^T\}$ *el conjunto de las matrices antisimétricas complejas de tamaño* 3 *y sea la matriz antisimétrica*

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

Consideramos su único autovalor $\lambda = 0$*, que posee un único bloque de Jordan de tamaño* 3*. Si calculamos el número de condición de Hölder no estructurado usando la Definición 2.3.2, se obtiene* $\kappa(A, 0) = (3, 2)$*. Puede comprobarse que los autovalores perturbados* $\widehat{\lambda}_k(\epsilon)$ *de* $A + \epsilon E$ *para cualquier matriz antisimétrica*

$$E = \begin{pmatrix} 0 & x & y \\ -x & 0 & z \\ -y & -z & 0 \end{pmatrix},$$

cumplen el desarrollo asintótico

$$\widehat{\lambda}_k(\epsilon) = \lambda + (2iz + 2x)^{1/2}\epsilon^{1/2} + o(\epsilon^{\frac{1}{2}}), \quad \textit{para} \quad k = 1, 2.$$

Por tanto, $n_{\mathbb{S}} = 2 < n_1 = 3$*, y el número de condición estructurado es*

$$\kappa(A, 0; \mathbb{S}) = (2, \max_{||E||_F=1} |2iz + 2x|) = (2, 2).$$

Ejemplo 2.4.2 *Consideramos la estructura* $\mathbb{S} = \{A \in \mathbb{C}^{3\times 3} : A\Sigma_3 = -\Sigma_3 A^T\}$ *para la matriz ortogonal simétrica* $\Sigma_3 = \begin{pmatrix} & & 1 \\ & -1 & \\ 1 & & \end{pmatrix}$*. Sea*

$$A = \begin{pmatrix} \sqrt{2} & i & 0 \\ i & 0 & i \\ 0 & i & -\sqrt{2} \end{pmatrix} \in \mathbb{S},$$

que tiene un solo autovalor $\lambda = 0$ con un único bloque de Jordan de tamaño 3. Si calculamos el número de condición de Hölder no estructurado usando la Definición 2.3.2, se tiene que $\kappa(A, 0) = (3, 4)$. Puede comprobarse que todas las matrices de $\mathbb{S}$ tienen la forma

$$E = \begin{pmatrix} x & y & 0 \\ z & 0 & y \\ 0 & z & -x \end{pmatrix}$$

y que los autovalores perturbados $\widehat{\lambda}_k(\epsilon)$ de $A + \epsilon E$ para cualquier matriz de esa forma cumplen el desarrollo

$$\widehat{\lambda}_k(\epsilon) = \lambda + (iz + 2iy + 2\sqrt{2}x)^{1/2}\epsilon^{1/2} + o(\epsilon^{1/2}), \quad \textit{para} \quad k = 1, 2.$$

Por tanto $n_{\mathbb{S}} = 2 < n_1 = 3$ y el número de condición estructurado es

$$\kappa(A, 0; \mathbb{S}) = (2, \max_{||E||_F=1} |iz + 2iy + 2\sqrt{2}x|) = \left(2, \frac{\sqrt{2}}{2}\right).$$

2.4.1. Estructuras lineales completamente no genéricas

Comenzamos por recordar nuestra definición de *estructuras completamente no genéricas*

Definición 2.4.3 *Sea λ un autovalor de $A \in \mathbb{C}^{n\times n}$ con forma de Jordan (2.10), y sean X e Y definidas como en (2.13) y (2.14), respectivamente. Sea $\mathbb{S} \subset \mathbb{C}^{n\times n}$. Se dice que la clase de matrices $\mathbb{S}$ es* **completamente no genérica** *para A y λ si para toda matriz $E \in \mathbb{S}$ se tiene que $Y^H EX = 0$.*

Nuestro objetivo en esta sección es ver qué relación debe haber entre las matrices de autovectores X e Y y la estructura $\mathbb{S}$ para asegurar que $Y^H EX = 0_{r_1\times r_1}$ para cualquier $E \in \mathbb{S}$ *siempre que $\mathbb{S}$ sea una estructura lineal.*

Si escribimos $X = \begin{pmatrix} x_1 & x_2 & \dots & x_{r_1} \end{pmatrix}$ e $Y = \begin{pmatrix} y_1 & y_2 & \dots & y_{r_1} \end{pmatrix}$, entonces las matrices de rango 1

$$M_{ij} = y_i x_j^H, \qquad i, j = 1, \dots, r_1 \tag{2.39}$$

van a jugar un papel importante en la discusión. Si llamamos *vec* al operador que concatena las columnas de una matriz $m \times n$ para producir un vector de $\mathbb{C}^{mn}$, una de las propiedades de este operador [48, §4.3] es que

$$\text{vec}(ABC) = (C^T \otimes A)\text{vec}(B) \tag{2.40}$$

para cualesquiera matrices A, B, C tales que su producto está bien definido. Haremos uso de esta propiedad en el siguiente lema, que da condiciones necesarias y suficientes para que se satisfaga (2.38).

Lema 2.4.4 *Sea $A \in \mathbb{C}^{n\times n}$ con forma canónica de Jordan* (2.10)*, sean X e Y las matrices definidas como en* (2.13) *y* (2.14) *y sean $M_{ij},\ i,j = 1,\ldots,r_1$, dadas por* (2.39)*. Entonces $Y^H EX = 0$ para $E \in \mathbb{C}^{n\times n}$ si y sólo si*

$$vec(M_{ij})^H vec(E) = 0, \qquad para\ todo \quad i,j\ \in \{1\ldots r_1\}. \tag{2.41}$$

Demostración:
El producto $Y^H EX$ valdrá cero si y sólo si cada una de las entradas $y_i^H E x_j$ son nulas. Usando la propiedad (2.40), esto es equivalente a que $(x_j^T \otimes y_i^H)vec(E) = 0$ para cada par de índices $i,j \in \{1\ldots r_1\}$, o lo que es lo mismo, a que sea $[vec(\overline{y_i}x_j^T)]^T vec(E) = 0$, lo que implica el resultado (2.41).

□

El Lema 2.4.4 demuestra que, dada una estructura de matrices $\mathbb{S}$, la condición (2.38) se satisface si y sólo si todas y cada una de las r_1^2 matrices M_{ij} en (2.39) son ortogonales a cualquier matriz de la estructura $\mathbb{S}$ con respecto al producto escalar matricial

$$\langle A, B\rangle = \mathrm{vec}(A)^H \mathrm{vec}(B). \tag{2.42}$$

Cuando el conjunto $\mathbb{S}$ de matrices estructuradas es un subespacio lineal, podemos hablar de la *antiestructura* correspondiente a $\mathbb{S}$.

Definición 2.4.5 *Sea $\mathbb{S}$ una estructura lineal de $\mathbb{C}^{n\times n}$, esto es, un subespacio vectorial de $\mathbb{C}^{n\times n}$. Se define la* **antiestructura** *asociada a $\mathbb{S}$ como*

$$\mathbb{S}^\perp = \{B \in \mathbb{C}^{n\times n} :\ vec(B)^H vec(A) = 0 \quad \forall A \in \mathbb{S}\}.$$

$\mathbb{S}^\perp$ es, obviamente, el complemento ortogonal del subespacio $\mathbb{S}$ en $\mathbb{C}^{n\times n}$ con respecto al producto escalar (2.42). Así, toda estructura lineal $\mathbb{S}$ tiene una única antiestructura asociada $\mathbb{S}^\perp$, que es también subespacio lineal, y si $\mathbb{S}$ tiene dimensión k entonces $\mathbb{S}^\perp$ tiene dimensión $n^2 - k$. Además, $(\mathbb{S}^\perp)^\perp = \mathbb{S}$. En la sección §2.4.2 veremos cuáles son las antiestructuras correspondientes a algunas de las estructuras lineales más comunes.

Con esta definición, la propiedad (2.38) es equivalente a decir que las r_1^2 matrices de rango uno M_{ij} en (2.39) pertenecen a la antiestructura $\mathbb{S}^\perp$. Este es el principal resultado de esta sección y es consecuencia directa del Lema 2.4.4 y la Definición 2.4.5:

Teorema 2.4.6 *Sea λ un autovalor de la matriz $A \in \mathbb{C}^{n\times n}$, sea $\mathbb{S} \subset \mathbb{C}^{n\times n}$ una estructura lineal, y sean $\kappa(A,\lambda) = (n_1, \alpha)$ y $\kappa(A,\lambda;\mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$ los números de condición no estructurado y estructurado, respectivamente, del autovalor λ. Sean las matrices $M_{ij},\ i,j = 1,\ldots,r_1$, definidas como en* (2.39)*. Entonces $n_{\mathbb{S}} < n_1$ si y sólo si $M_{ij} \in \mathbb{S}^\perp$ para todo $i,j = 1,\ldots,r_1$.*

En la sección anterior definimos la matriz de Lidskii (2.17)

$$W_\lambda = \frac{YX^H}{\|XY^H\|_2},$$

que, salvo por un factor constante, es igual a la suma $\sum_{i=1}^{r_1} M_{ii}$ de las matrices M_{ij} con $i = j$. Obtenemos así una condición necesaria para (2.38):

Corolario 2.4.7 *Sea λ un autovalor de la matriz $A \in \mathbb{C}^{n\times n}$, sea $\mathbb{S}$ una estructura lineal y sean $\kappa(A,\lambda) = (n_1, \alpha)$ y $\kappa(A,\lambda;\mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$ los números de condición no estructurado y estructurado, respectivamente, del autovalor λ. Si $n_{\mathbb{S}} < n_1$ entonces la perturbación W_λ definida en* (2.17) *pertenece a la clase de matrices $\mathbb{S}^\perp$.*

Ilustremos este resultado con los Ejemplos 2.4.1 y 2.4.2. Para el Ejemplo 2.4.1 su matriz de Lidskii es

$$W_\lambda = \begin{pmatrix} 1 & 0 & i \\ 0 & 0 & 0 \\ i & 0 & -1 \end{pmatrix},$$

que es simétrica. Como se verá en 2.4.2 la antiestructura asociada a las matrices antisimétricas es justamente el conjunto de matrices simétricas. Análogamente, en el Ejemplo 2.4.2 se puede ver que la matriz de Lidskii

$$W_\lambda = \begin{pmatrix} 1 & -\sqrt{2}\,i & -1 \\ -\sqrt{2}\,i & -2 & \sqrt{2}\,i \\ -1 & \sqrt{2}\,i & 1 \end{pmatrix},$$

pertenece a la antiestructura de las matrices $\mathbb{S} = \{A \in \mathbb{C}^{3\times 3} : A\Sigma_3 = -\Sigma_3 A^T\}$.

2.4.2. Estructuras y antiestructuras lineales

En esta subsección describiremos las antiestructuras asociadas a las estructuras lineales más comunes. Para ello debemos recordar algunas definiciones básicas.

Definición 2.4.8 *Se dice que $A \in \mathbb{C}^{n\times n}$ es* **simétrica** *(resp.,* **antisimétrica***) si $A^T = A$ (resp., si $A^T = -A$). Se dice que $A \in \mathbb{C}^{n\times n}$ es* **persimétrica** *(resp.,* **antipersimétrica***) si $A^T F_n = F_n A$ (resp., $A^T F_n = -F_n A$), donde F_n viene dada por (2.18).*

Definición 2.4.9 *Sea $n \in \mathbb{N}$, sean $p, q \in \mathbb{N}$ con $p + q = n$ y sea*

$$\Sigma_{p,q} = \begin{pmatrix} I_p & 0 \\ 0 & -I_q \end{pmatrix}.$$

Se dice que $A \in \mathbb{C}^{n\times n}$ es una matriz **pseudosimétrica** *(resp.,* **antipseudosimétrica***) con respecto a $\Sigma_{p,q}$ si $\Sigma_{p,q} A^T = A\Sigma_{p,q}$ (resp., $\Sigma_{p,q} A^T = -A\Sigma_{p,q}$).*

Definición 2.4.10 *Sea*

$$J_n = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}.$$

Se dice que $H \in \mathbb{R}^{2n\times 2n}$ *es* **hamiltoniana** *(resp.,* **antihamiltoniana***) si* HJ_n *es simétrica (resp., antisimétrica), es decir, si* $HJ_n = (HJ_n)^T$ *(resp., si* $HJ_n = -(HJ_n)^T$*). Toda matriz hamiltoniana* real $H \in \mathbb{R}^{n\times n}$ *se puede escribir de la forma*

$$H = \begin{pmatrix} A & B \\ C & -A^T \end{pmatrix}$$

con B *y* C *simétricas, mientras que toda matriz antihamiltoniana real* $\mathcal{H} \in \mathbb{R}^{n\times n}$ *se puede escribir de la forma*

$$\mathcal{H} = \begin{pmatrix} A & B \\ C & A^T \end{pmatrix}$$

con B *y* C *antisimétricas*

Definición 2.4.11 *Se dice que* $A \in \mathbb{C}^{n\times n}$ *es una matriz* **circulante** *(resp.,* **cocirculante***) si es de Toeplitz (resp., de Hankel) y cada una de sus filas (resp., de sus columnas) es una traslación cíclica de la fila (resp., de la columna) anterior.*

Por ejemplo, para $n = 4$

$$\begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ a_4 & a_1 & a_2 & a_3 \\ a_3 & a_4 & a_1 & a_2 \\ a_2 & a_3 & a_4 & a_1 \end{pmatrix}$$

es una matriz circulante y

$$\begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ a_2 & a_3 & a_4 & a_1 \\ a_3 & a_4 & a_1 & a_2 \\ a_4 & a_1 & a_2 & a_3 \end{pmatrix}$$

es una matriz cocirculante.

Definición 2.4.12 *Sea una matriz* $A = (a_{ij}) \in \mathbb{C}^{n\times n}$*. Se dice que*

i) A *tiene* **sumas diagonales nulas** *(o abreviando,* **sumas-d nulas***) si, para cada* $k \in \{-n+1, -n+2, \ldots, 0, \ldots, n-2, n-1\}$*,*

$$\sum_{i-j=k} a_{ij} = 0.$$

ii) A *tiene* **sumas antidiagonales nulas** *(o abreviando,* **sumas-ad nulas***) si, para cada* $k \in \{2, \ldots, 2n\}$*,*

$$\sum_{i+j=k} a_{ij} = 0.$$

iii) *A tiene* **sumas diagonales extendidas nulas** *(o abreviando,* **sumas-de nulas***) si, para cada* $k \in \{0, \ldots, n-2, n-1\}$,

$$\sum_{i-j=k} a_{ij} + \sum_{j-i=n-k} a_{ij} = 0.$$

iv) *A tiene* **sumas antidiagonales extendidas nulas** *(o abreviando,* **sumas-ade nulas***) si, para cada* $k \in \{2, \ldots, n, n+1\}$,

$$\sum_{i+j=k} a_{ij} + \sum_{j+i=n-k} a_{ij} = 0.$$

Sean las siguientes matrices:

$$A_1 = \begin{pmatrix} 1 & 2 & 0 \\ 4 & 2 & -2 \\ 0 & -4 & -3 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 0 & 2 & 1 \\ -2 & 2 & 4 \\ -3 & -4 & 0 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} 1 & 4 & 9 \\ -3 & 2 & 2 \\ -6 & -6 & -3 \end{pmatrix} \quad \text{y} \quad A_3 = \begin{pmatrix} 1 & 4 & 9 \\ -3 & 7 & 5 \\ -16 & -6 & -1 \end{pmatrix}.$$

Entonces A_1 tiene sumas-d nulas, A_2 tiene sumas-ad nulas, A_3 tiene sumas-de nulas y A_4 tiene sumas-ade nulas.

Después de estas definiciones, la siguiente tabla describe las antiestructuras correspondientes a las estructuras lineales más usuales:

$\mathbb{S}$	$\mathbb{S}^{\perp}$
simétrica	antisimétrica
pseudosimétrica	antipseudosimétrica
persimétrica	antipersimétrica
hamiltoniana real	antihamiltoniana real
Toeplitz	sumas d-nulas
Hankel	sumas ad-nulas
circulante	sumas de-nulas
cocirculante	sumas dea-nulas

La idea de la demostración es la misma para todas las estructuras: se sabe por [42] que para cada estructura lineal $\mathbb{S}$ de dimensión t existe una matriz $F \in \mathbb{R}^{n^2 \times t}$ tal que toda $A \in \mathbb{S}$ satisface

$$\operatorname{vec}(A) = Fp$$

para algún vector $p \in \mathbb{C}^t$ apropiado. A la vista de ello, una matriz $B \in \mathbb{C}^{n \times n}$ estará en $\mathbb{S}^{\perp}$ si y sólo si

$$\operatorname{vec}(B)^H F = 0_{1 \times t}.$$

Por tanto, las condiciones sobre las entradas de la matriz B se extraen de esta ecuación. Veremos sólo dos ejemplos, las matrices simétricas 3×3 y las cocirculantes 4×4, a modo de ilustración. El resto de los casos son totalmente análogos, y pasar a dimensión arbitraria no reviste complicación alguna.

1. **Simétrica,** $n = 3$: Sea $A \in \mathbb{C}^{3\times 3}$ una matriz simétrica. Como la dimensión del conjunto de matrices simétricas 3×3 es $t = 6$, sabemos que existe una matriz $F \in \mathbb{R}^{9\times 6}$ y un vector $p \in \mathbb{C}^6$ tal que $vec(A) = Fp$. Dicha matriz F es la que se muestra en la siguiente ecuación.

$$vec(A) = F\triangle p = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{22} \\ a_{23} \\ a_{33} \end{pmatrix}.$$

   Sea ahora $B = (b_{ij})_{i,j=1}^{3} \in \mathbb{C}^{3\times 3}$ una matriz cualquiera. Entonces, $vec(B)^H F = 0_{1\times 6}$ si y sólo si

$$\begin{pmatrix} \overline{b_{11}} & \overline{b_{21}} & \overline{b_{31}} & \overline{b_{12}} & \overline{b_{22}} & \overline{b_{32}} & \overline{b_{13}} & \overline{b_{23}} & \overline{b_{33}} \end{pmatrix} F = 0_{1\times t}$$

   Basta hacer el producto matricial e igualar los elementos de ambas matrices para obtener que $B \in \mathbb{S}^{\perp}$ si y sólo si $\Leftrightarrow b_{ii} = 0$ para i=1,2,3 y $b_{ij} = -b_{ji}$ para $i, j \in \{1, \ldots, 3\}$. En otras palabras, la antiestructura de las matrices simétricas es la clase de matrices antisimétricas.

2. **Circulante,** $n = 4$: Sea $A = \begin{pmatrix} a & d & c & b \\ b & a & d & c \\ c & b & a & d \\ d & c & b & a \end{pmatrix}$ una matriz circulante. Como la dimensión del conjunto de matrices circulantes 4×4 es $t = 4$, sabemos que existe una matriz $F \in \mathbb{R}^{16\times 4}$ y un vector $p \in \mathbb{C}^4$ tales que

$$vec(A) = Fp = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}$$

para cualquier matriz circulante A

Sea ahora $B = (b_{ij})_{i,j=1}^{3} \in \mathbb{C}^{3\times 3}$ una matriz cualquiera. Entonces, puede comprobarse que $vec(B)^H F = 0_{1\times 6}$ si y sólo si

$$\begin{cases} b_1 + b_6 + b_{11} + b_{16} = 0 \\ b_2 + b_7 + b_{12} + b_{13} = 0 \\ b_3 + b_8 + b_9 + b_4 = 0 \\ b_4 + b_5 + b_{10} + b_{15} = 0 \end{cases},$$

es decir, la antiestructura asociada a las matrices circulantes son las matrices que tienen sumas-de nulas.

2.4.3. Números de condición estructurados vía el diagrama de Newton

En esta subsección identificaremos el exponente principal $n_{\mathbb{S}}$ en el número de condición estructurado $\kappa(A, \lambda; \mathbb{S})$ para una estructura $\mathbb{S}$ completamente no genérica, esto es, para una estructura que satisfaga la condición (2.38). Nuestra herramienta fundamental será el *diagrama de Newton* [3, Appendix A7] (a veces llamado también *polígono de Puiseux-Newton*), una herramienta ideada y desarrollada por Newton, aunque rigurosamente justificada mucho más tarde por Puiseux [79]. Describiremos el diagrama de Newton adaptado al contexto del problema que nos ocupa: sean las matrices $A, E \in \mathbb{C}^{n\times n}$, sea λ_0 un autovalor de A, sean λ, ϵ dos parámetros y sea J_A la forma canónica de Jordan de la matriz $A = P_A J_A P_A^{-1}$, como en (2.10). Sin pérdida de generalidad, supondremos en toda esta sección que

$$\lambda_0 = 0$$

(de no ser así, basta reemplazar el parámetro λ por $\lambda - \lambda_0$ en lo que sigue). Con esta simplificación, podemos escribir la ecuación característica de la matriz perturbada $A + \epsilon E$ como

$$p(\lambda, \epsilon) = \det(A + \epsilon E - \lambda I) = \det(J_A + \epsilon P^{-1} E P - \lambda I) = \\ = \lambda^n + \alpha_1(\epsilon)\lambda^{n-1} + \ldots + \alpha_{n-1}(\epsilon)\lambda + \alpha_n(\epsilon)$$

donde

$$\alpha_k(\epsilon) = \widehat{\alpha}_k \epsilon^{a_k} + \ldots, \;\; k = 1, \ldots, n,$$

siendo a_k el exponente principal y $\widetilde{\alpha}_k$ el coeficiente director de $\alpha_k(\epsilon)$, es decir, $\widetilde{\alpha}_k \neq 0$ y no hay términos de orden menor que a_k en el desarrollo de $\alpha_k(\epsilon)$. Se sabe [3, 56, 72] que en esta situación las raíces de $p(\lambda, \epsilon)$ vienen dadas por desarrollos

$$\lambda(\epsilon) = \mu \epsilon^{p/q} + \ldots$$

en potencias fraccionarias de ϵ. Los exponentes principales p/q de estos desarrollos pueden determinarse por medio de la siguiente construcción geométrica: dibujamos los puntos (k, a_k) para $k = 1, \ldots, m$ y el punto $(0, 0)$ correspondiente a la potencia λ^n. Entonces dibujamos los segmentos que forman el borde inferior de la envolvente convexa de esos puntos. Estos segmentos constituyen lo que se denomina *Diagrama de Newton* asociado a $p(\lambda, \epsilon)$. Se puede demostrar que:

1. Las pendientes de los distintos segmentos son los exponentes principales de los desarrollos en ϵ de las raíces $\lambda(\epsilon)$ de $p(\lambda, \epsilon)$.

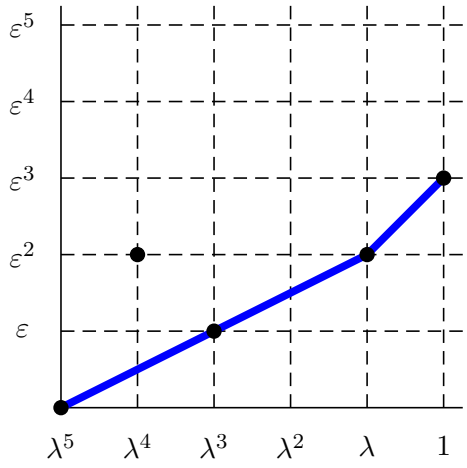
2. El número de raíces que corresponde a cada pendiente coincide con la longitud de la proyección sobre el eje horizontal del segmento con dicha pendiente.

3. Los coeficientes directores μ asociados a un determinado exponente, y por tanto al segmento S que tiene ese exponente por pendiente, son las soluciones de la ecuación

$$\sum_{(k, a_k) \in S} \mu^{m-k} \widehat{\alpha}_k = 0.$$

Ilustramos estas propiedades con el siguiente ejemplo.

Ejemplo 2.4.13 *Sea* $P(\lambda, \epsilon) = \lambda^5 - \epsilon^2 \lambda^4 + (\epsilon - 3\epsilon^2)\lambda^3 + \epsilon^2 \lambda - \epsilon^3$. *El diagrama de Newton asociado a* $P(\lambda, \epsilon)$ *es*



El diagrama de Newton está compuesto por dos segmentos, uno de pendiente $1/2$ *y otro de pendiente* 1. *Esto significa que hay cuatro autovalores perturbados cuyo exponente principal es* $1/2$, *y cuyo coeficiente director viene dado por las soluciones de la ecuación*

$$2\mu^3 - \mu = 0.$$

El quinto autovalor tiene exponente principal 1, *y el coeficiente director viene dado por la solución de*

$$-\mu - 1 = 0.$$

Por tanto, las raíces de $p(\lambda, \epsilon)$ *son de la forma*

$$\lambda_k(\epsilon) = \pm\sqrt{\frac{1}{2}}\,\epsilon^{\frac{1}{2}} + \ldots \quad k = 1, 2, 3, 4$$

y

$$\lambda_5(\epsilon) = -\epsilon + \ldots .$$

Emplearemos esta teoría para obtener los términos dominantes de los desarrollos en serie de potencias fraccionarias en ϵ de los autovalores de la matriz perturbada $A + \epsilon E$ para $E \in \mathbb{S}$ cuando $\mathbb{S}$ es una estructura completamente no genérica.

Para ello, sea $A \in \mathbb{C}^{n\times n}$ con forma de Jordan (2.10). Para cada $i = 1, \ldots, q$, las columnas de P contienen r_i cadenas de Jordan linealmente independientes de longitud n_i. Consideramos el *segundo vector* de cada cadena de Jordan derecha de longitud n_1 asociada a λ, y reunimos todos estos segundos vectores en la matriz

$$X_2 = \left(\; Pe_2, \; Pe_{n_1+2}, \; Pe_{2n_1+2}, \ldots, \; Pe_{(r_1-1)n_1+2} \; \right)$$

de $n \times r_1$. Análogamente, se reúnen en la matriz

$$Y_2 = \left(\; Q^H e_{n_1-1}, \; Q^H e_{2n_1-1}, \; \ldots, \; Q^H e_{r_1 n_1-1} \; \right)$$

de $n \times r_1$ todos los segundos vectores de cadenas de Jordan izquierdas de longitud n_1. Si X e Y son las matrices de autovectores definidas como en (2.13) y (2.14), y suponemos que hay r_2 bloques de Jordan asociados a λ del segundo mayor tamaño n_2, consideramos la matriz

$$W_2 = \left[Y \mid Q^H e_{r_1 n_1 + n_2},\ Q^H e_{r_1 n_1 + 2n_2},\ \ldots,\ Q^H e_{r_1 n_1 + r_2 n_2} \right],$$

de tamaño $n \times (r_1 + r_2)$, que contiene todos los autovectores izquierdos asociados a las cadenas de Jordan de longitud n_1 y n_2, y la matriz

$$Z_2 = \left[X \mid Pe_{r_1 n_1 + 1},\ Pe_{r_1 n_1 + n_2 + 1},\ Pe_{r_1 n_1 + 2n_2 + 1}, \ldots,\ Pe_{(r_1 n_1 + (r_2 - 1)n_2 + 1} \right]$$

de $n \times (r_1 + r_2)$, que hace lo mismo para autovectores derechos. Finalmente, se define

$$\Phi_1 = Y^H E X \in \mathbb{C}^{r_1 \times r_1}, \quad \Phi_2 = W_2^H E Z_2 \in \mathbb{C}^{(r_1 + r_2) \times (r_1 + r_2)} \tag{2.43}$$

y las matrices de tamaño $r_1 \times r_1$

$$\Phi_{12} = Y^H E X_2, \quad \Phi_{21} = Y_2^H E X. \tag{2.44}$$

Queremos hallar el exponente principal $1/n_{\mathbb{S}}$ de $\kappa(A, \lambda; \mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$ para estructuras $\mathbb{S}$ que cumplan (2.38). Para ello debemos identificar la menor pendiente posible del diagrama de Newton asociado al polinomio característico de $A + \epsilon E$ cuando $E \in \mathbb{S}$. La restricción $Y^H E X = 0$ sobre $\mathbb{S}$ equivale en términos del diagrama de Newton a que el punto $P_1 = (n_1 r_1, r_1)$ no aparece en el diagrama (véanse las Figuras 2.1 y 2.2), puesto que el coeficiente de $p(\lambda, \epsilon)$ en $\lambda^{n - r_1 n_1} \epsilon^{r_1}$ es $\det \Phi_1 = \det(Y^H E X) = 0$. De hecho, como el rango de la matriz $\Phi_1 = Y^H E X$ es cero, en el diagrama de Newton no puede aparecer ningún segmento de pendiente $1/n_1$, luego los puntos

$$P_2^j = (n_1 r_1 - j n_1, r_1 - j), \qquad j \in \{1, \ldots, r_1 - 1\}$$

en las Figuras 2.1 y 2.2 no serán candidatos para el análisis en nuestra discusión.

En cualquier caso, *el segmento del diagrama de Newton con menor pendiente provendrá de un segmento de pendiente mayir que* $1/n_1$ *que conecte el origen con alguno de los puntos de la cuadrícula más cercanos a* P_1. Distinguiremos dos situaciones diferentes:

1. La matriz A tiene forma de Jordan (2.10) con $q = 1$, es decir, sólo posee bloques de Jordan de tamaño n_1, y los puntos más cercanos a P_1 (véase Figura 2.2) son

$$P_3 = (n_1 r_1 - 1, r_1) \tag{2.45}$$

$$P_4 = (n_1 r_1 - 2, r_1) \tag{2.46}$$

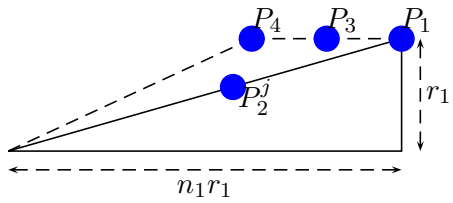
2. La matriz A tiene forma de Jordan (2.10) con $q > 1$, es decir, hay bloques de Jordan de tamaños menores que n_1 y los puntos más cercanos (véase Figura 2.1) son

$$P_3 = (n_1 r_1 - 1, r_1) \tag{2.47}$$

$$P_4^j = (n_1 r_1 + n_2 j, r_1 + j) \quad \text{para algún} \quad j \in \{1, \ldots, r_2\} \tag{2.48}$$

$$P_5 = (n_1 r_1, r_1 + 1) \tag{2.49}$$

Figura 2.1:



Estudiemos primero el segundo caso, es decir, el caso en que $q > 1$: denotamos por m_3, m_4^j y m_5, respectivamente, a las pendientes de los segmentos que unen el origen $(0,0)$ con cada uno de los puntos P_3, P_4^j y P_5 respectivamente. Obviamente, $m_4^1 \leq m_5$, independientemente de los valores de n_1, n_2 y r_1. El siguiente resultado compara las restantes pendientes.

Lema 2.4.14 *Sea A una matriz con forma de Jordan* (2.10)*, sean los puntos P_3, P_4^j, $j = 1, \ldots r_2$ y P_5, definidos como en (2.47), (2.48) y (2.49), respectivamente, y sean m_3, m_4^j y m_5 las pendientes respectivas del segmento que une el origen con cada uno de esos puntos. Entonces*

(i) $m_3 \leq m_5$ *si y sólo si* $n_1 \geq 2$.

(ii) $m_3 \leq m_4^j$ *para algún* $j \in \{1, \ldots, r_2\}$ *si y sólo si* $n_1 - n_2 \geq \frac{r_1+j}{r_1 j}$.

(iii) $m_4^j \leq m_5$ *, para algún* $j \in \{1, \ldots, r_2\}$ *si y sólo si* $\dfrac{n_1}{n_2} \leq \dfrac{(r_1+1)j}{r_1(j-1)}$.

Demostración:

El apartado (i) es trivial. Para ver el apartado (ii), basta comparar las pendientes correspondientes, esto es, m_3 es más pequeña que m_4^j si

$$\frac{r_1}{n_1 r_1 - 1} \leq \frac{r_1 + j}{n_1 r_1 + j n_2}.$$

Un simple cálculo demuestra que esto es equivalente a

$$n_1 - n_2 \geq \frac{r_1 + j}{r_1 j}.$$

Finalmente, la pendiente m_4^j es más pequeña que m_5 cuando

$$\frac{r_1+j}{n_1r_1+jn_2} \leq \frac{r_1+1}{n_1r_1}$$

lo que ocurre sólo si

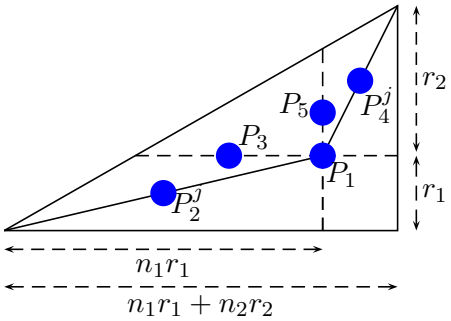
$$\frac{n_1}{n_2} \leq \frac{r_1+1}{r_1}\frac{j}{j-1},$$

concluyendo la demostración.

□

Ahora queremos determinar los coeficientes del polinomio característico $p(\lambda,\epsilon)$ asociados a los puntos P_3, P_4^j y P_5. Si denotamos a una submatriz de tamaño $k\times k$ de una matriz $\mathcal{M}$ formada por las filas $i_1,\ldots,i_k$ y las columnas $j_1,\ldots,j_k$ por $\mathcal{M}([i_1,\ldots,i_k],[j_1,\ldots,j_k])$, y a la submatriz principal de filas y columnas $i_1,\ldots,i_k$ por $\mathcal{M}(i_1,\ldots,i_k)$, tenemos el siguiente resultado.

Figura 2.2:



Teorema 2.4.15 *Sea la matriz $A\in\mathbb{C}^{n\times n}$ con forma canónica de Jordan (2.10), sea $E\in\mathbb{C}^{n\times n}$ y sea $p(\lambda,\epsilon)$ el polinomio característico de la matriz $A+\epsilon E$. Sean los puntos P_3, P_4^j, $j=1,\ldots,r_2$ y P_5 definidos como en (2.47), (2.48) y (2.49) respectivamente. Entonces*

1. *El coeficiente de $\lambda^{n-n_1r_1+1},\epsilon^{r_1}$ en $p(\lambda,\epsilon)$, correspondiente al punto P_3 en la Figura 2.1, es:*

   a) *si $n_1-n_2\geq 2$, el coeficiente es*

$$\begin{aligned} C_3=(-1)^{(n_1-1)r_1-1}[&\sum_{j=1}^{r_1}\det(\Phi([1:r_1]\,,\,[1:j-1\,,j+r_1\,,j+1:r_1])) \\ &+\sum_{j=1}^{r_1}\det(\Phi([1:j-1\,,j+r_1\,,j+1:r_1],[1:r_1]))] \end{aligned} \tag{2.50}$$

donde $\Phi = \begin{pmatrix} \Phi_1 & \Phi_{12} \\ \Phi_{21} & 0 \end{pmatrix}$, *y las matrices* Φ_1, Φ_{12} *y* Φ_{21} *están dadas por* (2.43) *y* (2.44).

b) *si* $n_2 = n_1 - 1$,

$$\widetilde{C}_3 = C_3 + (-1)^{(n_1-1)r_1-1} \sum_{j=1}^{r_1} \sum_{k=1}^{r_2} \det(\Phi_2([1:j-1, j+1:r_1,\ r_1+k])),$$

donde C_3 *viene dado por* (2.50) *y la matriz* Φ_2 *está definida en* (2.43).

2. *Para cada* $j \in \{1, \ldots, r_2\}$, *el coeficiente de* $\lambda^{n-n_1r_1-jn_2}\epsilon^{r_1+j}$ *en* $p(\lambda, \epsilon)$, *correspondiente al punto* P_4^j *en la Figura 2.1, es*

$$C_4^j = (-1)^{(n_1-1)r_1+(n_2-1)j} \sum_{\{k_1,\ldots,k_j\}\subset\{1,\ldots,r_2\}} \det(\Phi_2([1 \ldots r_1\ ,\ r_1+k_1,\ \ldots\ , r_1+k_j]))$$

Demostración:

Sea $J_A = P_A^{-1}AP_A$ la forma de Jordan de A. El coeficiente del término en $\lambda^j\epsilon^k$ del polinomio característico de $A+\epsilon E$ es igual a la suma de los coeficientes de los términos de orden ϵ^k que aparecen en los menores principales de tamaño $n-j$ de la matriz

$$\mathcal{G} = J_A - \epsilon P_A^{-1} E P_A. \tag{2.51}$$

La única manera de construir un menor principal de tamaño $n-j$ que contenga un producto de ese orden es elegir el menor de modo que el producto contenga exactamente $n-j-k$ términos que no dependan de ϵ, esto es, que el menor contenga $n-j-k$ de las posiciones supradiagonales de $\mathcal{G}$ que contienen un 1 en la forma de Jordan J_A. A estas posiciones las llamaremos *posiciones ϵ-constantes*[3] en el resto de la demostración.

En vista de esta observación, el coeficiente del término en $\lambda^{n-n_1r_1+1}\epsilon^{r_1}$, correspondiente al punto P_3, es la suma de los coeficientes de orden ϵ^{r_1} que aparecen en los menores principales de tamaño n_1r_1-1 de $\mathcal{G}$. Esto equivale a elegir menores principales de tamaño n_1r_1-1 que contengan exactamente $(n_1-1)r_1-1$ posiciones ϵ-constantes. Veamos en primer lugar que los menores con esta propiedad *tienen que estar formados por filas que provienen de* r_1 *bloques de Jordan distintos de la matriz* J_A: en efecto, cada posición ϵ-constante es de la forma $(j, j+1)$. Sea s el número de bloques de Jordan distintos de entre los que se eligen las filas para generar productos de orden ϵ_1^r, y sea $p_i,\ i \in \{1, \ldots, s\}$, el número de posiciones ϵ-constantes que aporta al producto cada uno de los s bloques. Las p_i posiciones ϵ-constantes de cada bloque determinan como mínimo p_i+1 filas del menor principal. Por tanto, el número de filas del menor principal que contienen posiciones ϵ-constantes es al menos $\sum_{i=1}^s p_i + s$. Como los menores son de tamaño n_1r_1-1 y además $\sum_{i=1}^s p_i = (n_1-1)r_1-1$, concluimos que tiene que ser $s = r_1$.

Distinguimos los siguientes casos dependiendo de cómo se elijan las filas en cada uno de los menores principales:

[3]Como estamos suponiendo que $\lambda_0 = 0$, ésas son las únicas posiciones que contienen términos que no dependen de ϵ en la matriz.

1. Si escogemos las $n_1r_1 - 1$ filas de entre los r_1 bloques de tamaño máximo n_1, éstos contienen como máximo $(n_1-1)r_1$ posiciones ϵ-constantes. Para obtener $(n_1-1)r_1-1$ posiciones ϵ-constantes eligiendo n_1r_1-1 filas debemos dejar fuera del menor principal una fila que sea o la primera o la última asociada a uno de los bloques de Jordan de tamaño r_1: dejar fuera del menor cualquier otra fila intermedia eliminaría dos posiciones ϵ-constantes en lugar de sólo una. Por tanto, el coeficiente depende exclusivamente de lo que llamaremos las *primeras y segundas esquinas* de la perturbación $P_A^{-1}EP_A$. Podemos ilustrar lo que son con el siguiente ejemplo: sea una matriz de 6 $\times$ 6, con un único autovalor $\lambda = 0$, cuya forma canónica de Jordan tiene dos bloques de Jordan de tamaño 3. Entonces, para toda perturbación E de 6×6, los coeficientes del término en $\lambda\epsilon^2$ dependen sólo de los elementos marcados con un símbolo en la matriz

$$P_A^{-1}EP_A = \left(\begin{array}{ccc|ccc} * & * & * & * & * & * \\ \clubsuit_1 & * & * & \clubsuit_2 & * & * \\ \heartsuit_1 & \spadesuit_1 & * & \heartsuit_2 & \spadesuit_2 & * \\ \hline * & * & * & * & * & * \\ \clubsuit_3 & * & * & \clubsuit_4 & * & * \\ \heartsuit_3 & \spadesuit_3 & * & \heartsuit_4 & \spadesuit_4 & * \end{array}\right).$$

Los elementos marcados con $\heartsuit$ son las esquinas primeras, y los marcados con $\clubsuit$ y $\spadesuit$ son las esquinas segundas.

De hecho, puede comprobarse fácilmente que, en general, el coeficiente del término $\lambda^{n-n_1r_1+1}\epsilon^{r_1}$ es la suma de los menores principales

$$\mathcal{G}([1 : jn_1 - 1, jn_1 + 1 : n_1r_1])$$

y

$$\mathcal{G}([1 : (j-1)n_1, (j-1)n_1 + 2 : n_1r_1]).$$

El primer caso corresponde a dejar fuera del menor principal la *última fila* del j-ésimo bloque de Jordan de tamaño r_1 (esto es, la jn_1-ésima fila de $\mathcal{G}$), mientras que la segunda corresponde a dejar fuera la fila $(jn_1 + 1)$-ésima de $\mathcal{G}$, es decir, la primera fila del j-ésimo bloque. Un simple cálculo permite escribir esta suma como suma de los menores que se indica en (2.50) de la matriz $\begin{pmatrix} \Phi_1 & \Phi_{12} \\ \Phi_{21} & 0 \end{pmatrix}$ donde Φ_{12}, Φ_{21} están definidos como en (2.43) y (2.44). Esto concluye la demostración del apartado 1.a). En el ejemplo anterior, la matriz Φ es

$$\Phi = \left(\begin{array}{cc|cc} \heartsuit_1 & \heartsuit_2 & \clubsuit_1 & \clubsuit_2 \\ \heartsuit_3 & \heartsuit_4 & \clubsuit_3 & \clubsuit_4 \\ \hline \spadesuit_1 & \spadesuit_2 & 0 & 0 \\ \spadesuit_3 & \spadesuit_4 & 0 & 0 \end{array}\right),$$

y el coeficiente correspondiente es

$$-\det\begin{pmatrix} \heartsuit_1 & \clubsuit_2 \\ \heartsuit_3 & \clubsuit_4 \end{pmatrix} - \det\begin{pmatrix} \clubsuit_1 & \heartsuit_2 \\ \clubsuit_3 & \heartsuit_4 \end{pmatrix} - \det\begin{pmatrix} \heartsuit_1 & \heartsuit_2 \\ \spadesuit_3 & \spadesuit_4 \end{pmatrix} - \det\begin{pmatrix} \spadesuit_1 & \spadesuit_2 \\ \heartsuit_3 & \heartsuit_4 \end{pmatrix}.$$

2. Si escogemos las $n_1r_1 - 1$ filas de entre $r_1 - k$ bloques de Jordan de tamaño n_1, y k bloques de tamaños menores, $n_{i_1}, \ldots, n_{i_k}$, tendremos a lo sumo $(n_1 - 1)(r_1 - k) + \sum_{j=1}^{k}(n_{i_j} - 1)$ posiciones ϵ-constantes disponibles. Para que este número sea mayor o igual que $(n_1 - 1)r_1 - 1$ tiene que ser $\sum_{j=1}^{k} n_{i_j} \geq n_1k - 1$ con $n_{i_j} < n_1$ y $k \geq 1$. La única posibilidad de que esto ocurra es que sea $k = 1$ y $n_{i_1} = n_1 - 1$, es decir, sólo se pueden conseguir términos del orden deseado si el segundo mayor tamaño n_2 de bloques de Jordan asociados a λ es justamente $n_2 = n_1 - 1$. En tal caso, las filas del menor principal provendrán de $r_1 - 1$ bloques de Jordan completos de tamaño n_1 y un solo bloque de Jordan completo de tamaño $n_1 - 1$.

   En consecuencia, la parte del coeficiente principal que se añade al coeficiente C_3 en (2.50) dependerá sólo de los elementos de Φ_2, esto es, con las *primeras esquinas* de la perturbación $P_A^{-1}EP_A$ asociadas con los bloques de Jordan de *ambos tamaños* n_1 y $n_1 - 1$. Si consideramos, por ejemplo, una matriz de 10×10 con $n_1 = 3$, $n_2 = 2$ y $r_1 = r_2 = 2$, el coeficiente del término en $\lambda^5\epsilon^2$ dependerá sólo de las entradas marcadas con un símbolo en

$$P_A^{-1}EP_A = \left(\begin{array}{ccc|ccc|cc|cc} * & * & * & * & * & * & * & * & * & * \\ * & * & * & * & * & * & * & * & * & * \\ \heartsuit_1 & * & * & \heartsuit_2 & * & * & \clubsuit_1 & * & \clubsuit_2 & * \\ \hline * & * & * & * & * & * & * & * & * & * \\ * & * & * & * & * & * & * & * & * & * \\ \heartsuit_3 & * & * & \heartsuit_4 & * & * & \clubsuit_3 & * & \clubsuit_4 & * \\ \hline * & * & * & * & * & * & * & * & * & * \\ \clubsuit_5 & * & * & \clubsuit_6 & * & * & \clubsuit_7 & * & \clubsuit_8 & * \\ \hline * & * & * & * & * & * & * & * & * & * \\ \clubsuit_9 & * & * & \clubsuit_{10} & * & * & \clubsuit_{11} & * & \clubsuit_{12} & * \end{array}\right) \tag{2.52}$$

   Más concretamente, un simple cálculo muestra que el coeficiente es la suma de los menores principales

$$G([1 : (j-1)n_1\ ,\ jn_1 + 1 : n_1r_1,\ n_1r_1 + (k-1)n_2 + 1 : n_1r_1 + kn_2])$$

   para $j \in \{1 \ldots r_1\}$, $k \in \{1 \ldots r_2\}$, esto es, incluimos en el menor las filas correspondientes a *todos* los bloques de Jordan de tamaño n_1 *salvo uno* (el j-ésimo), y las filas correspondientes a *un solo* bloque de tamaño $n_1 - 1$ (el k-ésimo). Equivalentemente, la parte del coeficiente principal que se añade a C_3 es la suma de los menores principales $[1 \ldots j-1, j+1 : r_1,\ r_1 + k]$ con $j \in \{1, \ldots, r_1\}$, $k \in \{1, \ldots, r_2\}$ de la matriz

Φ_2. Para el anterior ejemplo este coeficiente es

$$-\det\begin{pmatrix} \heartsuit_1 & \clubsuit_1 \\ \clubsuit_5 & \clubsuit_7 \end{pmatrix} - \det\begin{pmatrix} \heartsuit_1 & \clubsuit_2 \\ \clubsuit_9 & \clubsuit_{12} \end{pmatrix} - \det\begin{pmatrix} \heartsuit_4 & \clubsuit_3 \\ \clubsuit_6 & \clubsuit_7 \end{pmatrix} - \det\begin{pmatrix} \heartsuit_4 & \clubsuit_4 \\ \clubsuit_{10} & \clubsuit_{12} \end{pmatrix}$$

.

Esto concluye la demostración del apartado 1.b). Estudiamos ahora el punto P_4^j, que se corresponde con los términos en $\lambda^{n-n_1r_1-jn_2}\epsilon^{r_1+j}$, esto es, con la suma de los coeficientes de orden ϵ^{r_1+j} que aparecen en los menores principales de tamaño $n_1r_1+jn_2$ de la matriz $\mathcal{G}$. Por tanto, esos menores principales deben contener exactamente $(n_1-1)r_1+(n_2-1)j$ posiciones ϵ-constantes. Para ello es necesario que las filas de estos menores principales provengan de exactamente r_1+j bloques de Jordan de la matriz J_A, razonando de manera análoga al caso anterior. La única posibilidad es elegir r_1 bloques de Jordan completos de tamaño n_1 y j bloques completos del segundo mayor tamaño n_2. Esta combinación es la única que contiene exactamente el número de posiciones ϵ-constantes que se necesitan. Por tanto, el coeficiente será la suma de aquellos menores principales de tamaño r_1+j de la matriz Φ_2 que contengan a la matriz Φ_1. En el ejemplo (2.52), el coeficiente de P_4^1 es

$$-\det\begin{pmatrix} \heartsuit_1 & \heartsuit_2 & \clubsuit_1 \\ \heartsuit_3 & \heartsuit_4 & \clubsuit_3 \\ \clubsuit_5 & \clubsuit_6 & \clubsuit_7 \end{pmatrix} - \det\begin{pmatrix} \heartsuit_1 & \heartsuit_2 & \clubsuit_2 \\ \heartsuit_3 & \heartsuit_4 & \clubsuit_4 \\ \clubsuit_9 & \clubsuit_{10} & \clubsuit_{12} \end{pmatrix},$$

y el coeficiente de P_4^2 es

$$-\det\begin{pmatrix} \heartsuit_1 & \heartsuit_2 & \clubsuit_1 & \clubsuit_2 \\ \heartsuit_3 & \heartsuit_4 & \clubsuit_3 & \clubsuit_4 \\ \clubsuit_5 & \clubsuit_6 & \clubsuit_7 & \clubsuit_8 \\ \clubsuit_9 & \clubsuit_{10} & \clubsuit_{11} & \clubsuit_{12} \end{pmatrix}.$$

□

El Teorema 2.4.15 es válido para perturbaciones no estructuradas cualesquiera. Imponiendo ahora la condición (2.38), esto es, que la perturbación pertenezca a una estructura completamente no genérica, podemos extraer los coeficientes $\alpha_{\mathbb{S}}$ para una estructura dada, y así escribir una expresión simplificada para el número de condición de Hölder estructurado. Esto es lo que muestra el siguiente teorema.

Teorema 2.4.16 *Sea λ un autovalor de una matriz $A \in \mathbb{C}^{n\times n}$ con forma canónica de Jordan* (2.10)*, y sean X e Y matrices de autovectores de A definidas como en* (2.13) *y* (2.14)*. Sea $\mathbb{S}$ un conjunto de matrices tales que $Y^HEX = 0$ para toda perturbación $E \in \mathbb{S}$ y sean Φ_1, Φ_2 las matrices dadas por* (2.43) *y Φ_{12}, Φ_{21} definidas como en* (2.44)*.*

1. *Si $n_1 - n_2 \geq 2$ y $r_1 = 1$, entonces*

$$\kappa(A, \lambda; \mathbb{S}) = (n_1 - 1, \widetilde{\alpha})$$

donde

$$\widetilde{\alpha} = \sup_{E\in\mathbb{S}, ||E||\leq 1} \left|\Phi_{12} + \Phi_{21}\right|.$$

2. *Si* $n_1 = n_2 + 1$ *y* $r_1 = 1$, *entonces*

$$\kappa(A, \lambda; \mathbb{S}) = (n_1 - 1, \widetilde{\alpha})$$

donde

$$\widetilde{\alpha} = \sup_{E \in \mathbb{S}, ||E|| \leq 1} |\Phi_{12} + \Phi_{21} + \text{traza}(\Phi_2)|.$$

3. *Si* $n_1 = n_2 + 1$ *y* $r_1 = 2$, *entonces*

$$\kappa(A, \lambda; \mathbb{S}) = \left(\frac{n_1 r_1 - 1}{r_1}, \widetilde{\alpha}\right)$$

donde

$$\widetilde{\alpha} = \sup_{E \in \mathbb{S}, \|E\| \leq 1} \left| \sum_{k=1}^{r_2} \det \Phi_2([1\ ,\ 2+k]) + \sum_{k=1}^{r_2} \det \Phi_2([2\ ,\ 2+k]) \right|.$$

4. *Si* $n_1 - n_2 \geq 2$, $r_1 > 1$ *y* $r_2 > 1$, *o si* $n_1 = n_2 + 1$, $r_1 > 2$ *y* $r_2 > 1$, *entonces*

$$\kappa(A, \lambda; \mathbb{S}) = \left(\frac{n_1 r_1 + 2n_2}{r_1 + 2}, \widetilde{\alpha}\right)$$

donde

$$\widetilde{\alpha} = \sup_{E \in \mathbb{S}, ||E|| \leq 1} \left| \sum_{\substack{k_p \in \{1 : r_2\} \\ p \in \{1:2\}}} \det \Phi_2([1 \ldots r_1\ ,\ r_1 + k_1, r_1 + k_2]) \right|.$$

Demostración: Basta reeemplazar Φ_1 por la matriz nula en el Teorema 2.4.15 y tener en cuenta el Lema 2.4.14.

□

Para completar nuestro estudio debemos analizar el caso en que la forma canónica de Jordan de la matriz A tiene bloques de Jordan de un único tamaño. Este caso queda reflejado en la Figura 2.2, que reproducimos de nuevo para comodidad del lector.

El primer candidato para formar el segmento de menor pendiente del diagrama de Newton es el punto P_3, que corresponde al término en $\lambda\epsilon^{r_1}$ del polinomio característico. Por tanto, el coeficiente asociado será la suma de los menores principales de tamaño $n-1$ de $\mathcal{G}$ que contengan $n - r_1 - 1$ posiciones ϵ-constantes. Ahora bien, como $q = 1$, se tiene que $n = n_1 r_1$ por lo que habremos de buscar menores principales de tamaño $n_1 r_1 - 1$ que contengan $(n_1 - 1)r_1 - 1$ posiciones $\epsilon - constantes$. Eso sólo lo podemos obtener dejando fuera del menor una única fila, que tendrá que ser o la primera o la última de un bloque de Jordan. Ahora bien, si $r_1 > 1$ entonces al dejar fuera del menor una de esas filas, el menor resultante siempre contiene una fila de ceros, debido a la condición (2.38) de la estructura $\mathbb{S}$. Por tanto, *el punto* $P_3 = (n_1 r_1 - 1, r_1)$ *desaparece del diagrama de Newton*

si $r_1 > 1$. Por un argumento similar, $P_4 = (n_1 r_1 - 2, r_1)$ no puede aparecer si $r_1 > 2$, y así sucesivamente con los puntos $(n_1 r_1 - j, r_1),\ j = 3, \ldots$. Por tanto, la menor pendiente posible en el diagrama de Newton es siempre $(n_1 r_1 - r_1)/r_1 = n_1 - 1$.

El siguiente Teorema describe los dos primeros casos, para $r_1 = 1$ y $r_1 = 2$. A medida que r_1 crece, las expresiones para $\widetilde{\alpha}$ se complican más y más. Obviamos el caso $n_1 = 1$, puesto que es el caso trivial en que λ sigue siendo autovalor de $+\epsilon B$.

Teorema 2.4.17 *Sea* λ *un autovalor de la matriz* $A \in \mathbb{C}^{n\times n}$ *con forma de Jordan* (2.10) *tal que* $q = 1$*, es decir, A sólo tiene bloques de Jordan de tamaño* $n_1 \geq 2$*. Sea* $\mathbb{S}$ *un conjunto estructurado de matrices tales que* $Y^H E X = 0$ *para toda perturbación* $E \in \mathbb{S}$*, siendo* X *e* Y *las matrices de autovectores definidas en* (2.13)*,* (2.14)*, y* Φ_{12}*,* Φ_{21} *las matrices definidas en* (2.43) *y* (2.44)*. Entonces,*

$$\kappa(A, \lambda; \mathbb{S}) = (n_1 - 1, \widetilde{\alpha})\,,$$

1. *Si, además,* $r_1 = 1$*, entonces*

$$\widetilde{\alpha} = \sup_{E\in\mathbb{S}, ||E||\leq 1} \left|\Phi_{12} + \Phi_{21}\right|.$$

2. *Si, además,* $r_1 = 2$*, entonces*

$$\widetilde{\alpha} = \sup_{E\in\mathbb{S}, ||E||\leq 1} |\det\, \Phi_{12} + \det\, \Phi_{21} + [\Phi_{12}]_{11}[\Phi_{21}]_{22} + [\Phi_{12}]_{22}[\Phi_{21}]_{11}]\,|.$$

Demostración:

El primer apartado se sigue de la demostración del Teorema 2.4.16 del cálculo del coeficiente no nulo del punto P_3 para $r_1 = 1$. Cuando $r_1 > 1$ el coeficiente de P_3 en las condiciones de nuestro teorema es nulo ya que Φ_1 es la matriz nula. De este modo el siguiente punto más cercano que tenemos que estudiar es P_4 (véase la Figura 2.2).

El coeficiente del término en $\lambda^{n-n_1 r_1+2}\epsilon^{r_1}$, correspondiente al punto P_4, es la suma de los coeficientes de orden ϵ^{r_1} que aparecen en los menores principales de tamaño $n_1 r_1 - 1$ de (2.51). Esto equivale a elegir menores principales de tamaño $n_1 r_1 - 2$ que contengan exactamente $(n_1-1)r_1-2$ posiciones ϵ-constantes. El único modo de obtener esto es escoger menores principales de la matriz total (2.51) eliminando dos filas primeras, dos últimas o una primera y otra última asociadas a cualquiera de los bloques de Jordan de tamaño n_1; cualquier otra elección de filas eliminaría más posiciones ϵ-constantes de las necesarias. De hecho, los menores principales resultantes de eliminar o bien las dos primeras o bien las dos últimas filas de cualesquiera bloques de Jordan de tamaño r_1 serían nulos por ser la matriz Φ_1 nula, obteniéndose el último apartado.

En particular para el caso $r_1 = 2$ la suma de los menores anteriores pueden simplificarse como siguen. Sea

$$P_A^{-1}EP_A = \left(\begin{array}{ccccc|ccccc} * & * & * & \vdots & * & * & * & * & \vdots & * \\ \dots & \dots & \dots & \vdots & * & \dots & \dots & \dots & \vdots & * \\ * & * & * & \vdots & * & * & * & * & \vdots & * \\ \clubsuit_1 & * & * & \vdots & * & \clubsuit_2 & * & * & \vdots & * \\ \heartsuit_1 & \spadesuit_1 & * & \vdots & * & \heartsuit_2 & \spadesuit_2 & * & \vdots & * \\ \hline * & * & * & \vdots & * & * & * & * & \vdots & * \\ \dots & \dots & \dots & \vdots & * & \dots & \dots & \dots & \vdots & * \\ * & * & * & \vdots & * & * & * & * & \vdots & * \\ \clubsuit_3 & * & * & \vdots & * & \clubsuit_4 & * & * & \vdots & * \\ \heartsuit_3 & \spadesuit_3 & * & \vdots & * & \heartsuit_4 & \spadesuit_4 & * & \vdots & * \end{array}\right).$$

De este modo el coeficiente de P_4 es igual a

$$-\det\begin{pmatrix} \clubsuit_1 & \clubsuit_2 \\ \clubsuit_3 & \clubsuit_4 \end{pmatrix} - \det\begin{pmatrix} \spadesuit_1 & \spadesuit_2 \\ \spadesuit_3 & \spadesuit_4 \end{pmatrix} - \clubsuit_1\spadesuit_4 - \spadesuit_1\clubsuit_4,$$

obteniéndose el segundo apartado. □

Concluimos esta sección aplicando los teoremas anteriores a las matrices que se ofrecieron en los Ejemplos 2.4.1 y 2.4.2 comprobando que el resultado coincide con el ya visto. En ambos casos se cumple que $n_1 = 1$, $r_1 = 1$, y escribimos la matriz $P_A^{-1}EP_A$ de la forma

$$P_A^{-1}EP_A = \begin{pmatrix} * & * & * \\ \clubsuit & * & * \\ \heartsuit & \spadesuit & * \end{pmatrix}$$

Por tanto, por el apartado 1. del Teorema 2.4.16 se tiene que $\kappa(A, \lambda; \mathbb{S}) = (2, \tilde{\alpha})$ con

$$\tilde{\alpha} = \sup_{E\in\mathbb{S}, ||E||\leq 1} |\clubsuit + \spadesuit|.$$

En el caso de la matriz antisimétrica compleja

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

se tiene

$$P_A^{-1}EP_A = \begin{pmatrix} * & * & * \\ -iz - x & * & * \\ 0 & -iz - x & * \end{pmatrix}$$

y para la matriz

$$A = \begin{pmatrix} \sqrt{2} & i & 0 \\ i & 0 & i \\ 0 & i & -\sqrt{2} \end{pmatrix}$$

se tiene

$$P_A^{-1} E P_A = \begin{pmatrix} * & * & * \\ iy + \sqrt{2}x & * & * \\ 0 & \sqrt{2}x + iy + iz & * \end{pmatrix}.$$

Como se ve, los resultados obtenidos coinciden con los calculados en los Ejemplos 2.4.1 y 2.4.2.

Otro ejemplo completamente no genérico es la estructura de matrices

$$\mathbb{S} = \{A \in \mathbb{C}^{n\times n} : A^T M = -MA\}$$

donde M es una matriz ortogonal simétrica. Sabemos gracias al apartado (I) del Teorema 2.2.7 que si $A \in \mathbb{S}$ tiene un autovalor $\lambda = 0$ con n_1 impar y $r_1 = 1$, entonces $n_{\mathbb{S}} < n_1$. Si, además, $n_2 \neq n_1 - 1$, el apartado 1 del Teorema 2.4.16 nos dice que

$$\kappa(A, \lambda; \mathbb{S}) = (n_1 - 1, \widetilde{\alpha}),$$

donde

$$\widetilde{\alpha} = \sup_{E\in\mathbb{S}, ||E||\leq 1} |\Phi_{12} + \Phi_{21}|.$$

Además, como $r_1 = 1$,

$$\Phi_{12} + \Phi_{21} = y_1^H E x_2 + y_2^H E x_1 = \text{traza}(\begin{pmatrix} y_1^H \\ y_2^H \end{pmatrix} E \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}),$$

donde x_1 y x_2 (resp., y_1 e y_2) son los dos primeros vectores de la única cadena derecha (resp., izquierda) de Jordan asociada al bloque de tamaño n_1. Usando como antes [71, Theorem 7.3], se comprueba fácilmente que la relación entre los vectores de Jordan izquierdos y derechos es

$$\begin{pmatrix} y_1^H \\ y_2^H \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1^T \\ x_2^T \end{pmatrix} M,$$

de modo que

$$\widetilde{\alpha} = \sup_{E\in\mathbb{S}, ||E||\leq 1} |\Phi_{12} + \Phi_{21}| = \sup_{E\in\mathbb{S}, ||E||\leq 1} \left| \text{traza}\left(ME \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1^T \\ x_2^T \end{pmatrix} \right) \right|.$$

Si llamamos

$$K = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1^T \\ x_2^T \end{pmatrix}, \tag{2.53}$$

entonces la perturbación $E = \frac{1}{||K||_F} M\overline{K}$ es tal que $E \in \mathbb{S}$, $\|E\|_F = 1$ y

$$\tilde{\alpha} \geq \frac{1}{||K||_F} |\text{traza}(KK^H)| \geq ||K||_F.$$

Por otra parte $\tilde{\alpha} \leq \sqrt{n}||K||_F$. Así tenemos el siguiente resultado.

Lema 2.4.18 *Sea* $\mathbb{S} = \{A \in \mathbb{C}^{n\times n} : A^T M = -MA\}$, *sea* $A \in \mathbb{S}$ *con un autovalor* $\lambda = 0$ *y forma de Jordan (2.10) con* n_1 *impar,* $r_1 = 1$ *y* $n_2 \neq n_1 - 1$, *y sea la matriz* K *como en (2.53). Entonces* $\kappa(A, \lambda; \mathbb{S}) = (n_1 - 1, \tilde{\alpha})$ *con*

$$||K||_F \leq \tilde{\alpha} \leq \sqrt{n}||K||_F.$$

2.5. Conclusiones y trabajos futuros

Hemos introducido una definición de número de condición de tipo Hölder para autovalores múltiples, tanto para haces regulares de matrices como para matrices. En ambos casos se han comparado los números de condición estructurado y no estructurado para diversos tipos de estructuras. También hemos propuesto una definición de condicionamiento para autovalores de polinomios matriciales regulares vía la linealización en forma *compañera*. De acuerdo con los resultados obtenidos, el comportamiento bajo perturbaciones estructuradas de los autovalores múltiples difiere poco del comportamiento de los autovalores simples descrito en [16, 55, 82], en el sentido de que la estructura influye poco sobre el condicionamiento, salvo en situaciones muy particulares. Todas estas situaciones especiales parecen provenir de la combinación de simetría con antisimetría, tanto en el caso de matrices que son antisimétricas con respecto a un producto escalar simétrico (Teorema 2.2.7, apartados (ii) y (iv)) como en el caso de haces de matrices simétricos/antisimétricos (Teorema 2.3.5 (iv)) o el de haces palindrómicos (Teorema 2.3.6). Entender por qué ocurre esto es una de las cuestiones que quedan abiertas para trabajos futuros.

También hemos llevado a cabo un análisis detallado, vía el diagrama de Newton, de la teoría de perturbaciones que se requiere en el caso que hemos llamado completamente no genérico, esto es, aquél en que las perturbaciones estructuradas inducen un comportamiento cualitativamente mejor en los autovalores que las perturbaciones arbitrarias, no estructuradas. Hemos caracterizado las estructuras lineales completamente genéricas a través de una condición matricial de ortogonalidad (Teorema 2.4.6) y hemos obtenido, usando el diagrama de Newton, expresiones simplificadas para los números de condición estructurados correspondientes (Teoremas 2.4.16 y 2.4.17).

Sin embargo, estas expresiones no son aún explícitas, sino que vienen dadas como la solución de un problema no trivial de optimización. Queda como problema abierto el obtener expresiones explícitas para los números de condición como se hizo en la sección 2.2.1 o, si eso no fuera posible, llegar a estimaciones como la obtenida en el Lema 2.4.18.

Capítulo 3

Algoritmos espectrales de alta precisión relativa para matrices simétricas estructuradas

3.1. Preliminares

Ya vimos en el Capítulo 1 que, si denotamos por λ_i y v_i, respectivamente, a los autovalores y autovectores exactos de una matriz simétrica $A \in \mathbb{R}^{n\times n}$, y denotamos por $\widehat{\lambda}_i$ y $\widehat{v}_i$ a los autovalores y autovectores calculados por un algoritmo convencional (como QR, o divide-y-vencerás), entonces $\widehat{\lambda}_i$ y $\widehat{v}_i$ son los autovalores y autovectores exactos de una matriz $\widehat{A} = A + E$ tal que

$$\frac{||E||_2}{||A||_2} = O(\varepsilon), \tag{3.1}$$

donde ε representa el épsilon-máquina, una cantidad que mide la precisión con la que trabaja la aritmética finita de nuestro ordenador. Entonces, el Teorema de Weyl [87, Corollary 4.10, p. 203] asegura que

$$|\lambda_i - \widehat{\lambda}_i| \leq ||E||_2 \quad i = 1, 2, \ldots, n.$$

En cuanto a los autovectores, el Teorema de Davis y Kahan [19] nos dice que

$$\frac{1}{2} sen(2\theta) \leq \frac{||E||_2}{\text{mín}_{j\neq i} |\lambda_i - \widehat{\lambda}_i|} \quad i = 1, 2, \ldots, n$$

donde θ representa el ángulo entre el autovector exacto v_i y el calculado $\widehat{v}_i$. Combinando estas fórmulas con la estabilidad regresiva dada en (3.1), se obtiene

$$\frac{|\lambda_i - \widehat{\lambda}_i|}{|\lambda_i|} \leq O(\varepsilon) \frac{\text{máx}\, |\lambda_j|}{|\lambda_i|} \quad i = 1, 2, \ldots, n$$

y

$$\frac{1}{2}\text{sen}(2\theta) \leq O(\varepsilon)\frac{\text{máx}\,|\lambda_j|}{\text{mín}_{j\neq i}\,|\lambda_i - \widehat{\lambda}_i|} \quad i = 1, 2, \ldots, n \quad .$$

Por tanto, se puede asegurar alta precisión relativa, tal y como se definió en (1.3), sólo para los autovalores más grandes de la matriz A, pero no se tiene información sobre la precisión de los autovalores más pequeños. Un ejemplo de esta situación es la matriz

$$\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1'000001 \end{pmatrix}.$$

Si se calculan los autovalores con el comando *eig* de MATLAB, que emplea el algoritmo QR con una precisión de $\varepsilon \approx 10^{-16}$, se obtiene un error relativo para todos los autovalores del orden de ε *excepto* para el autovalor más pequeño $\lambda = 3{,}333331 \times 10^{-7}$, para el cual el valor calculado $\widehat{\lambda} = 0{,}00000033333315$ cumple que $\dfrac{|\lambda - \widehat{\lambda}|}{|\lambda|} \approx 10^{-10} \neq O(10^{-16})$.

En diversas situaciones se necesitan algoritmos que sean capaces de calcular autovalores y autovectores con tantas cifras significativas correctas como sea posible. Así, dada una matriz $A \in \mathbb{R}^{n\times n}$, lo mejor que se puede esperar es que todos sus autovalores calculados $\widehat{\lambda}_i$ y sus correspondientes autovectores $\widehat{v}_i,\ i = 1, \ldots, n$, cumplan

$$\begin{aligned} \frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} &= O(\kappa\varepsilon), \\ \Theta(\widehat{v}_i, v_i) &= \frac{O(\kappa\varepsilon)}{\text{relgap}(\lambda_i)}, \end{aligned} \tag{3.2}$$

donde λ_i son los autovalores exactos de A, q_i los autovectores exactos, $\Theta(\cdot,\cdot)$ representa el ángulo entre dos vectores, $\text{relgap}(\lambda_i)$ es la separación relativa definida como

$$\text{relgap}(\lambda_i) = \min\left\{\text{min}_{j\neq i}\frac{|\lambda_j - \lambda_i|}{|\lambda_i|}, 1\right\},\ i = 1, \ldots, n, \tag{3.3}$$

y κ es una constante de tamaño moderado, que por lo general depende de la dimensión de la matriz A, pero que es siempre independiente del número de condición de A.

Los algoritmos que calculan los autovalores y autovectores con esta precisión se llaman, como ya se dijo en el capítulo introductorio, *algoritmos espectrales de alta precisión relativa*. Una definición similar puede darse para el problema de valores singulares (SVD).

Existen algoritmos de alta precisión relativa tanto para el problema espectral como para el problema de valores singulares. Todos estos algoritmos tienen en común el hecho de que la alta precisión relativa se alcanza sólo para clases específicas de matrices, y cada clase de matrices requiere un algoritmo distinto, adaptado a ese tipo de matrices. La propiedad común a todas estas clases de matrices es que es posible calcular con alta precisión un cierto tipo de factorización, que definiremos en breve (véase la Definición 3.1.1).

De entre todos estos métodos nos centraremos, como ya se anunció en el Capítulo 1, en una familia de algoritmos espectrales [22, 30, 84] que constan de dos etapas:

Primera etapa: cálculo de una factorización inicial de la matriz.

Segunda etapa: aplicación de un algoritmo de tipo Jacobi a los factores de la descomposición obtenida en la primera etapa.

Para obtener las cotas de alta precisión relativa (3.2), ambas etapas deben llevarse a cabo con una precisión adecuada. La precisión de la segunda etapa suele garantizarse mediante un análisis de errores que supone asegurada la alta precisión en la primera etapa. Por tanto, se puede afirmar que la precisión de esta familia de algoritmos depende esencialmente de la precisión obtenida en la factorización de la primera etapa. En otras palabras, *las clases de matrices para las que estos algoritmos calculan autovalores y autovectores con alta precisión relativa son aquellas clases que permiten calcular con alta precisión relativa la factorización inicial.*

En el problema de valores singulares se pueden encontrar muchas clases de matrices (véase [22]) tales que eliminación gaussiana con pivote completo, adaptada convenientemente a cada una de las clases, conduce a una factorización *no simétrica* de la forma $A = X\Delta Y^T$, calculada con alta precisión relativa. Dos de estas familias son las matrices DSTU (totalmente unimodulares diagonalmente escaladas) y las matrices TSC (*total signed compound*), veánse secciones 3.3.1 y 3.3.2 para sus respectivas definiciones. Nuestra principal aportación en este capítulo es el desarrollo de algoritmos de cálculo de una factorización de la forma $A = X\Delta X^T$ de las matrices *simétricas* A en las estructuras DSTU y TSC. Estos algoritmos calculan la factorización con la precisión requerida para que sus autovalores y autovectores se calculen después con la alta precisión relativa (3.2). Así, para cualquier matriz simétrica perteneciente a una de las clases DSTU o TSC, sus autovalores y autovectores podrán calcularse con alta precisión relativa.

Dos algoritmos de alta precisión relativa que calculan autovalores y autovectores de matrices simétricas posiblemente indefinidas son los siguientes:

1. *Algoritmo J-ortogonal* [85, 84], el cual consta de dos etapas:

   a) La matriz simétrica $A \in \mathbb{R}^{n\times n}$ se descompone de la forma

   $$A = GJG^T, \quad J = I_{npos} \oplus -I_{r-npos} \tag{3.4}$$

   siendo *npos* el número de autovalores positivos, r el rango de la matriz A y $G \in \mathbb{R}^{n\times r}$ una matriz de rango completo.

   b) Aplicación sucesiva de transformaciones de la forma

   $$G_{m+1} = G_m J_m, \tag{3.5}$$

donde $G_0 = G$ y J_m es una matriz ortogonal que contiene *rotaciones hiperbólicas* de la forma

$$\begin{pmatrix} cosh(\theta) & -senh(\theta) \\ senh(\theta) & cosh(\theta) \end{pmatrix}.$$

Esta etapa calcula de modo implícito los autovalores y autovectores de G a partir de los autovalores y autovectores del haz (G^TG, J).

2. *Algoritmo SVD con signos* [30], compuesto por las etapas:

   a) Dada la matriz simétrica $A \in \mathbb{R}^{n\times n}$ se calcula una descomposición de la forma $A = XDY^T$, donde X, Y son matrices bien condicionadas y D es una matriz diagonal.

   b) Cálculo de una descomposición en valores singulares de XDY^T.

   c) Cálculo de los autovalores y autovectores de A a partir de sus valores y vectores singulares.

Ninguno de estos algoritmos, estrictamente hablando, produce errores de la forma (3.2). Para el método J-ortogonal, la constante κ en (3.2) es el máximo de los números de condición de matrices intermedias que se utilizan en el algoritmo, y estos números de condición podrían ser, en principio, magnitudes grandes. El segundo algoritmo, el SVD con signos, tiene una cota de error para el cálculo de autovectores de la forma (3.2), pero con una cantidad más pequeña, relgap*, en el denominador (véase (3.12)). Pese a todo, en la práctica ambos algoritmos son capaces de calcular autovalores y autovectores con alta precisión relativa.

Como se ha observado antes, salvo una ligera diferencia, ambos algoritmos tienen en común la primera etapa. Para ser más precisos en la descripción de esa primera etapa comenzaremos definiendo lo que es una *descomposición que revela el rango* (en inglés, *rank-revealing decomposition*).

Definición 3.1.1 *Dada una matriz $A \in \mathbb{R}^{m\times n}$ con $m \geq n$ y rango r, una* **descomposición que revela el rango** *de A es una factorización de la forma $A = X\Delta Y^T$ con $X \in \mathbb{R}^{m\times r}$, $Y \in \mathbb{R}^{n\times r}$, $\Delta \in \mathbb{R}^{r\times r}$, donde Δ es una matriz diagonal y no singular y las matrices X, Y son matrices bien condicionadas. Para abreviar, diremos que cualquier descomposición de este tipo es una* **RRD** *de A.*

La descomposición en valores singulares, por ejemplo, es una RRD. Pero se pueden usar otros métodos para la obtención de una RRD, como eliminación gaussiana con pivote completo, QR con pivote completo o el método de pivote diagonal.

La descomposición inicial (3.4) que utiliza el método J-ortogonal no es exactamente una descomposición RRD, pero si se escala la matriz Δ a ambos lados con $|\Delta|^{-1/2} = diag(|\Delta_{ii}^{-1/2}|)$, obtenemos una factorización simétrica de la forma $A = GJG^T$ con $G = X|\Delta|^{1/2}$ y $J = |\Delta|^{-1/2}\Delta|\Delta|^{-1/2}$.

Como las matrices que se van a estudiar son matrices simétricas, se estudiarán descomposiciones RRD simétricas, con la idea de preservar la estructura.

El camino que se utilizará para obtener una descomposición simétrica RRD es la factorización LDL^T por bloques [43, Capítulo 11]. Dada una matriz simétrica $A \in \mathbb{R}^{n\times n}$ existe una matriz de permutación P tal que

$$PAP^T = LDL^T, \tag{3.6}$$

donde la matriz L es una matriz triangular inferior con unos en la diagonal y D es una matriz diagonal con bloques diagonales de tamaños 1×1 y 2×2. Esta factorización no es una RRD ya que la matriz D no es diagonal, pero si diagonalizamos los bloques de tamaño 2×2 mediante rotaciones de Givens, entonces podemos conseguir una descomposición RRD simétrica como se muestra en [85, 86] y explicamos a continuación.

Sea $Q \in \mathbb{R}^{n\times n}$ una matriz ortogonal, diagonal por bloques particionada de manera conforme con la matriz D: cada bloque diagonal de tamaño 2×2 de Q es una rotación de Givens 2×2 que diagonaliza el correspondiente bloque de la matriz D. Por lo tanto,

$$A = X\Delta X^T, \tag{3.7}$$

con

$$X = P^T LQ, \qquad \text{y} \qquad \Delta = Q^T DQ. \tag{3.8}$$

Puesto que la matriz Q es una matriz ortogonal, el número de condición de la matriz X para cualquier norma unitariamente invariante coincide con el número de condición del factor L. Por tanto, para demostrar que la matriz X está bien condicionada basta estudiar el condicionamiento de la matriz L.

Se puede demostrar fácilmente que la descomposición simétrica RRD determina la descomposición espectral con alta precisión relativa [29]. Más en concreto, se puede demostrar que si $A = XDX^T$ y $\widetilde{A} = \widetilde{X}\widetilde{D}\widetilde{X}^T$ son, respectivamente, descomposiciones RRD simétricas de dos matrices simétricas A y $\widetilde{A}$, y el error relativo en norma

$$\|\widetilde{X} - X\|/\|X\|$$

en el factor X, y el error relativo componente a componente

$$|\widetilde{D}_{ii} - D_{ii}|/|D_{ii}|$$

en el factor D están acotados por una cantidad β menor que 1, entonces, denotando por

$$\eta = \beta(2+\beta)\kappa(X),$$

el error relativo de los autovalores está acotado por $O(\eta)$ y el seno de los ángulos canónicos entre los autovectores de las matrices A y $\widetilde{A}$ está acotado por $O(\eta)$ dividido por la separación relativa (para más detalles véase [29, §2]). Esto equivale a decir que cambios pequeños en los factores de la descomposición RRD producen cambios pequeños en los autovalores y autovectores.

A la vista de ello, se concluye que para asegurar alta precisión relativa para ambos algoritmos, el J-ortogonal o el SVD con signos, basta probar que la factorización inicial se calcula con una precisión suficiente. A ello dedicamos el presente capítulo.

Para determinar la precisión con la que se ha obtenido una descomposición RRD, procederemos del siguiente modo:

Etapa 1
Veremos en primer lugar cómo se puede calcular una factorización LDL^T por bloques de matrices simétricas DSTU y TSC con error relativo pequeño componente a componente, esto es, si $\widehat{L}$ y $\widehat{D}$ son los factores calculados en aritmética finita, y L, D son los factores exactos, se demostrará que

$$|\widehat{l}_{ij} - l_{ij}| = O(\varepsilon)|l_{ij}|, \qquad |\widehat{d}_{ij} - d_{ij}| = O(\varepsilon)|d_{ij}| \tag{3.9}$$

para cada $i, j \in \{1, \ldots, n\}$. Para ello *es suficiente probar que no se producen restas en todo el proceso de la factorización*, ya que los productos, cocientes, raíces cuadradas y sumas de cantidades del mismo signo en coma flotante no producen errores progresivos grandes y *la única posibilidad de producir un error progresivo grande es la cancelación*, producida por la resta de números cercanos.

Etapa 2
Analizaremos si la descomposición RRD, obtenida a partir de la factorización por boques LDL^T vía rotaciones de Givens, como en (3.8), cumple los requisitos que aseguran la precisión (3.2). De acuerdo con el análisis de errores en [30], estos requisitos son que el factor Δ en (3.7) se calcule con errores relativos pequeños *componente a componente* y que el factor X se calcule con error relativo pequeño en *norma* para cualquier norma unitariamente invariante, esto es,

$$\|\widehat{X} - X\| = O(\varepsilon)\|X\|, \qquad |\widehat{\Delta}_{ii} - \Delta_{ii}| = O(\varepsilon)|\Delta_{ii}|, \quad i = 1, \ldots, n, \tag{3.10}$$

siendo $\widehat{X}, \widehat{\Delta}$ los factores calculados en aritmética en coma flotante y X, Δ los factores exactos.

Todo esto conducirá a probar que los autovalores y respectivos autovectores de matrices simétricas con estructura DSTU o TSC se calculan con alta precisión relativa mediante el algoritmo SVD con signos. Para ser más precisos, el error relativo en los autovalores calculados será de la forma (3.2), donde κ viene dado por

$$\kappa = \kappa(R')\kappa(X), \tag{3.11}$$

donde $\kappa(\cdot)$ denota el número de condición en norma dos, X es el factor no diagonal en (3.7), y R' es el mejor escalamiento por filas del factor triangular R de una descomposición QR con pivote por columnas de la matriz $X\Delta$. En [22, Teorema 3.2] se demuestra que $\kappa(R')$ es a lo sumo de orden $O(n^{3/2}\kappa(X))$, luego el orden de la constante κ depende sólo del número de condición del factor X de la RRD simétrica . De ahí la importancia de demostrar que

el número de condición de X no es grande. Veremos en el Teorema 3.3.5 que para matrices simétricas DSTU el número de condición $\kappa(X)$ es del orden de n^2.

En cuanto a los autovectores, se calculan con un error de la forma (3.2), pero reemplazando la separación relativa (3.3) usual por

$$\operatorname{relgap}^*(|\lambda_i|) = \text{mín} \left\{ \min_{\substack{j\in\mathcal{S} \\ j\neq i}} \left| \frac{|\lambda_j| - |\lambda_i|}{\lambda_i} \right|, 1 \right\}, \tag{3.12}$$

donde el conjunto de índices $\mathcal{S}$ es igual a $\{1, \ldots, n\}$, salvo que el autovalor, digamos λ_{j_0}, con el valor absoluto más cercano a $|\lambda_i|$ tenga signo opuesto al de λ_i. En tal caso, $\mathcal{S}$ se obtiene del conjunto $\{1, \ldots, n\}$ eliminando el índice j_0 y el índice k de cualquier otro autovalor que se encuentre a una distancia de orden $O(\kappa\varepsilon)$ del autovalor λ_{j_0}.

Finalmente, debemos hacer notar que nuestro análisis de errores es válido sólo para matrices simétricas *no singulares*: si A es singular, el número de autovalores nulos se obtiene de cualquier RRD que cumpla (3.10), y puede modificarse el algoritmo SVD con signos para calcular una base del núcleo, usando la factorización QR con pivote completo. Sin embargo, este procedimiento queda fuera de nuestro análisis de errores.

3.2. Factorización LDL^T por bloques de matrices simétricas

La descomposición por bloques LDL^T (3.6) es una versión simétrica de la factorización LU. Cualquier matriz simétrica admite una factorización de esta forma [43, Capítulo 11], y el procedimiento más común para calcularla es el *método de pivote diagonal*: este comienza por elegir una matriz de permutación P, un entero $s = 1$ ó $s = 2$, y un pivote no singular E de tamaño $s \times s$ tal que

$$PAP^T = \begin{pmatrix} E & C^T \\ C & B \end{pmatrix},$$

de modo que

$$PAP^T = \begin{pmatrix} I_s & 0 \\ CE^{-1} & I_{n-s} \end{pmatrix} \begin{pmatrix} E & 0 \\ 0 & B - CE^{-1}C^T \end{pmatrix} \begin{pmatrix} I_s & E^{-1}C^T \\ 0 & I_{n-s} \end{pmatrix}. \tag{3.13}$$

La factorización por bloques LDL^T de A se obtiene repitiendo el proceso sobre los sucesivos complementos de Schur $B - CE^{-1}C^T$. El coste en operaciones aritméticas del proceso es $n^3/3$ más el coste de elegir las permutaciones.

Para la elección del pivote en eliminación gaussiana existen estrategias de pivote parcial, pivote completo o pivote mixto. Para la factorización simétrica LDL^T también existen estrategias simétricas para la elección del pivote E. Como nuestro objetivo es una descomposición RRD, usaremos la estrategia de pivote de Bunch–Parlett [14], que es un análogo simétrico de la estrategia de pivote completo. Esta estrategia produce, en la práctica, factores L bien condicionados (véase [13] para un análisis detallado del factor de crecimiento del factor L). Esta estrategia se resume como sigue.

$$
\begin{array}{l}
\alpha = (1+\sqrt{17})/8 \approx 0{,}64 \\
\mu_0 = \max_{i,j}|a_{ij}| =: |a_{pq}| \\
\mu_1 = \max_{i}|a_{ii}| =: |a_{rr}| \\
\textbf{If } \mu_1 \geq \alpha\mu_0 \textbf{ then} \\
\qquad \text{elegir } E = [a_{rr}] \text{ como pivote } 1\times 1 \\
\textbf{else} \\
\qquad \text{elegir } E = \begin{pmatrix} a_{pp} & a_{pq} \\ a_{pq} & a_{qq} \end{pmatrix} \text{ como pivote } 2\times 2.
\end{array}
\tag{3.14}
$$

Esto significa que elegimos pivote 2×2 cuando el elemento de mayor valor absoluto de la matriz es bastante más grande que el elemento de mayor valor absoluto de la diagonal. Cualquier pivote 2×2 elegido con esta estrategia es una matriz simétrica indefinida y bien condicionada, ya que su número de condición en norma dos está acotada por

$$(1+\alpha)/(1-\alpha) \approx 4{,}6.$$

Se elige el valor $(1+\sqrt{17})/8$ de la constante α para que el factor de crecimiento de dos pasos consecutivos con pivotes 1×1 sea igual al de un solo paso con pivote 2×2 (véase [43, Capítulo 11]).

Es bien conocido que cualquier valor intermedio o final del proceso de eliminación gaussiana con cualquier estrategia de pivote es o un menor, o un cociente de menores de la matriz original [22, Lemma 5.1]. Descomponiendo una matriz cuadrada A de la forma $\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$, y denotando el complemento de Schur del bloque A_{11} en A como

$$\mathcal{C}_k = A_{22} - A_{21}A_{11}^{-1}A_{12}, \tag{3.15}$$

tenemos el siguiente resultado.

Lema 3.2.1 *Sea $A = PLDUQ$ una factorización de A, calculada mediante eliminación gaussiana con cualquier estrategia de pivotaje, donde P y Q son matrices de permutación, L es una matriz triangular inferior, U es triangular superior, ambas con unos en la diagonal, y D es una matriz diagonal. Reescribiendo $A = P\overline{L}\,\overline{U}Q$ con $\overline{L} = L$ y $\overline{U} = DU$ y denotando por $\mathcal{C}_k$ cualquier complemento de Schur de A como en (3.15), entonces*

1. *Cualquier menor de $\mathcal{C}_k$ es un cociente de menores, o simplemente un menor, de A.*
2. *Cualquier menor de A^{-1} es, salvo signos, un cociente de menores de A.*
3. *Cualquier entrada de L, D, U o $\overline{U}$ es o bien cero, o bien un menor de A, o bien un cociente de menores de A.*
4. *Cualquier menor de L, D, U o $\overline{U}$ formado por filas consecutivas es o bien cero, o bien un menor de A, o bien un cociente de menores de A.*

5. *Cualquier entrada de L^{-1}, D^{-1}, U^{-1} o $\overline{U}^{-1}$ es o bien cero, o bien un menor de A, o bien un cociente de menores de A.*

Veamos ahora cuáles de las propiedades dadas en el lema se cumplen también para la descomposición LDL^T por bloques; veremos que las propiedades 1, 2 y 3 del Lema 3.2.1 se preservan, dando de hecho sus fórmulas explícitas. Sin embargo las propiedades 4 y 5, no se cumplen, como se demostrará con un contraejemplo.

Antes de escribir estos resultados, necesitamos un resultado auxiliar, consecuencia de la identidad de Sylvester (véase, por ejemplo, [18, §2]).

Lema 3.2.2 *Sea M cualquier matriz cuadrada, particionada de la forma*

$$M = \begin{pmatrix} B & * \\ * & * \end{pmatrix}, \qquad con \qquad B = \begin{pmatrix} B_1 & * \\ * & * \end{pmatrix}$$

donde B y B_1 son submatrices cuadradas y no singulares. Sea $\mathcal{C}_1^B$ (resp. $\mathcal{C}_1^M$) el complemento de Schur de B_1 en B (resp. en M). Entonces, el complemento de Schur de B en M es igual al complemento de Schur de $\mathcal{C}_1^B$ en $\mathcal{C}_1^M$.

Para no complicar la notación, la Fórmula 3.15 y el Lema 3.2.3 vienen dados en términos del complemento de Schur del bloque principal superior izquierdo. Por supuesto, todos los resultados son ciertos para cualquier submatriz cuadrada y no singular de la matriz A. Para enunciar los siguientes resultados se utilizará la notación de MATLAB: $A([r_1, \ldots, r_p], [c_1, \ldots, c_p])$ denota la submatriz de tamaño $p \times p$ de A que contiene los elementos de las filas $r_1, \ldots, r_p$ y las columnas $c_1, \ldots, c_p$. Además, $1 : k$ denota la lista de enteros desde 1 hasta k.

Lema 3.2.3 *Sea $A \in \mathbb{R}^{n\times n}$, sea $k < n$ y sea $A_k = A(1 : k, 1 : k)$ la submatriz superior izquierda de A. Entonces,*

(a) el elemento (i, j) del complemento de Schur $\mathcal{C}_k$ de tamaño $(n - k) \times (n - k)$ de A_k en A viene dado por

$$\mathcal{C}_k(i, j) = \frac{\det A([1 : k, k + i], [1 : k, k + j])}{\det A([1 : k], [1 : k])} \tag{3.16}$$

para cada $i, j \in \{1, \ldots, n - k\}$.

(b) para cualquier $s \leq n - k$, el menor de $\mathcal{C}_k$ que contiene las filas $i_1 < \ldots < i_s$ y las columnas $j_1 < \ldots < j_s$ viene dado por

$$\begin{aligned} &\det \mathcal{C}_k([i_1, \ldots, i_s], [j_1, \ldots, j_s]) = \\ &\qquad = \frac{\det A([1 : k, k + i_1, \ldots, k + i_s], [1 : k, k + j_1, \ldots, k + j_s])}{\det A[1 : k], [1 : k]} \end{aligned} \tag{3.17}$$

En particular, cada menor del complemento de Schur $\mathcal{C}_k$ es cociente de menores de la matriz original A.

Demostración:
Probaremos (a) por inducción sobre k. Si $k = 1$, los elementos de $\mathcal{C}_1$ son

$$\mathcal{C}_1(i,j) = a_{i+1,j+1} - \frac{a_{i+1,1}\, a_{1,j+1}}{a_{11}} = \frac{a_{i+1,j+1}\, a_{11} - a_{i+1,1}\, a_{1,j+1}}{a_{11}} =$$

$$= \frac{\det A([1, i+1], [1, j+1])}{\det A([1], [1])},$$

es decir, la ecuación (3.16) con $k = 1$. Supongamos que (3.16) es cierto para algún $k \in \{1, 2 \ldots n\}$. Demostraremos que también es cierto para $k + 1$. Según el Lema 3.2.2, el complemento de Schur $\mathcal{C}_{k+1}$ de A_{k+1} en A es el resultado de realizar dos complementos de Schur sucesivos: primero, el complemento de Schur $\mathcal{C}_k$ de A_k en A, y después el complemento de Schur del elemento $(1, 1)$ en $\mathcal{C}_k$. Por la hipótesis de inducción, sabemos que

$$\mathcal{C}_k(i,j) = \frac{\det A([1:k, k+i], [1:k, k+j])}{\det A([1:k], [1:k])}.$$

Sustituyendo en la fórmula

$$\mathcal{C}_{k+1}(i,j) = \mathcal{C}_k(i+1, j+1) - \frac{\mathcal{C}_k(i+1,1)\mathcal{C}_k(1,j+1)}{\mathcal{C}_k(1,1)}$$

para los elementos de $\mathcal{C}_{k+1}$ nos lleva a

$$\mathcal{C}_{k+1}(i,j) = \frac{\begin{vmatrix} \det A([1:k, k+i+1], [1:k, k+j+1]) & \det A([1:k, k+1], [1:k, k+j+1]) \\ \det A([1:k, k+i+1], [1:k, k+1]) & \det A([1:k, k+1], [1:k, k+1]) \end{vmatrix}}{\det A([1:k], [1:k])\, \det A([1:k, k+1], [1:k, k+1])}$$

Basta aplicar la identidad de Sylvester [48, p. 22] al numerador para obtener (3.16) con $k + 1$ en lugar de k.

Una vez probado (a), todos los elementos de $\mathcal{C}_k([i_1, \ldots, i_s], [j_1, \ldots, j_s])$ pueden escribirse como cocientes de menores de la matriz A, con el mismo denominador

$$d_k = \det A([1:k], [1:k]).$$

Así, la submatriz puede escribirse como $(1/d_k)\, M$, donde para cada $l, m \in \{1, \ldots, s\}$, el elemento (l, m) de M es $\det A([1:k, k+i_l], [1:k, k+j_m])$.Aplicando de nuevo la identidad de Sylvester a M se demuestra la parte (b).

□

Como consecuencia del Lema 3.2.3, se demuestra el siguiente teorema:

Teorema 3.2.4 *Sea A una matriz real simétrica y sea $PAP^T = LDL^T$ una factorización LDL^T por bloques de la matriz A como en (3.6), obtenida usando cualquier estrategia de pivotaje. Entonces cualquier entrada del factor L o el factor D es o bien cero, o bien un cociente de menores ede A o simplemente un menor de A.*

Demostración:

La demostración es similar a la del Lema 5.1 en [22, p. 52]. Las entradas de D son o elementos de la matriz A o elementos de un complemento de Schur de la matriz A. Por el Lema 3.2.3, cada elemento de D es un elemento de A o un cociente de menores de A. Por otra parte, cada elemento l_{ij} de L generado por un pivote 1×1 es un cociente de dos elementos del correspondiente complemento de Schur de A y, como ambos elementos han sido creados en la misma etapa de la factorización, por el apartado (a) del Lema 3.2.3 son cocientes de la forma (3.16) con el mismo denominador. Por lo tanto, ambos denominadores se cancelan y se obtiene que el elemento l_{ij} es un cociente de menores de A. El argumento es similar para las entradas de L generadas por un pivote de tamaño 2×2, usando el apartado (b) del Lema 3.2.3.

□

Por otra parte, la factorización LDL^T no cumple las propiedades 4 y 5 del Lema 3.2.1, es decir, todo menor formado por filas consecutivas de la matriz L no es necesariamente un cociente de menores de la matriz original A. Si consideramos, por ejemplo, la matriz

$$A = \begin{pmatrix} 13 & 39 & 65 \\ 39 & 128 & 274 \\ 65 & 274 & 903 \end{pmatrix} = LDL^T$$

con

$$L = \begin{pmatrix} 1 & & \\ 3 & 1 & \\ 5 & 7 & 1 \end{pmatrix}, \qquad D = \left(\begin{array}{c|cc} 13 & & \\ \hline & 11 & 2 \\ & 2 & 11 \end{array}\right)$$

Puede comprobarse por enumeración que

$$\det L([2,3],[1,2]) = \begin{vmatrix} 3 & 1 \\ 5 & 7 \end{vmatrix} = 16$$

no es cociente de menores de A. Nótese que este menor solapa en parte con un pivote 1×1 y en parte con otro de tamaño 2×2.

Concluimos derivando fórmulas explícitas para los elementos de L y D como cociente de menores de A. Estas fórmulas, que necesitaremos más adelante para recalcular las entradas de L, son nuevas en la literatura y extienden las fórmulas clásicas para eliminación gaussiana (véase, por ejemplo,[50, §1.4]). Por simplicidad, las fórmulas están escritas sin tener en cuenta las matrices de permutación necesarias para el pivotaje, pero el resultado es válido trivialmente para la factorización (3.6) sin más que renombrar filas y columnas.

Lema 3.2.5 *Sea $A \in \mathbb{R}^{n\times n}$ una matriz simétrica factorizada como $A = LDL^T$ con $L \in \mathbb{R}^{n\times n}$ triangular inferior con unos sobre la diagonal y $D \in \mathbb{R}^{n\times n}$ diagonal por bloques de tamaños 1×1 y 2×2. Sea $D = diag(\mathrm{D}_1, \ldots, \mathrm{D}_\mathrm{r})$ con $D_k \in \mathbb{R}^{s_k\times s_k}$, $s_k = 1$ ó 2, $k = 1, \ldots, r$, y una partición conforme de L, $L = [L_1 \mid \cdots \mid L_r]$, con $L_k \in \mathbb{R}^{n\times s_k}$. Para cada $k \in \{1, \ldots, r\}$,*

sea $n_k = s_1 + \ldots + s_k$. *Entonces, para cada* $k \in \{1, \ldots, r\}$, *los elementos* (i, j) *de* L *con* $j \in \{n_k - s_k + 1, \ldots, n_k\}$ *vienen dados por*

$$L(i,j) = \frac{\det A([1 : j-1, i, j+1 : n_k], [1 : n_k])}{\det A([1 : n_k], [1 : n_k])}, \qquad i = n_k + 1, \ldots, n, \tag{3.18}$$

y, si $n_0 = 0$, *los elementos* (i, j) *de* D *con* $i, j \in \{n_k - s_k + 1, \ldots, n_k\}$ *vienen dados por*

$$D(i,j) = \frac{\det A([1 : n_{k-1}, i], [1 : n_{k-1}, j])}{\det A([1 : n_{k-1}], [1 : n_{k-1}])}. \tag{3.19}$$

Demostración:
Se distinguen los casos $s_k = 1$ y $s_k = 2$. Si $s_k = 1$ entonces $n_k = n_{k-1} + 1$ y

$$L(i, n_k) = \frac{\mathcal{C}_{k-1}(i, 1)}{\mathcal{C}_{k-1}(1, 1)}, \qquad i \in \{n_k + 1, \ldots, n\},$$

que, según el apartado (a) del Lema 3.2.3, y una vez simplificado, es igual a

$$L(i, n_k) = \frac{\det A([1 : n_{k-1}, n_{k-1} + i], [1 : n_{k-1}, n_{k-1} + 1])}{\det A([1 : n_{k-1}, n_{k-1} + 1], [1 : n_{k-1}, n_{k-1} + 1])} =$$

$$= \frac{\det A([1 : n_{k-1}, i], [1 : n_k])}{\det A([1 : n_k], [1 : n_k])}$$

Si $s_k = 2$ entonces $n_k = n_{k-1} + 2$ y, para cada $i \in \{n_k + 1, \ldots, n\}$, tenemos

$$L(i, n_k - 1) = \frac{\begin{vmatrix} \mathcal{C}_{k-1}(i,1) & \mathcal{C}_{k-1}(i,2) \\ \mathcal{C}_{k-1}(1,2) & \mathcal{C}_{k-1}(2,2) \end{vmatrix}}{\begin{vmatrix} \mathcal{C}_{k-1}(1,1) & \mathcal{C}_{k-1}(1,2) \\ \mathcal{C}_{k-1}(2,1) & \mathcal{C}_{k-1}(2,2) \end{vmatrix}}$$

y

$$L(i, n_k) = \frac{\begin{vmatrix} \mathcal{C}_{k-1}(1,1) & \mathcal{C}_{k-1}(2,1) \\ \mathcal{C}_{k-1}(i,1) & \mathcal{C}_{k-1}(i,2) \end{vmatrix}}{\begin{vmatrix} \mathcal{C}_{k-1}(1,1) & \mathcal{C}_{k-1}(1,2) \\ \mathcal{C}_{k-1}(2,1) & \mathcal{C}_{k-1}(2,2) \end{vmatrix}}.$$

En ambos casos, la identidad de Sylvester, combinada con el Lema 3.2.3, conduce a la fórmula del enunciado. Finalmente, las fórmulas para los elementos de D se obtienen trivialmente usando el hecho de que los elementos de D son o bien elementos de la matriz original o bien elementos de alguno de sus complementos de Schur.

□

3.3. Factorización con alta precisión para matrices simétricas DSTU y TSC

En esta sección se demostrará que, para ambas clases de matrices DSTU y TSC, la descomposición LDL^T por bloques (3.6) puede calcularse con un error relativo pequeño *componente a componente*, como se indicó en (3.9). Para demostrarlo, se modificará el método de pivote diagonal, a fin de evitar posibles sustracciones en el proceso factorización (el efecto de cancelación es el único que puede producir una inestabilidad en el análisis de errores regresivos).

3.3.1. Matrices DSTU

Definición 3.3.1 *Se dice que una matriz Z con entradas enteras es* totalmente unimodular (TU) *si todos sus menores son iguales a -1, 0 ó 1. En particular, las entradas de Z sólo pueden tomar los valores -1, 0 ó 1. Una matriz A es* diagonalmente escalada totalmente unimodular (DSTU) *si $A = \mathcal{D}_L Z \mathcal{D}_R$, con Z totalmente unimodular y $\mathcal{D}_L$, $\mathcal{D}_R$ matrices diagonales.*

Un ejemplo de matriz diagonalmente escalada totalmente unimodular es el siguiente:

$$A = \begin{pmatrix} 2\sqrt{2} & 2 & 0 & 0 & 0 \\ 3\sqrt{2} & 0 & 3 & 0 & -12 \\ 0 & 4 & 0 & 12 & 0 \\ 0 & 0 & 7 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & 0 & -2 \end{pmatrix} = \mathcal{D}_L Z \mathcal{D}_R =$$

$$= \begin{pmatrix} 2 & & & & \\ & 3 & & & \\ & & -4 & & \\ & & & 7 & \\ & & & & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 \\ 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{2} & & & & \\ & 1 & & & \\ & & -1 & & \\ & & & 3 & \\ & & & & 4 \end{pmatrix}.$$

La clase de matrices TU contiene algunas clases de matrices como las matrices acíclicas, matrices de elementos finitos asociadas a sistemas lineales de masa-resorte, o las matrices de incidencia nodo-arco (vease [22, Sección 10]). Estamos interesados solamente en las matrices *simétricas* DSTU, esto es, matrices simétricas A que pueden escribirse como

$$A = \mathcal{D} Z \mathcal{D}, \tag{3.20}$$

donde Z es una matriz simétrica TU, y $\mathcal{D} = \mathrm{diag}(\mathrm{d}_1, \ldots, \mathrm{d}_\mathrm{n})$ es una matriz diagonal. La matriz Z se supone conocida exactamente, pero los elementos de la matriz $\mathcal{D}$ son sólo conocidos con alta precisión relativa, es decir, con error relativo del orden del épsilon-máquina.

Una primera propiedad relevante de las matrices DSTU es que *cualquier complemento de Schur de una matriz DSTU es también trivialmente DSTU*. Esta es una de las propiedades fundamentales que utilizaremos en cada una de las etapas de la factorización. La segunda propiedad que veremos de las matrices DSTU, y que es la clave para evitar cancelaciones, es la siguiente.

Lema 3.3.2 *Cualquier menor de una matriz simétrica DSTU es un monomio con coeficientes* $0, 1$ *ó* -1 *con los elementos diagonales* d_i *de la matriz* $\mathcal{D}$ *como variables.*

Como consecuencia de este lema, cualquier menor de una matriz DSTU puede calcularse con alta precisión relativa, y lo mismo sucede con los elementos de cualquier complemento de Schur, puesto que, según el Teorema 3.2.4, cada uno de estos elementos es un cociente de menores de la matriz original.

Otra propiedad interesante de las matrices simétricas DSTU es que los pivotes de tamaño 2×2 seleccionados por la estrategia de Bunch–Parlett tienen la siguiente estructura especial.

Lema 3.3.3 *Cualquier pivote de tamaño* 2×2 *elegido por la estrategia de pivotaje de Bunch–Parlett de una matriz simétrica DSTU tiene, al menos, una entrada nula en la diagonal.*

Demostración:
Un pivote de tamaño 2×2 se elige siempre que

$$\mu_1 = \max_i|a_{ii}| := |a_{rr}| < \alpha\,|a_{pq}| =: \alpha \max_{i,j}|a_{ij}| = \alpha\,\mu_0$$

Si suponemos que los elementos diagnales a_{pp} y a_{qq} son ambos no nulos, entonces

$$|a_{pp}| \leq \mu_1 < \alpha\,|a_{pq}| \quad \text{y} \quad |a_{qq}| \leq \mu_1 < \alpha\,|a_{pq}|, \tag{3.21}$$

de modo que $a_{pq} = d_p d_q z_{pq}$ es no nulo y consecuentemente, también lo son d_p y d_q. Esto, junto con (3.21), implica que

$$|d_p| < \alpha^2 |d_p|, \tag{3.22}$$

en contradicción con el hecho de que $\alpha < 1$. Por tanto, o bien $a_{pp} = 0$ ó $a_{qq} = 0$.

□

Nótese que la desigualdad (3.22) está también en contradicción con $\alpha = 1$. Por tanto, el *Lema 3.3.3 se cumple incluso si* $\alpha = 1$, un hecho que se utilizará en §3.3.1 al modificar la estrategia de pivote.

Como la estrategia de Bunch-Parlett puede elegir pivotes de tamaño 1×1 o tamaño 2×2, el estudio de la factorización se basará en la distinción de estos dos casos.

- **Caso 1:** *El pivote elegido* $E = [a_{rr}]$ *es* 1×1

  Los elementos de L calculados en una cierta etapa de la factorización son un cociente

$$l_{ir} = (CE^{-1})_{ir} = \frac{a_{ir}}{a_{rr}}$$

de elementos del complemento de Schur de la anterior etapa. Así, l_{ir} se calcula con un error relativo pequeño, suponiendo que los elementos del complemento de Schur a_{ir} y a_{rr} se hayan calculado también con un error relativo pequeño. Al calcular los elementos del factor D, por otro lado, podrían darse sustracciones, puesto que vienen dados por las fórmulas

$$(B - CE^{-1}C^T)_{ij} = a_{ij} - l_{ir}a_{rj}. \tag{3.23}$$

A la vista del Teorema 3.2.4 y el Lema 3.3.2, cada uno de los operandos de la sustracción es o bien cero o bien, salvo signo, un producto de potencias de los elementos de la matriz diagonal $\mathcal{D}$. De hecho, un simple cálculo demuestra que cada operando es un monomio con coeficientes ± 1 ó 0 en las variables d_i y d_j. Por lo tanto podemos reescribir (3.23) como

$$m_1 = m_2 + m_3$$

donde cada operando m_i, para $i = 1, 2, 3$, es un monomio con coeficientes ± 1 ó 0 en las dos variables d_i y d_j. Podemos distinguir cuatro posibilidades para esta operación aritmética, dependiendo del signo de m_2 y m_3. Si alguno de los sumandos es cero no hay tal operación. Si los dos operandos son no nulos entonces m_1 *por fuerza es nulo* puesto que la única forma de obtener un monomio en las mismas variables con coeficientes $1, -1$ ó 0 en el operando m_1 es que los coeficientes de los monomios m_2 y m_3 sean ± 1 y se cancelen entre ellos. En otras palabras, *cuando la operación arimética (3.23) tiene dos operandos no nulos podemos asignar cero al resultado sin realizar la operación aritmética.* Así evitaremos la cancelación que se hubiera podido producir.

- **Caso** 2**:** *El pivote elegido E es 2×2 con elemento no diagonal a_{pq}, $p < q$.*

  Por el Lema 3.3.3, las entradas en las dos columnas de L son

$$l_{ip} = (CE^{-1})_{ip} = \frac{-a_{ip}a_{qq}}{a_{pq}^2} + \frac{a_{iq}a_{pq}}{a_{pq}^2}\ ; \tag{3.24}$$

$$l_{iq} = (CE^{-1})_{iq} = \frac{a_{ip}a_{pq}}{a_{pq}^2} - \frac{a_{iq}a_{pp}}{a_{pq}^2}\ . \tag{3.25}$$

  De nuevo, ambas expresiones pueden escribirse de la forma $m_1 = m_2 + m_3$, donde cada operando m_i es un monomio con coeficientes $0, 1$ ó -1 en tres variables: d_i y los recíprocos de d_p y d_q. El mismo argumento que en el caso 1 puede emplearse, esto es, podemos evitar cualquier sustracción asignando $m_1 = 0$ cuando los operandos m_2 y m_3 son no nulos.

  Algo similar ocurre con los elementos del factor D: los elementos del complemento de Schur de una matriz arbitraria son de la forma

$$(B - CE^{-1}C^T)_{ij} =$$

$$= a_{ij} - \frac{a_{ip}a_{qq}a_{pj} - a_{iq}a_{pq}a_{pj} - a_{ip}a_{pq}a_{qj} + a_{iq}a_{pp}a_{qj}}{a_{pp}a_{qq} - a_{pq}^2}, \tag{3.26}$$

pero en nuestro caso, según el Lema 3.3.3, se tiene $a_{pp}a_{qq} = 0$. Reemplazando este resultado en (3.26) obtenemos que las entradas del complemento de Schur son una suma

$$m_1 = m_2 + m_3 + m_4 + m_5 \tag{3.27}$$

de cuatro operandos, cada uno de los cuales es un monomio en las variables d_i, d_j con coeficientes ± 1 ó 0. Por tanto m_1 es también un monomio en las mismas variables y con coeficiente ± 1 ó 0 por ser un elemento de un complemento de Schur de una matriz DSTU. Si todos los operandos son cero, o sólo uno de ellos es no nulo, entonces no se realiza ninguna operación. Si tenemos exactamente dos o cuatro operandos no nulos en (3.27), el mismo razonamiento empleado en el caso 1 nos dice que el valor de m_1 debe ser cero, ya que la única posibilidad de que dos o cuatro operandos ± 1 sumen una cantidad perteneciente a $\{1, -1, 0\}$ es que sumen cero. Finalmente, en el caso de que tengamos exactamente tres operandos no nulos ocurrirá que necesariamente dos de ellos deberán cancelarse y el resultado de m_1 será igual al tercer operando. Por tanto, si los tres operandos no nulos son m_r, m_s, m_t , podemos asignar

$$m_1 = -|m_t|\text{sign}(m_r m_s m_t)$$

donde m_t es cualquiera de los tres operandos, ya que los tres tienen el mismo valor absoluto $|d_i d_j|$.

El análisis anterior sugiere el siguiente algoritmo de coste $O(n^3)$.

Algoritmo 1: LDL^T

Input: A $n \times n$

Output: L D

1×1 2×2 P

$PAP^T = LDL^T$

```
1. for i = 1 to n
2.  elegir pivote de acuerdo con la estrategia de pivote de Bunch–Parlett
3.     if pivote 1 × 1 a_ii
4.          D_ii = a_ii
5.          for j = i + 1 to n
6.               l_ji = a_ji/a_ii
7.          endfor
8.          for j = i + 1 to n
9.               for k = i + 1 to n
10.                   a_jk = a_jk − a_ji a_ik / D_ii
```

11. (*)Si la resta tiene dos operandos no nulos, asignamos el valor $a_{jk} = 0$
12. **endfor**
13. **endfor**
14. **else** pivote 2×2 , $\begin{pmatrix} a_{ii} & a_{i,i+1} \\ a_{i+1,i} & a_{i+1,i+1} \end{pmatrix}$
15. $D_{ii} = a_{ii}, D_{i,i+1} = D_{i+1,i} = a_{i,i+1}, D_{i+1,i+1} = a_{i+1,i+1}$
16. **for** $j = i + 1$ **to** n
17. $l_{ji} = \dfrac{a_{j,i+1}a_{i,i+1}}{a^2_{i,i+1}} - \dfrac{a_{ji}a_{i+1,i+1}}{a^2_{i,i+1}}$
18. (*)Si la resta tiene dos operandos no nulos, asignamos el valor $l_{ji} = 0$
19. **endfor**
20. **for** $j = i + 2$ **to** n
21. $l_{j,i+1} = \dfrac{a_{j,i}a_{i,i+1}}{a^2_{i,i+1}} - \dfrac{a_{j,i+1}a_{ii}}{a^2_{i,i+1}}$
22. (*)Si la resta tiene dos operandos no nulos, asignamos el valor $l_{j,i+1} = 0$
23. **endfor**
24. **for** $j = i + 1$ **to** n
25. **for** $k = i + 1$ **to** n
26. $m_2 = a_{jk}, m_3 = -\dfrac{a_{jq}a_{pq}a_{pk}}{a^2_{pq}}, m_4 = -\dfrac{a_{jp}a_{pq}a_{qk}}{a^2_{pq}}$
27. **if** $a_{pp} = 0$
28. $m_5 = \dfrac{a_{jp}a_{qq}a_{pk}}{a^2_{pq}}$
29. **else** $m_5 = \dfrac{a_{jq}a_{pp}a_{qk}}{a^2_{pq}}$
30. **endif**
31. $a_{jk} = m_2 + m_3 + m_4 + m_5$
32. (*) Si la suma tiene **dos** o **cuatro** operandos no nulos, asignamos el valor $a_{jk} = 0$
33. (*) Si la suma tiene **tres** operandos no nulos m_r, m_s, m_t,
34. asignamos el valor $a_{jk} = -|m_r|\text{sign}(m_r m_s m_t)$
35. **endfor**
36. **endfor**
37. **endif**
38. **endfor**

Además, el análisis llevado a cabo anteriormente sirve como demostración del siguiente resultado.

Teorema 3.3.4 *El Algoritmo 1 calcula todas las entradas de los factores L y D de la factorización LDL^T por bloques de una matriz simétrica DSTU con alta precisión relativa, es decir,*

$$|\widehat{l}_{ij} - l_{ij}| = O(\varepsilon)|l_{ij}|, \quad |\widehat{d}_{ij} - d_{ij}| = O(\varepsilon)|d_{ij}|, \; \forall\, i, j \in \{1, \ldots, n\}$$

donde $\widehat{L}$ y $\widehat{D}$ son los factores calculados en aritmética en coma flotante por el Algoritmo 1, y L, D son los factores exactos que el método de pivote diagonal calcularía en aritmética exacta eligiendo los pivotes con las mismas dimensiones y posiciones que los elegidos en aritmética en coma flotante para calcular $\widehat{L}$ y $\widehat{D}$.

Cualquier intento de estimar teóricamente las constantes dentro de la $O(\varepsilon)$ en el Teorema 3.3.4 producirá cotas pesimistas. Sea p_k el número de operaciones en coma flotante necesarias para calcular los elementos de D en la etapa k-ésima de la factorización por bloques LDL^T. Esta cantidad viene dada por la fórmula recursiva $p_{k+1} = 2p_k + 1$, por lo que $p_k = 2^{k+1} - 1$. Por ello, la constante en la $O(\varepsilon)$ obtenida por un análisis elemental de errores es exponencial. Sin embargo, esto no se ha observado nunca en la práctica.

Una nueva estrategia de pivotaje

Un característica adicional de eliminación gaussiana con pivote completo sobre matrices no simétricas DSTU de $n \times n$ es que el número de condición de los factores L y U crece cuadráticamente con la dimensión n [22, Teorema 10.2]. Esto es importante, pues garantiza que los factores están bien condicionados y, por ello, la factorización obtenida es una RRD. Con el fin de probar lo mismo para el factor triangular L en la descomposición LDL^T por bloques, debemos hacer un ligero cambio en la estrategia de pivotaje. Se considera la siguiente estrategia,

$$\begin{array}{ll} \textbf{if} & \mu_0 = \max_{i,j}|a_{ij}| = \max_i|a_{ii}| = \mu_1 \\ & \quad \text{elegir pivote } 1 \times 1 \\ \textbf{else} & \\ & \quad \text{elegir pivote } 2 \times 2 \end{array} \tag{3.28}$$

Con esta estrategia de pivotaje, la entradas de L están trivialmente acotadas por 1 en valor absoluto, mientras que lo mejor que se puede decir utilizando la estrategia de Bunch–Parlett es que $|l_{ij}| \leq 1/\alpha \approx 1'6$ (para los elementos generados por pivotes de tamaño 1×1). Además, el número de condición en norma 2 de los pivotes está acotado por $4'6$ para Bunch–Parlett, y por $2'6$ para esta nueva estrategia. Nótese que el cambio en el valor de α no afecta a la validez de los resultados en §3.3.1, ya que el Lema 3.3.3 es cierto para $\alpha \leq 1$.

Veamos que con esta estrategia modificada de pivote, el número de condición del factor L de una descomposición por bloques LDL^T (3.6) crece cuadráticamente con la dimensión de la matriz factorizada.

Teorema 3.3.5 *Sea A una matriz simétrica DSTU. Existe una matriz B, también DSTU, cuyo factor triangular inferior calculado por eliminación gaussiana con pivote completo coincide con el factor triangular de la descomposición por bloques LDL^T de A obtenida usando la estrategia de pivotaje (3.28). Por tanto, el número de condición del factor L de la factorización LDL^T por bloques crece cuadráticamente con la dimensión de la matriz A.*

Demostración:

Sin pérdida de generalidad, podemos restringirnos a comparar los dos primeros pasos de eliminación gaussiana con pivote completo con los correspondientes pasos del método de pivote diagonal. Si el primer pivote elegido por el método de pivote diagonal es de tamaño 1×1, se aplican a la matriz A las mismas permutaciones que se harían con eliminación gaussiana, y la primera columna de ambos factores triangulares trivialmente coincide. Por otro lado, como el pivote elegido en este caso está en la diagonal, la matriz se permuta simétricamente, y los complementos calculados por ambos métodos son iguales.

Ahora, supongamos que el primer pivote elegido por el método de pivote diagonal es de tamaño 2×2, por ejemplo,

$$\begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & 0 \end{pmatrix},$$

donde hemos supuesto, a la vista del Lema 3.3.3, que $a_{qq} = 0$. El caso $a_{pp} = 0$ es completamente análogo. Denotando por P_{ij} a la matriz de permutación que intercambia la fila i-ésima con la j-ésima, el método de pivotaje diagonal permuta la matriz A de la forma $P_{2q}P_{1p}\ A\ P_{1p}P_{2q}$ para colocar el pivote 2×2 en la esquina superior izquierda de la matriz. Eliminación gaussiana, en cambio, permuta A de la forma $P_{1p}AP_{1q}$ para colocar el pivote a_{pq} en la parte superior izquierda de la matriz. Va a ser conveniente aplicar una permutación de filas adicional, P_{2q}, que coloque la entrada nula a_{qq} en la posición $(2,1)$. Finalmente, para colocar el elemento $a_{qp} = a_{pq}$ en la posición $(2,2)$ realizamos los cambios por columnas P_{2q} y P_{2p}. De esta manera, si renombramos

$$M_1 = P_{2q}P_{1p}\ A\ P_{1p}P_{2q}, \qquad M_2 = P_{2q}P_{1p}AP_{1q}P_{2q}P_{2p},$$

resultan dos matrices idénticas excepto por las dos primeras columnas, que están intercambiadas. Si denotamos por m_{ij} al elemento (i,j) de la matriz simétrica M_1 entonces $m_{22} = 0$, y m_{12} es la entrada de M_1 con mayor valor absoluto de la matriz. Así el método de pivote diagonal calcula las entradas de las dos primeras columnas de su respectivo factor L como

$$\begin{aligned} l_{i1} &= \frac{m_{i2}}{m_{12}}, \quad i = 3, \ldots, n, \\ l_{i2} &= \frac{m_{i1}m_{12} - m_{i2}m_{11}}{m_{12}^2}, \quad i = 3, \ldots, n \end{aligned}$$

y las entradas del complemento de Schur de tamaño $(n-2) \times (n-2)$ Schur como

$$(B - CE^{-1}C^T)_{ij} = m_{ij} + \frac{-m_{i2}m_{12}m_{1j} - m_{i1}m_{12}m_{2j} + m_{i2}m_{11}m_{2j}}{m_{12}^2}, \tag{3.29}$$

para $i, j = 3, \ldots, n$. Veamos que las dos primeras etapas de eliminación Gaussiana con pivote completo sobre M_2 producen exactamente las dos mismas columnas de L y el mismo complemento de Schur de tamaño $(n-2) \times (n-2)$. El primer paso de eliminación gaussiana sobre M_2 produce una primera columna con un cero en la primera posición, y $l_{i1} = m_{i2}/m_{12},\ i = 3, \ldots, n$ como antes. El complemento de Schur $(n-1) \times (n-1)$ resultante de esta primera etapa tiene la forma

$$\left(\begin{array}{c|ccc} m_{12} & \cdots & m_{2j} & \cdots \\ \hline \vdots & \vdots & \vdots & \vdots \\ m_{i1} - \dfrac{m_{i2}m_{11}}{m_{12}} & \vdots & m_{ij} - \dfrac{m_{i2}m_{1j}}{m_{12}} & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{array}\right) \tag{3.30}$$

En la segunda etapa, eliminación gaussiana busca la entrada con mayor valor absoluto en esta matriz. Veamos que esta entrada es de nuevo m_{12}; en caso contrario, debería haber un nuevo elemento más grande que m_{12} en valor absoluto. Cualquier entrada del complemento de Schur (3.30) desaparece si es el resultado de una resta con operandos no nulos, luego la única posibilidad de elementos nuevos son cocientes de la forma $-(m_{i2}m_{1j})/(m_{12})$. Estos cocientes son también monomios con coeficientes ± 1 en dos variables. Estas dos variables son los elementos diagonales de $\mathcal{D}$ correspondientes a los índices i, j antes de que la matriz A se permutase para obtener M_2. En cualquier caso, como ambos i, j son diferentes de 1 y 2, los $\widetilde{d}_i$ y $\widetilde{d}_j$ correspondientes, son diferentes de d_p y d_q. En consecuencia, el valor absoluto de cualquier nuevo elemento aparecido en (3.30) es estrictamente menor que el máximo $|m_{12}| = |d_p d_q|$.

Luego en esta segunda etapa de eliminación gaussiana de la matriz M_2 no hace falta permutar. La segunda columna del factor L correspondiente tiene entradas

$$l_{i2} = \frac{m_{i1} - \dfrac{m_{11}m_{i2}}{m_{12}}}{m_{12}} = \frac{m_{i1}m_{12} - m_{11}m_{i2}}{m_{12}^2}$$

que coinciden con las entradas l_{i2} anteriores reeemplazando i por j (recuérdese que $m_{ij} = m_{ji}$). Finalmente, el complemento de Schur de tamaño $(n-2) \times (n-2)$ calculado por eliminación gaussiana en este segundo paso tiene las entradas

$$\left(m_{ij} - \frac{m_{i2}m_{1j}}{m_{12}}\right) - \frac{\left(m_{1i} - \dfrac{m_{11}m_{i2}}{m_{12}}\right) m_{2j}}{m_{12}}\ .$$

Un simple cálculo demuestra que esta fórmula coincide con (3.29). Esto prueba que, como se demostró, dos pasos de eliminación gaussiana sobre M_2 producen las dos mismas columnas para L que un paso del método de pivote diagonal con un pivote de tamaño 2×2 sobre la matriz M_1. Además el complemento de Schur $(n-2) \times (n-2)$ resultante en ambos casos es el mismo.

Repitiendo el argumento para los pasos subsiguientes de la descomposición por bloques, obtenemos que el factor L de una matriz A simétrica DSTU, es igual al factor triangular inferior de una factorización LDU de una matriz no simétrica $B = AQ$, para una matriz de permutación Q apropiada. Nótese que la matriz B es DSTU si A es DSTU, ya que

$$B = D_L \widetilde{Z} D_R,$$

con $D_L = D$, $D_R = Q^T D Q$, $\widetilde{Z} = ZQ$, y esta última es TU. Por tanto el número de condición del factor L crece cuadráticamente con la dimensión n de la matriz original A por el Teorema 10.2 de [22].

□

La demostración del Teorema 3.3.5 se basa indirectamente en [22, Teorema 10.2]. Una demostración directa análoga a la de [22, Teorema 10.2] no es posible ya que falta el ingrediente principal: que los elementos de cualquier menor del factor L sean cocientes de menores de la matriz original A y, por lo tanto, los elementos de la matriz L^{-1} sean cocientes de menores de A. Y esto no es cierto para la factorización LDL^T por bloques como se señaló con el ejemplo 3×3 dado en §2.

Cálculo de menores de una matriz DSTU con alta precisión relativa

Se sabe por el Lema 3.3.2 que cualquier menor de una matriz simétrica DSTU es un monomio con coeficientes $0, 1$ o -1 en los elementos diagonales d_i de la matriz $\mathcal{D}$. En el siguiente lema damos una fórmula explícita de estos coeficientes para una matriz DSTU cualquiera.

Lema 3.3.6 *Sea $A \in \mathbb{R}^{m \times n}$ una matriz DSTU cualquiera, esto es, $A = \mathcal{D}_L Z \mathcal{D}_R$, donde Z es totalmente unimodular y $\mathcal{D}_L = diag(l_1, \ldots, l_m)$, $\mathcal{D}_R = diag(r_1, \ldots, r_n)$ son matrices diagonales. Entonces*

$$det(A([i_1, \ldots, i_s], [j_1, \ldots, j_s])) = l_{i_1} \ldots l_{i_s} r_{j_1} \ldots r_{j_s} K \tag{3.31}$$

donde $K = det(Z([i_1, \ldots, i_s], [j_1, \ldots, j_s]))$.

Un problema del Álgebra Lineal Numérica es calcular el determinante de una matriz con alta precisión relativa. En el caso de las matrices DSTU esto es posible en los siguientes casos:

1. Sea A una matriz DSTU tal que se conocen los factores $A = \mathcal{D}_L Z \mathcal{D}_R$. Las entradas de Z se suponen conocidas exactamente, y las entradas de $\mathcal{D}_L$ y $\mathcal{D}_R$ conocidas con alta precisión relativa en la aritmética empleada (esto es, con un error relativo del orden del épsilon-máquina). Entonces se puede calcular cualquier menor de A con alta precisión relativa utilizando la fórmula (3.31).

2. Sea A una matriz *simétrica* DSTU. Aplicando el Algoritmo 1 obtenemos que $PAP^T = LDL^T$ con $D = diag(D_1, \ldots, D_r)$, donde las matrices D_i para $i \in \{1, \ldots, r\}$ son

matrices de tamaño 1×1 ó 2×2, cuyos elementos están calculados con alta precisión relativa por el Teorema 3.3.4. De este modo

$$det(A) = \prod_{i=1}^{r} det(D_i).$$

Este producto se realiza con alta precisión relativa: es producto de cantidades calculadas con alta precisión relativa ya que las entradas de los pivotes de tamaño 1×1 están calculadas con dicha precisión por el Algoritmo 1 y los determinantes de los pivotes de tamaño 2×2, por ejemplo, $D_i = \begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{pp} \end{pmatrix}$, cumplen que $det(D_i) = -a_{pq}^2$ gracias al Lema 3.3.3.

3.3.2. Matrices TSC

Definición 3.3.7 *Sea $\mathcal{S}$ un conjunto de matrices con un patrón de signos dado, es decir, todas las matrices pertenecientes a $\mathcal{S}$ tienen sus entradas no nulas en las mismas posiciones y con el mismo signo. Se dice que el conjunto $\mathcal{S}$ es TSC (*total signed compound*) si para toda matriz $A \in \mathcal{S}$ y para toda submatriz cuadrada M de A, la expresión Laplaciana de su determinante*

$$\det M = \sum_{\pi} [\text{sign}(\pi)\, m_{1,\pi(1)} m_{2,\pi(2)} \dots m_{s,\pi(s)}] \tag{3.32}$$

es o bien suma de monomios del mismo signo, con por lo menos un monomio no nulo, o bien idénticamente cero (es decir, todos los monomios en la expresión son nulos).

Hay patrones bien conocidos entre las matrices TSC, como el patrón tridiagonal

$$\begin{pmatrix} + & + & & & \\ & + & - & + & \\ & + & + & + & \\ & & + & - & + \\ & & & + & + \end{pmatrix}$$

o el patrón en punta de flecha

$$\begin{pmatrix} + & + & + & + & + \\ + & - & & & \\ + & & - & & \\ + & & & - & \\ + & & & & - \end{pmatrix}.$$

Las matrices TSC son además matrices *huecas* (hay a lo sumo $3n-2$ entradas no nulas en una matriz TSC de $n \times n$ (véase [21, Lema 7.1]). Esta propiedad se observa a simple vista en la Figura 3.1, en la que se muestran dos matices TSC generadas aleatoriamente tal como se describe en la Sección 3.5. Un cuadrado gris representa una entrada nula de la

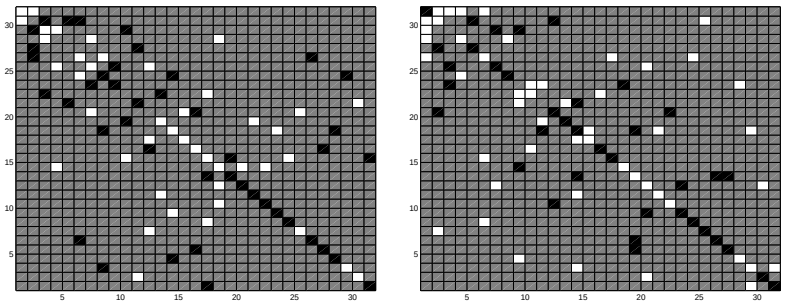


Figura 3.1: Ejemplos de matrices simétricas TSC

matriz, un cuadrado blanco una entrada positiva y un cuadrado negro una entrada negativa de una matriz simétrica TSC.

El hecho de tener tantos elementos nulos permite calcular determinantes de matrices TSC de tamaño n mediante algoritmos de coste $O(n)$. Además, tales algoritmos calculan el determinante con alta precisión relativa, debido a la propiedad que define las matrices TSC, ya que en su cálculo no puede haber cancelación al ser suma de monomios del mismo signo. El coste de $O(n)$ se alcanza haciendo uso de una definición alternativa de las matrices TSC: cada matriz TSC puede construirse empezando con una matriz no nula de tamaño 1×1 y repitiendo una o varias de entre cuatro reglas prefijadas de construcción (véanse [11, 22] para más detalles). Si nos restringimos al conjunto de matrices TSC *simétricas*, se puede probar que sólo se necesitan tres reglas de construcción para generar cualquier matriz TSC. El siguiente resultado es una versión simétrica del Lema 7.2 de [22].

Teorema 3.3.8 *Una matriz simétrica TSC se puede generar partiendo de una matriz no nula de tamaño* 1×1 *y aplicando repetidamente algunas de las siguientes reglas, repetidas en cualquier orden:*

1. *Si A es una matriz TSC simétrica, entonces la permutación de dos filas y las correspondientes columnas, o multiplicar por -1 una fila y su correspondiente columna, no cambia el carácter TSC simétrico.*

2. *Si A_1 y A_2 son ambas matrices simétricas TSC, entonces también lo es su suma directa*
$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

3. *Si una matriz $\widetilde{A}$ de tamaño $n \times n$ es simétrica TSC, con $\tilde{a}_{ii} \neq 0$, entonces también*

es TSC simétrica la matriz A siguiente de tamaño $(n+1)\times(n+1)$

$$A=\left(\begin{array}{ccccccc|c} & & & & & & & 0\\ & & & & & & & \vdots\\ & & & & & & & 0\\ & & & \widetilde{A} & & & & a_{i,n+1}\\ & & & & & & & 0\\ & & & & & & & \vdots\\ & & & & & & & 0\\ \hline 0 & \dots & 0 & a_{i,n+1} & 0 & \dots & 0 & a_{n+1,n+1}\end{array}\right)$$

donde se puede asignar cero a la entrada $\widetilde{a}_{ii}$. Las nuevas entradas no nulas $a_{i,n+1}$ y $a_{n+1,n+1}$ deben elegirse de tal modo que los dos monomios que aparecen en el menor $a_{n+1,n+1}\widetilde{a}_{i,i}-a_{i,n+1}a_{n+1,i}$ tengan el mismo signo (o bien sean ambos nulos).

Estas reglas nos permitirán generar matrices TSC en § 3.5 a un coste computacional muy bajo.

Factorización LDL^T por bloques con alta precisión relativa para matrices simétricas TSC

El interés por las matrices simétricas TSC proviene del hecho que todos sus menores pueden calcularse sin sustracciones, y por lo tanto sin efectos de cancelación. De acuerdo con el Teorema 3.2.4, cualquier cantidad intermedia en la descomposición LDL^T por bloques de una matriz es un cociente de menores o simplemente un menor de la matriz original. Aunque las fórmulas estándar del método de pivote diagonal podrían requerir alguna sustracción, y por tanto dar lugar a una posible cancelación, las restas pueden evitarse si el correspondiente elemento de L o del complemento de Schur se recalcula como cociente de menores de la matriz original. Estas fórmulas aparecen el Lema 3.2.3 para los elementos del complemento de Schur , y en el Lema 3.2.5 para los elementos de L. Por otro lado, la ausencia de cancelación implica que estos menores se calculan con alta precisión relativa.

A continuación se escribe en pseudocódigo el correspondiente algoritmo. Por supuesto la operación de recalcular supone un coste adicional. Un menor de tamaño $s\times s$ supone $O(s)$ operaciones, y como cualquier submatriz de una matriz TSC es también TSC, la versión modificada del método de pivotaje diagonal puede costar en el peor de los casos $O(n^4)$ operaciones aritméticas, el mismo orden que el algoritmo propuesto en [22] (véase [22, Teorema 7.2, p. 60]) para calcular RRD *no simétricas* de matrices TSC. Sobre esta cuestión, véase también el experimento al final de la sección 3.5.2.

Algoritmo 2: LDL^T

Input: A $n\times n$

Output: L D
1×1 2×2 P
$PAP^T = LDL^T$

1. **for** $i = 1$ **to** n
2. elegir pivote de acuerdo con la estrategia de pivote de Bunch–Parlett
3. **if** pivote 1×1 , a_{ii}
4. $D_{ii} = a_{ii}$
5. **for** $j = i + 1$ **to** n
6. $l_{ji} = a_{ji}/a_{ii}$
7. **endfor**
8. **for** $j = i + 1$ **to** n
9. **for** $k = i + 1$ **to** n
10. $a_{jk} = a_{jk} - \dfrac{a_{ji}a_{ik}}{D_{ii}}$
11. (*) Si la resta tiene dos operandos no nulos del mismo signo, recalcular a_{jk} como cociente de dos menores de A según la fórmula 3.16 en Lema 3.2.3
12. **endfor**
13. **endfor**
14. **else** pivote 2×2 , $\begin{pmatrix} a_{ii} & a_{i,i+1} \\ a_{i+1,i} & a_{i+1,i+1} \end{pmatrix}$
15. $D_{ii} = a_{ii}, D_{i,i+1} = D_{i+1,i} = a_{i,i+1}, D_{i+1,i+1} = a_{i+1,i+1}$
16. **for** $j = i + 1$ **to** n
17. $dpiv = a_{ii}a_{i+1,i+1} - a_{i,i+1}^2$
18. (*)Si la resta tiene dos operandos no nulos del mismo signo recalcular $dpiv$ como cociente de dos menores de A según la fórmula 3.17 en Lema 3.2.3
19. $l_{ji} = \dfrac{a_{ji}a_{i+1,i+1}}{dpiv} - \dfrac{a_{j,i+1}a_{i,i+1}}{dpiv}$
20. (*) Si la resta tiene dos operandos no nulos del mismo signo, recalcular l_{ji} como cociente de dos menores de A según la fórmula 3.18 en Lema 3.2.5
21. **endfor**
22. **for** $j = i + 2$ **to** n
23. $l_{j,i+1} = -\dfrac{a_{j,i+1}a_{i,i+1}}{dpiv} + \dfrac{a_{j,i+1}a_{ii}}{dpiv}$
24. (*) Si la resta tiene dos operandos no nulos del mismo signo, recalcular $l_{j,i+1}$ como cociente de dos menores de A según la fórmula 3.18 en Lema 3.2.5
25. **endif**
26. **for** $j = i + 1$ **to** n
27. **for** $k = i + 1$ **to** n

28. $$a_{jk} = a_{jk} - \frac{a_{jp}a_{qq}a_{pk}}{dpiv} - \frac{a_{jq}a_{pq}a_{pk}}{dpiv} - \frac{a_{jp}a_{pq}a_{qk}}{dpiv} + \frac{a_{jq}a_{pp}a_{qk}}{dpiv}$$
29. (*) Si la resta tiene dos operandos no nulos del signo, recalcular a_{jk} como cociente de dos menores de A según la fórmula 3.16 en Lema 3.2.3
30. **endfor**
31. **endfor**
32. **endif**
33. **endfor**

El argumento anterior demuestra la alta precisión relativa con que se calculan, elemento a elemento, los factores de la descomposición LDL^T de una matriz TSC. Esto se refleja en el siguiente teorema.

Teorema 3.3.9 *El Algoritmo 2 calcula todas las entradas de los factores L y D de una descomposición LDL^T por bloques de una matriz simétrica TSC con alta precisión relativa, esto es,*

$$|\widehat{l}_{ij} - l_{ij}| = O(\varepsilon)|l_{ij}|, \quad |\widehat{d}_{ij} - d_{ij}| = O(\varepsilon)|d_{ij}|,$$

donde $\widehat{L}$ y $\widehat{D}$ son los factores calculados en aritmética en coma flotante por el Algoritmo 1, y L, D son los factores exactos que el método de pivote diagonal calcularía en aritmética exacta eligiendo los pivotes con las mismas dimensiones y posiciones que en aritmética en coma flotante para calcular $\widehat{L}$ y $\widehat{D}$.

3.4. Paso de LDL^T a RRD: Análisis de errores

Una vez calculada la factorización por bloques LDL^T, ya vimos en (3.8) cómo obtener una descomposición simétrica RRD mediante diagonalización por rotaciones de Givens. Es fácil demostrar que como L está calculado con un error relativo pequeño componente a componente y $X = P^T LQ$ es, salvo permutaciones, el resultado de una transformación de Givens en coma flotante (véase, por ejemplo, [20, Lema 3.1]), el factor calculado X cumplirá (3.10). De hecho, obtendremos cotas de error más finas en el Teorema 3.4.1, donde se muestra que el factor X se calcula con un error relativo pequeño *columna a columna*. Observemos, además, que X y L tienen el mismo número de condición, luego si el factor L está bien condicionado, también lo estará el factor X, garantizando una de las propiedades esenciales de una descomposición RRD (esto queda probado, por el momento, para las matrices DSTU utilizando la nueva estrategia de pivotaje y el Teorema 3.3.5).

A continuación presentamos el análisis de errores que demuestra que la descomposición LDL^T por bloques dada por los Algoritmos 1 ó 2, seguida por una diagonalización de Givens, conduce a una descomposición RRD que cumple (3.10), requisito esencial para poder asegurar alta precisión relativa en el cálculo de los autovalores y autovectores de las matrices simétricas DSTU y TSC mediante el algoritmo SVD con signos. No se hará distinción entre matrices DSTU y TSC, ya que el análisis de errores es válido para cualquier

matriz de la que se pueda calcular una descomposición LDL^T por bloques con error relativo pequeño componente a componente como en (3.9). Este análisis de errores es muy similar al llevado a cabo en [29]. De hecho, se tomará de [29] algunos resultados necesarios para el análisis. Para ser más precisos, introduciremos la siguiente notación; partimos del modelo convencional de aritmética en coma flotante,

$$\mathtt{fl}(a \odot b) = (a \odot b)(1+\delta), \tag{3.33}$$

donde a y b son números reales en coma flotante, $\odot \in \{+,-,\times,/\}$, y $|\delta| \leq \varepsilon$, donde ε es la precisión de la máquina. Por otro lado, se supone que no se produce *overflow* ni *underflow*. Para cada $k > 0$ llamamos

$$\gamma_k = \frac{k\varepsilon}{1-k\varepsilon} \tag{3.34}$$

y, como en [43, § 3.4], denotamos por θ_k cualquier cantidad positiva acotada por γ_k. Finalmente, dada una matriz real y simétrica de tamaño 2×2, el método de diagonalización de Jacobi se escribe como

$$\begin{pmatrix} a & c \\ c & b \end{pmatrix} = \begin{pmatrix} cs & sn \\ -sn & cs \end{pmatrix} \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix} \begin{pmatrix} cs & -sn \\ sn & cs \end{pmatrix}, \tag{3.35}$$

con $\lambda_1 = a - ct,\ \lambda_2 = b + ct$, donde

$$t = \frac{\text{signo}(\zeta)}{|\zeta| + \sqrt{1+\zeta^2}} \qquad \text{para} \qquad \zeta = \frac{b-a}{2c} \tag{3.36}$$

y

$$cs = \frac{1}{\sqrt{1+t^2}}, \qquad sn = cs \cdot t. \tag{3.37}$$

El principal resultado de esta sección, escrito con esta notación, es el siguente.

Teorema 3.4.1 *Sea $A \in \mathbb{R}^{n\times n}$ una matriz simétrica y sean $\widehat{L}, \widehat{D}$ los factores calculados de una factorización LDL^T por bloques de A, obtenida a través del método de pivotaje diagonal usando la estrategia de pivotaje de Bunch–Parlett (3.14) (es decir, con $\alpha = (1+\sqrt{17})/8$). Supongamos que $\widehat{L}, \widehat{D}$ se han calculado con error relativo pequeño componente a componente*

$$\begin{aligned} \widehat{l}_{ij} &= l_{ij}(1+\theta^{(ij)}_{K_L}), \qquad i,j = 1,\ldots,n \\ \widehat{d}_{ij} &= l_{ij}(1+\theta^{(ij)}_{K_D}), \qquad i,j = 1,\ldots,n \end{aligned} \tag{3.38}$$

para ciertas constantes $K_L, K_D > 0$, y sean $\widehat{X}, \widehat{\Delta}$ (respectivamente X, Δ) los factores calculados (respectivamente exactos) de una descomposición RRD simétrica obtenida por diagonalización mediante rotaciones de Givens (3.8) en aritmética finita (respectivamente en aritmética exacta), usando las fórmulas (3.35)–(3.37). Entonces,

$$\frac{|\widehat{\Delta}_{jj} - \Delta_{jj}|}{|\Delta_{jj}|} \leq 4\,\frac{1+\alpha}{1-\alpha}\,\gamma_{K_D+29}, \qquad j = 1,\ldots,n \tag{3.39}$$

y

$$\frac{\|\widehat{X}(:,j) - X(:,j)\|_2}{\|X(:,j)\|_2} \leq \sqrt{2nC^2 + 1}\ \gamma_M, \qquad j = 1, \ldots, n \tag{3.40}$$

donde C *es*

$$C = \frac{1}{1-\alpha} + O(\varepsilon).$$

y

$$M = \max\{48K_{_L} + 141, K_{_L} + 48K_{_D} + 143\}. \tag{3.41}$$

Para demostrar este resultado utilizaremos el siguiente lema auxiliar extraído de [29, Appendix A.3].

Lema 3.4.2 (Dopico & Koev [29]) *Sea*

$$\widetilde{A} = \begin{pmatrix} \widetilde{a} & \widetilde{c} \\ \widetilde{c} & \widetilde{b} \end{pmatrix} = \begin{pmatrix} a(1+\delta_a) & c(1+\delta_c) \\ c(1+\delta_c) & b(1+\delta_b) \end{pmatrix}$$

ua matriz de números reales en coma flotante, con $\max\{|\delta_a|, |\delta_b|, |\delta_c|\} \leq \gamma_k$ *y* $\alpha|\widetilde{c}| \geq \max\{|\widetilde{a}|, |\widetilde{b}|\}$. *Sea*

$$A = \begin{pmatrix} a & c \\ c & b \end{pmatrix},$$

con autovalores $\lambda_1 \geq \lambda_2$, *y autovectores ortonormales correspondientes* $v_1 = [cs, -sn]^T$ *y* $v_2 = [sn, cs]$. *Sean* $\widehat{\lambda}_1, \widehat{\lambda}_2, \widehat{cs}$ *y* $\widehat{sn}$ *las versiones de* λ_1, λ_2, cs *y* sn *calculadas en aritmética en coma flotante para* $\widetilde{A}$ *según las fórmulas (3.35)–(3.37). Si*

$$4\sqrt{2}\,\frac{1+\alpha}{1-\alpha}\gamma_{k+29} \leq 1, \qquad \text{y} \qquad \gamma_{141+48k} \leq 1,$$

entonces

$$\frac{|\widehat{\lambda}_i - \lambda_i|}{|\lambda_i|} \leq 4\,\frac{1+\alpha}{1-\alpha}\gamma_{k+29}, \quad i = 1, 2 \tag{3.42}$$

y

$$\widehat{cs} = cs(1 + \theta_{16k+113}), \qquad \widehat{sn} = cs(1 + \theta_{48k+141}). \tag{3.43}$$

Demostración: (del Teorema 3.4.1)

En primer lugar, se supone sin pérdida de generalidad que $P = I$, ya que las permutaciones no introducen error alguno. Sea $\widehat{Q}$ la matriz ortogonal calculada que diagonaliza $\widehat{D}$, esto es, si $\widehat{D} = diag(\widehat{D}_1, \ldots, \widehat{D}_r)$ con $\widehat{D}_k \in \mathbb{R}^{s_k \times s_k}$, $s_k = 1$ ó 2, $k = 1, \ldots, r$, entonces $\widehat{Q} = diag(\widehat{Q}_1, \ldots, \widehat{Q}_r)$ con $\widehat{Q}_k \in \mathbb{R}^{s_k \times s_k}$, $k = 1, \ldots, r$. Los bloques de tamaño 1×1 de $\widehat{Q}_k$ son iguales a 1, y cada bloque de tamaño 2×2

$$\widehat{Q}_k = \begin{pmatrix} \widehat{cs} & -\widehat{sn} \\ \widehat{sn} & \widehat{cs} \end{pmatrix} \tag{3.44}$$

es la versión calculada en aritmética finita de la rotación de Jacobi que diagonalizaría en aritmética exacta el bloque $\widehat{D}_k$ de tamaño 2×2. Análogamente, $Q = \mathrm{diag}(\mathrm{Q}_1, \ldots, \mathrm{Q}_\mathrm{r})$, donde

$$Q_k = \begin{pmatrix} cs & -sn \\ sn & cs \end{pmatrix}$$

es la rotación de Jacobi exacta que diagonaliza el bloque diagonal D_k de D. Para aquellas columnas j correspondientes a un pivote con $s_k = 1$ se tiene que $\widehat{\Delta}_{jj} = \widehat{d}_{jj}$, $\Delta_{jj} = d_{jj}$ y $\widehat{X}(:,j) = \widehat{L}(:,j)$, $X(:,j) = L(:,j)$, por lo que (3.39) y (3.40) se satisfacen trivialmente. Sólo queda considerar el caso de las columnas correspondientes a pivotes de tamaño 2×2. Supongamos que la j-ésima y la $(j+1)$-ésima son dos de tales columnas. Primero, la desigualdad (3.39) se sigue directamente de aplicar el Lema 3.4.2, es decir, tomando $\widetilde{A}$, A, λ_1, λ_2 , y k iguales, respectivamente, a $\widehat{D}$, D, Δ_{jj}, $\Delta_{j+1,j+1}$ y $K_{_D}$. Con esta elección, la desigualdad (3.42) se reduce a (3.39). Para demostrar (3.40), se observa que

$$X = LQ, \qquad\qquad \widehat{X} = \mathtt{fl}(\widehat{L}\widehat{Q}),$$

donde $\mathtt{fl}(expr)$ denota el resultado calculado en precisión finita de la expresión $expr$. Leyendo estas igualdades componente a componente para las columnas j y $j+1$ en cuestión, tenemos

$$\widehat{X}(i,j) = \begin{cases} 0 & , si \quad i < j \ ; \\ \widehat{cs} & , si \quad i = j \ ; \\ \widehat{sn} & , si \quad i = j+1 \ ; \\ \mathtt{fl}(\widehat{l}_{ij}\,\widehat{cs} + \widehat{l}_{i,j+1}\,\widehat{sn}) & , si \quad i > j+1 \ , \end{cases} \tag{3.45}$$

$$\widehat{X}(i,j+1) = \begin{cases} 0 & , si \quad i < j \ ; \\ -\widehat{sn} & , si \quad i = j \ ; \\ \widehat{cs} & , si \quad i = j+1 \ ; \\ \mathtt{fl}(-\widehat{l}_{ij}\widehat{sn} + \widehat{l}_{i,j+1}\widehat{cs}) & , si \quad i > j+1 \ . \end{cases}$$

y las mismas igualdades sin los acentos circunflejos para los elementos de X. Por lo tanto,

$$||\widehat{X}(:,j) - X(:,j)||_2^2 = (\widehat{cs} - cs)^2 +$$

$$+(\widehat{sn} - sn)^2 + \sum_{i=j+2}^{n} [\mathtt{fl}(\widehat{l}_{ij}\widehat{cs} + \widehat{l}_{i,j+1}\widehat{sn}) - (l_{ij}cs + l_{i,j+1}sn)]^2.$$

Usando (3.38), (3.33) y (3.43), se puede escribir

$$\begin{aligned} \mathtt{fl}(\widehat{l}_{ij}\widehat{cs} + \widehat{l}_{i,j+1}\widehat{sn}) &= \Big[l_{ij}\, cs\, (1 + \theta_{K_L}^{(i,j)})\, (1 + \theta_{16K_D+113})\, (1+\delta_1) + \\ &\qquad + l_{i,j+1}\, sn\, (1 + \theta_{K_L}^{(i,j+1)})\, (1 + \theta_{48K_D+141})\, (1+\delta_2) \Big]\, (1+\delta_3) = \\ &= l_{ij}\, cs\, (1 + \theta_{K_L+16K_D+115}) + l_{i,j+1}\, sn\, (1 + \theta_{K_L+48K_D+143}). \end{aligned}$$

Volviendo hacer uso de (3.43), se tiene

$$\begin{aligned}\|\widehat{X}(:,j) - X(:,j)\|_2^2 &= (cs\,\theta_{16K_D+113})^2 + (sn\,\theta_{48K_D+141})^2 + \\ &\quad + \sum_{i=j+2}^{n} \left(l_{ij}\, cs\,\theta_{K_L+16K_D+115} + l_{i,j+1}\, sn\,\theta_{K_L+48K_D+143}\right)^2 \\ &\leq (\gamma_{48K_L+141})^2 + 2n\,(\gamma_{K_L+48K_D+143})^2 \max\{|l_{ij}|^2, |l_{i,j+1}|^2\},\end{aligned}$$

donde se ha usado la monotonía de γ_k en k. Llegados a este punto, observamos que aunque la estrategia de Bunch–Parlett asegura que las entradas del factor *calculado* $\widehat{L}$ cumplen $|\widehat{l}_{ik}| \leq 1/(1-\alpha)$ para todo i,k, esto puede no ser cierto para las entradas l_{ik} del factor *exacto* L. La cota componente a componente (3.38), sin embargo, implica tras llevar a cabo ciertos cálculos que $|l_{ik}| \leq C$ para una constante C igual a $1/(1-\alpha)$ salvo términos de primer orden en ε. Para un análisis más detallado veáse [29], donde se demuestra que

$$C = \frac{1}{(1-\alpha)(1-\gamma_{g(\alpha)})}, \qquad g(\alpha) = \left(32\left(\frac{1+\alpha}{1-\alpha}\right)^2 + 196\,\frac{1+\alpha}{1-\alpha}\right) K_D.$$

Así,

$$\|\widehat{X}(:,j) - X(:,j)\|_2 \leq \sqrt{1+2nC^2}\,\gamma_M$$

con M dado por (3.41), lo cual conduce trivialmente a (3.40), ya que

$$\|X(:,j)\|_2^2 \geq cs^2 + sn^2 \geq 1.$$

□

3.5. Experimentos numéricos

Hemos llevado a cabo pruebas numéricas que confirman la precisión predicha por nuestro análisis de errores para los Algoritmos 1 y 2. Todas estas pruebas numéricas se han realizado con el programa matemático MATLAB 5.3 usando un procesador AMD Athlon (tm) XP 2000+ con aritmética IEEE.

Para comparar los algoritmos 1 y 2 hemos usado como referencia los autovalores y autovectores calculados usando la aritmética de precisión variable MAPLE del paquete Symbolic Math Toolbox de MATLAB por medio del comando `vpa`. Para cada matriz A, la descomposición espectral "exacta" se obtiene usando el comando `eig` de MATLAB (es decir, con el algoritmo QR), pero eligiendo un valor suficientemente grande para la variable `digits`, que fija el número de cifras significativas con que se hacen los cálculos en MAPLE. Se elige *digits* igual a $18+d$ si el número de condición de la matriz A es $O(10^d)$. Denotamos por λ_i, q_i a los autovalores y autovectores calculados de esta manera y por $\widehat{\lambda}_i, \widehat{q}_i$ a los calculados por el algoritmo SVD con signos, implementado en MATLAB. Por tanto, $\widehat{\lambda}_i, \widehat{q}_i$ están calculados en aritmética de doble precisión, es decir, $\varepsilon \approx 2{,}2\cdot 10^{-16}$. La descomposición

inicial RRD está implementada en MATLAB, usando el Algoritmo 1 para matrices DSTU y el Algoritmo 2 para matrices TSC.

Para demostrar que los resultados obtenidos tienen la precisión deseada en todos los autovalores y autovectores analizaremos las siguientes cantidades:

1. El máximo error relativo en los autovalores

$$e_\lambda = \max_i \left| \frac{\lambda_i - \widehat{\lambda}_i}{\lambda_i} \right| \tag{3.46}$$

2. Una cantidad de control para los autovalores

$$\vartheta_\lambda = \frac{e_\lambda}{\kappa\varepsilon} \tag{3.47}$$

donde ε es el épsilon-máquina y $\kappa = \kappa(R')\,\kappa(X)$ como en (3.11). Si esta cantidad es del orden $O(1)$ significa que el resultado de los experimentos es del orden predicho por la teoría.

3. El máximo error en los autovectores

$$e_q = \max_i ||\widehat{q}_i - q_i||_2 \tag{3.48}$$

4. Una cantidad de control para los autovectores

$$\xi_q = \max_i \frac{||\widehat{q}_i - q_i||_2 \,\mathrm{relgap}^*(\widehat{\lambda}_i)}{\kappa\varepsilon} \tag{3.49}$$

con κ como antes y relgap* definido como en (3.12). De nuevo, ξ_q debería ser de orden $O(1)$ en los experimentos para confirmar el análisis.

3.5.1. Matrices DSTU

Se han generado matrices simétricas no singulares totalmente unimodulares (TU) de tamaños $6, 8, 10$ y 12. En principio no hemos podido generar matrices TU de tamaño mayor debido al alto coste que conlleva construirlas: generamos recursivamente matrices TU, comenzando con una matriz de tamaño 1, esto es o bien $-1, 1$ ó 0. Dada una matriz generada de tamaño s, el algoritmo construye una matriz TU de tamaño $(s+1) \times (s+1)$ TU orlando la matriz $s \times s$ con una nueva fila y columna, con entradas aleatorias elegidas entre ± 1 ó 0, y comprobando que todos los nuevos menores que contengan esta fila y/o columna cumplen que su valor es ± 1 ó 0. El coste computacional de comprobar estos menores es muy alto, de ahí la baja dimensión de las matrices generadas.

Una vez creada la matriz simétrica TU, la escalamos a ambos lados con una misma matriz diagonal con potencias de 10 sobre la diagonal y números de condición variando entre 10^5 y 10^{20}. Finalmente multiplicamos las potencias de 10 por valores aleatorios generados

por el comando `rand` de MATLAB . De este modo, se han generado matrices simétricas DSTU con un número de condición entre 10^{10} y 10^{40}. Para cada tamaño, dividimos los experimentos en tres grupos según sus números de condición: los que oscilan entre 10^{10} y 10^{20}, los que oscilan entre 10^{20} y 10^{30} y finalmente entre 10^{30} y 10^{40}. Generamos 100 matrices para cada uno de los tres grupos, por lo que las tablas dadas abajo reflejan el resultado para un total de 1200 matrices, 300 para cada dimensión elegida. La Tabla 1 muestra la cantidad de control ϑ_λ para los autovalores y la Tabla 2 la cantidad de control ξ_q para los autovectores. Para cada dimensión tenemos dos columnas, la izquierda nos da la media sobre 100 tests hechos en el rango correspondiente de número de condición, y la derecha nos da el valor máximo encontrado, esto es, el peor caso entre las 100 pruebas realizadas.

	$n=6$		$n=8$		$n=10$		$n=12$	
$\kappa(A)=O(10^d)$	Mean	Max	Mean	Max	Mean	Max	Mean	Max
$10 \le d \le 20$	1.412	6.689	1.746	32.34	1.879	19.14	1.425	9.310
$20 \le d \le 30$	1.460	16.34	1.652	38.14	1.432	13.49	1.696	45.45
$30 \le d \le 40$	1.699	26.65	1.338	11.34	1.157	3.949	1.719	33.02

Tabla 1: Datos estadísticos para el cálculo de autovalores de matrices simétricas DSTU: ϑ_λ

	$n=6$		$n=8$		$n=10$		$n=12$	
$\kappa(A)=O(10^d)$	Mean	Max	Mean	Max	Mean	Max	Mean	Max
$10 \le d \le 20$	0.508	2.653	0.508	1.886	0.579	1.989	0.605	2.364
$20 \le d \le 30$	0.502	1.914	0.518	1.716	0.623	2.214	0.603	1.928
$30 \le d \le 40$	0.447	1.884	0.582	2.795	0.571	2.840	0.621	2.697

Tabla 2: Datos estadísticos para el cálculo de autovectores de matrices simétricas DSTU: ξ_q

Finalmente, consideramos una subclase de matrices TU que se genera a bajo coste. Estas son las matrices simétricas tridiagonales TU eulerianas, matrices tridiagonales tales que la suma de los elementos de cualquier fila o columna es par. El coste de generar este tipo de matrices es $O(n)$ si la dimensión de la matriz generada es n. El determinante de estas matrices es siempre cero, pero el menor que resulta de eliminar la última fila y la última columna es no singular. Por tanto, generamos matrices simétricas tridiagonales TU eulerianas U de tamaño $n+1$ y realizamos las pruebas con las matrices $U(1:n,1;n)$, que son simétricas, TU, no singulares y de tamaño n. Utilizando estas matrices, volvemos a generar matrices DSTU del mismo modo que en los experimentos anteriores, denotémoslas DSTU*. Las Tablas 3 y 4 nos dan los resultados de las cantidades de control del cálculo de sus autovalores y autovectores, realizando 10 experimentos para cada dimensión. Observemos de nuevo que el resultado está en concordancia con los resultados teóricos.

	$n = 20$		$n = 40$		$n = 60$		$n = 100$	
$\kappa(A) = O(10^d)$	Mean	Max	Mean	Max	Mean	Max	Mean	Max
$10 \leq d \leq 20$	1.025	2.353	1.359	4.716	1.209	3.706	1.367	3.559
$20 \leq d \leq 30$	1.293	3.801	1.353	3.244	1.288	3.388	1.274	3.247
$30 \leq d \leq 40$	1.153	2.262	1.572	3.984	1.519	3.804	1.500	5.321

Tabla 3:Datos estadísticos para el cálculo de autovalores de matrices simétricas DSTU*: ϑ_λ

	$n = 20$		$n = 40$		$n = 60$		$n = 100$	
$\kappa(A) = O(10^d)$	Mean	Max	Mean	Max	Mean	Max	Mean	Max
$10 \leq d \leq 20$	0.534	0.841	0.670	2.578	0.601	2.330	0.720	3.132
$20 \leq d \leq 30$	0.523	1.454	0.746	2.720	0.635	2.221	0.594	2.494
$30 \leq d \leq 40$	0.588	1.835	0.770	2.125	0.641	1.855	0.776	2.826

Tabla 4: Datos estadísticos para el cálculo de autovectores de matrices simétricas DSTU*: ξ_q

3.5.2. Matrices TSC

Se han generado matrices simétricas TSC del siguiente modo: creamos una matriz aleatoria de tamaño 1×1 y repetidamente aplicamos las reglas 2 y 3 del Teorema 3.3.8 con una probabilidad dada. La regla 2 se aplica con una probabilidad del 50 %, eligiendo como A_2 uno de los bloques de la matriz A_1 calculada en la anterior etapa. La regla 3 se aplica generando los nuevos elementos $a_{i,n+1}$, $a_{n+1,n+1}$ con el comando `rand` de MATLAB, y asignando el valor cero al elemento $\widetilde{a}_{ii}$ con una probabilidad del 20 %. En la Figura 3.2 se representan dos matrices TSC de dimensión 100 generadas de ese modo.

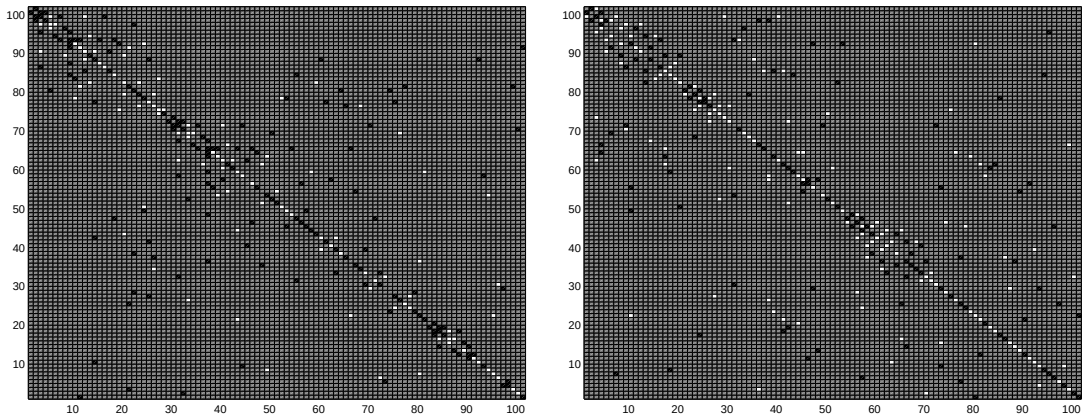


Figura 3.2: Ejemplos de matrices simétricas TSC

El uso del comando `vpa` utilizado para calcular los autovalores y autovectores de referencia para los algoritmos 1 y 2 viene limitado por el tamaño de las matrices. Por ello, a

pesar de poder generar matrices simétricas TSC de cualquier orden a un bajo coste, los experimentos los realizaremos con matrices de tamaño $10, 20, 40$, y 60. Una vez generadas estas matrices TSC las escalaremos a ambos lados con la misma matriz diagonal positiva mal condicionada, como hicimos con las matrices simétricas $DSTU$. Nótese que, como los escalamientos son positivos, el patrón de signos de la matriz TSC no cambia con el escalamiento. De nuevo, presentamos resultados para 1200 matrices, 100 de cada dimensión para cada uno de los tres rangos de números de condición. Los resultados se muestran en las tablas 5 y 6.

	$n = 10$		$n = 20$		$n = 40$		$n = 60$	
$\kappa(A) = O(10^d)$	Mean	Max	Mean	Max	Mean	Max	Mean	Max
$10 \leq d \leq 20$	1.446	10.024	1.449	4.196	1.940	8.802	2.280	9.639
$20 \leq d \leq 30$	1.332	6.579	2.170	38.68	2.033	5.172	2.278	9.528
$30 \leq d \leq 40$	1.362	5.973	1.591	7.411	2.841	44.70	2.502	9.583

Tabla 5: Datos estadísticos para el cálculo de autovalores de matrices simétricas TSC: ϑ_λ

	$n = 10$		$n = 20$		$n = 40$		$n = 60$	
$\kappa(A) = O(10^d)$	Mean	Max	Mean	Max	Mean	Max	Mean	Max
$10 \leq d \leq 20$	0.682	3.044	0.843	2.987	1.292	3.342	1.418	3.641
$20 \leq d \leq 30$	0.717	7.438	0.889	4.215	1.294	3.034	1.405	3.665
$30 \leq d \leq 40$	0.800	3.768	0.893	3.386	1.265	2.802	1.471	3.672

Tabla 6: Datos estadísticos para el cálculo de los autovectores de matrices simétricas TSC: ξ_q

Como se puede ver en las tablas 5 y 6, los resultados confirman nuestras predicciones teóricas. En paralelo, se calcularon los autovalores y autovectores con el comando `eig` de MATLAB. Como es de esperar, los errores relativos fueron extremadamente grandes, sin dígito correcto alguno para los autovalores más pequeños.

Para concluir con los experimentos numéricos, presentamos una última prueba para dar una idea del coste computacional de la factorización LDL^T por bloques de matrices simétricas TSC. Más concretamente para matrices de tamaños que van desde 10 a 100 en intervalos de diez, es decir, generamos cien matrices de tamaño 10, cien de 20, cien de 30 y así sucesivamente, un total de mil matrices en total. Para cada matriz se calcula su respectiva descomposición LDL^T por bloques usando el Algoritmo 2, y contamos el número de operaciones en coma flotante llevadas a cabo en el proceso. Cada estrella en la Figura 3.3 corresponde a un tamaño fijo y representa la media aritmética de las operaciones realizadas en las 100 matrices de ese tamaño, representandola en una gráfica a escala logarítmica, con el logaritmo del tamaño n de la matriz en el eje de abscisas. La línea continua corresponde a `flops`$= n^4$, y la línea discontinua a `flops`$= n^3$. Como puede verse en la figura, el coste parece estar entre ambos órdenes. Como no hemos estimado las constantes que aparecen en la o-mayúscula de Landau, es difícil llegar a conclusiones rigurosas.

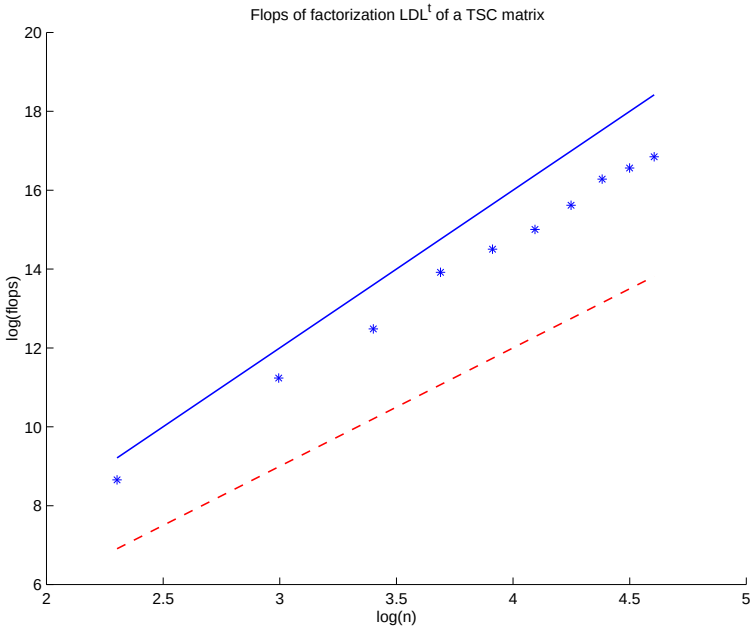


Figura 3.3: Coste computacional de la factorización LDL^T por bloques de matrices simétricas TSC de tamaños entre 10 y 100

3.6. Conclusiones y trabajos futuros

Hemos descrito y llevado a cabo un análisis de errores detallado de dos algoritmos que calculan todos los autovalores y autovectores con alta precisión relativa de cualquier matriz simétrica perteneciente a las clases DSTU o TSC.

Extensiones y cuestiones relacionadas con este trabajo y que podrán ser motivo de estudio en el futuro son:

1. Estudiar nuevos algoritmos de factorización de las clases de matrices estudiadas, empleando otras estrategias de pivote. Por ejemplo, la estrategia de pivote de Bunch-Kauffman.

2. Buscar nuevas clases de matrices, eventualmente no simétricas, cuyas propiedades permitan evitar cancelaciones en el proceso de factorización.

Bibliografía

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK user's guide, Third ed.* SIAM, Philadelphia, 1999.

[2] J. Barlow and J. Demmel. Computing accurate eigensystems of scaled diagonally dominant matrices. *SIAM J. Numer. Anal.*, 27:762–791, 1990.

[3] H. Baumgärtel. *Analytic Perturbation Theory for Matrices and Operators.* Birkhäuser, Basel, 1985.

[4] P. Benner and D. Kressner. Algorithm 854: Fortran 77 subroutines for computing the eigenvalues of Hamiltonian matrices II. *ACM Trans. Math. Softw.*, 32(2):352–373, 2006.

[5] P. Benner, V. Mehrmann, and H. Xu. A numerically stable, structure-preserving method for computing the eigenvalues of real Hamiltonian or symplectic pencils. *Num. Math.*, 78:329–358, 1996.

[6] P. Benner, V. Mehrmann, and H. Xu. A new method for computing the stable invariant subspace of a real Hamiltonian matrix. *J. Comput. Appl. Math.*, 86(1):17–43, 1997.

[7] Å. Björck and V. Pereyra. Solution of Vandermonde systems of equations. *Math. Comp.*, 24:893–903, 1970.

[8] S. Bora and V. Mehrmann. Linear perturbation theory for structured matrix pencils arising in control theory. *SIAM J. Matrix Anal. Appl.*, 28(1):148–169, 2006.

[9] S. Boyd, V. Balakrishnan, and P. Kabamba. A bisection method for computing the $\mathbf{H}_\infty$ norm of a transfer matrix and related problems. *Mathematics of Control, Signals, and Systems*, 2(3):207–219, 1989.

[10] R. P. Brent. Stability of fast algorithms for structured linear systems. In *Fast reliable algorithms for matrices with structure, SIAM*, pages 103–152, 1999.

[11] R. Brualdi and H. Ryser. *Combinatorial Matrix Theory.* Cambridge University Press, Cambridge, 1991.

[12] J. R. Bunch. The weak and strong stability of algorithms in numerical linear algebra. *Linear Algebra Appl.*, 88/89:49–66, 1987.

[13] J.R. Bunch. Analysis of the diagonal pivoting method. *SIAM J. Numer. Anal.*, 8(4):656–680, 1971.

[14] J.R. Bunch and B. Parlett. Direct methods for solving symmetric indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 8(3):639–655, 1971.

[15] J. V. Burke, A. S. Lewis, and M. L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM J. Optim.*, 15(3):751–779, 2005.

[16] R. Byers and D. Kressner. On the condition of a complex eigenvalue under real perturbations. *BIT*, 44(2):209–215, 2004.

[17] F. Chatelin. *Eigenvalues of Matrices*. Wiley, New York, 1993.

[18] R. W. Cottle. Manifestations of the schur complement. *Linear Algebra Appl.*, 8:189–211, 1974.

[19] C. Davis and W. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7:1–46, 1970.

[20] J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia.

[21] J. W. Demmel and A. Edelman. The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms. *Linear Algebra Appl.*, 230, 1995.

[22] J. W. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić, and Z. Drmač. Computing the singular value decomposition with high relative accuracy. *Linear Algebra Appl.*, 299:21–80, 1999.

[23] J. W. Demmel and W. Kahan. Accurate singular values of bidiagonal matrices. *SIAM J. Sci. Stat. Comp.*, 11:873–912, 1990.

[24] J. W. Demmel and K. Veselić. Jacobi's method is more accurate than QR. *SIAM J. Matrix Anal. Appl.*, 13:1204–1245, 1992.

[25] J.W. Demmel and W. Gragg. On computing accurate singular values and eigenvalues of matrices with acyclic graphs. *Linear Algebra Appl.*, 185:203–217, 1993.

[26] I. S. Dhillon. *A new $O(n^2)$ algorithm for the symmetric tridiagonal eigenvalue/eigenvector problem*. PhD thesis, University of California at Berkeley, Berkeley (USA), 1998.

[27] I.S. Dhillon and B.N. Parlett. Orthogonal eigenvectors and relative gaps. *SIAM J. Matrix Anal. Appl.*, 25(3):858–899, 2003.

[28] I.S. Dhillon, B.N. Parlett, and C. Vömel. The design and implementation of the MRRR algorithm. *ACM Trans. Math. Softw.*, 32(4):533–560, 2006.

[29] F. M. Dopico and P. Koev. Accurate symmetric rank-revealing and eigendecompositions of symmetric structured matrices. *SIAM J. Matrix Anal. Appl.*, 28:1126–1156, 2006.

[30] F. M. Dopico, J. M. Molera, and J. Moro. An orthogonal high relative accuracy algorithm for the symmetric eigenproblem. *SIAM J. Matrix Anal. Appl.*, 25(2):301–351, 2003.

[31] Z. Drmač and K. Veselić. LAPACK Working Note 169: New fast and accurate Jacobi SVD algorithm: I. Technical report, 2005.

[32] Z. Drmač and K. Veselić. LAPACK Working Note 170: New fast and accurate Jacobi SVD algorithm: II. Technical report, 2005.

[33] V. Olshevsky (editor). *Structured matrices in mathematics, computer science, and engineering: Proceedings of an AMS-IMS-SIAM joint Summer Research Conference, Boulder, June 27-July 1, 1999*. AMS, 2001.

[34] H. Faßbender, D. S. Mackey, N. Mackey, and X. Xu. Hamiltonian square roots of skew-Hamiltonian matrices. *Linear Algebra Appl.*, 287(1-3):125–159, 1999.

[35] K. V. Fernando and B. Parlett. Accurate singular values and differential qd algorithms. *Num. Math.*, 67:191–229, 1994.

[36] B. Friedlander, M. Morf, T. Kailath, and L. Ljung. New inversion formulas for matrices classified in terms of their distance from Toeplitz matrices. *Linear Algebra Appl.*, 27:31–60, 1979.

[37] I. Gohberg, M. A. Kaashoek, and P. Lancaster. General theory of regular matrix polynomials and band Toeplitz operators. *Integral Equations and Operator Theory*, 11:776–882, 1988.

[38] I. Gohberg, P. Lancaster, and L. Rodman. *Matrix polynomials*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1982.

[39] S. Graillat. A note on structured pseudospectra. *J. Comput. Appl. Math.*, 191(1):68–76, 2006.

[40] G. Heinig and K. Rost. *Algebraic methods for Toeplitz-like matrices and operators*. Birkhäuser, Basel, 1984. Operator Theory, vol. 13.

[41] D. J. Higham and N. J. Higham. Backward error and condition of structured linear systems. *SIAM J. Matrix Anal. Appl.*, 13(1):162–175, 1992.

[42] D. J. Higham and N. J. Higham. Structured backward error and condition of generalized eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 20(2):493–512, 1999.

[43] N. J. Higham. *Accuracy and Stability of Numerical Algorithms (2nd. ed.)*. SIAM, Philadelphia, 2002.

[44] N. J. Higham, R.-C. Li, and F. Tisseur. Backward error of polynomial eigenproblems solved by linearization. MIMS EPrint 2006.137, Manchester Institute for Mathematical Sciences, University of Manchester, Manchester, UK, 2006. See `http://eprints.ma.man.ac.uk/312`.

[45] N.J. Higham, D.S. Mackey, and F. Tisseur. The conditioning of linearizations of matrix polynomials. *SIAM Journal on Matrix Analysis and Applications*, 28(4):1005–1028, 2006.

[46] A. Hilliges, C. Mehl, and V. Mehrmann. On the solution of palindromic eigenvalue problems. In *Proceedings of ECCOMAS, Jyväskylä, Finland*, 2004.

[47] D. Hinrichsen and A.J. Pritchard. Stability radii of linear systems. *Syst. Control Lett.*, 7(1):1–10, 1986.

[48] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[49] R. A. Horn and V. V. Sergeichuk. Canonical forms for complex matrix congruence and *-congruence. *Linear Algebra Appl.*, 416(2-3):1010–1032, 2006.

[50] A. Householder. *The Theory of Matrices in Numerical Analysis*. Dover, 1975.

[51] K. D. Ikramov. Hamiltonian square roots of skew-Hamiltonian matrices revisited. *Linear Algebra Appl.*, 325(1-3):101–107, 2001.

[52] W. Kahan. Accurate eigenvalues of a symmetric tridiagonal matrix. Technical Report CS-41, Department of Computer Science, Stanford University, Palo Alto (USA), 1968.

[53] T. Kailath, S.-Y. Kung, and M. Morf. Displacement ranks of matrices and linear equations. *J. Math. Anal. Appl.*, 68:395–407, 1979.

[54] T. Kailath and A. H. Sayed. Displacement structure: theory and applications. *SIAM Rev.*, 37:297–386, 1995.

[55] M. Karow, D. Kressner, and F. Tisseur. Structured eigenvalue condition numbers. *SIAM J. Matrix Anal. Appl.*, 28(4):1056–1058, 2006.

[56] T. Kato. *Perturbation Theory for Linear Operators*. Springer, Berlin, 1980.

[57] P. Koev. Accurate eigenvalues and SVDs of totally nonnegative matrices. *SIAM J. Matrix Anal. Appl.*, 27(1):1–23, 2005.

[58] P. Koev and F. M. Dopico. Accurate eigenvalues of certain sign-regular matrices, 2006. preprint.

[59] D. Kressner, M. J. Peláez, and J. Moro. Structured Hölder condition numbers for multiple eigenvalues, 2006. preprint.

[60] P. Lancaster and P. Psarrakos. A note on weak and strong linearizations of regular matrix polynomials. Technical Report No. 470, Manchester Institute for Mathematical Sciences, University of Manchester, June 2005.

[61] H. Langer and B.Ñajman. Remarks on the perturbation of analytic matrix functions. II. *Integral Equations and Operator Theory*, 12(3):392–407, 1989.

[62] H. Langer and B.Ñajman. Remarks on the perturbation of analytic matrix functions. III. *Integral Equations and Operator Theory*, 15(5):796–806, 1992.

[63] H. Langer and B.Ñajman. Leading coefficients of the eigenvalues of perturbed analytic matrix functions. *Integral Equations and Operator Theory*, 16(4):600–604, 1993.

[64] V. B. Lidskiĭ. On the theory of perturbations of nonselfadjoint operators. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 6(1):52–60, 1966.

[65] C. F. Van Loan. A symplectic method for approximating all the eigenvalues of a Hamiltonian matrix. *Linear Algebra Appl.*, 61:233–252, 1984.

[66] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann. Palindromic polynomial eigenvalue problems: Good vibrations from good linearizations. *SIAM J. Matrix Anal. Appl.*, 28(4):1029–1051, 2006.

[67] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann. Vector spaces of linearizations for matrix polynomials. *SIAM J. Matrix Anal. Appl.*, 28(4):971–1004, 2006.

[68] D. S. Mackey, N. Mackey, and F. Tisseur. $\mathbb{G}$-reflectors: analogues of Householder transformations in scalar product spaces. *Linear Algebra Appl.*, 385:187–213, 2004.

[69] D. S. Mackey, N. Mackey, and F. Tisseur. Structured mapping problems for matrices associated with scalar products part I: Lie and Jordan algebras. MIMS EPrint 2006.44, Manchester Institute for Mathematical Sciences, University of Manchester, Manchester, UK, 2006. See `http://eprints.ma.man.ac.uk/197/`.

[70] R. Mathias. The singular values of the Hadamard product of a positive semidefinite and a skew-symmetric matrix. *Linear and Multilinear Algebra*, 31(1-4):57–70, 1992.

[71] C. Mehl. On classification of normal matrices in indefinite inner product spaces. *Electron. J. Linear Algebra*, 15:50–83, 2006.

[72] J. Moro, J. V. Burke, and M. L. Overton. On the Lidskii-Vishik-Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure. *SIAM J. Matrix Anal. Appl.*, 18(4):793–817, 1997.

[73] S.Ñoschese and L. Pasquini. Eigenvalue condition numbers: zero-structured versus traditional. *J. Comput. Appl. Math.*, 185(1):174–189, 2006.

[74] V. Olshevsky. Pivoting for structured matrices and rational tangential interpolation. In *Contemporary mathematics: theory and applications, AMS*, pages 1–73, 2003.

[75] C. C. Paige and C. Van Loan. A Schur decomposition for Hamiltonian matrices. *Linear Algebra Appl.*, 41:11–32, 1981.

[76] M. J. Peláez and J. Moro. Accurate factorization and eigenvalue algorithms for symmetric DSTU and TSC matrices. *SIAM J. Matrix Anal. Appl.*, 28(4):1173–1198, 2006.

[77] M. J. Peláez and J. Moro. Structured condition numbers of multiple eigenvalues. *Proceedings in Applied Mathematics and Mechanics (PAMM)*, 6(1):67–70, 2006.

[78] M. J. Peláez and J. Moro. Hölder condition numbers for eigenvalues under fully nongeneric structured perturbation, 2007. preprint.

[79] V. Puiseux. Recherches sur les fonctions algébriques. *J. Math Pures Appl.*, 15:365–480, 1850.

[80] L. Rodman. Bounded and stably bounded palindromic difference equations of first order. *Electron. J. Linear Algebra*, 15:22–49, 2006.

[81] S. M. Rump. Structured perturbations. I. Normwise distances. *SIAM J. Matrix Anal. Appl.*, 25(1):1–30, 2003.

[82] S. M. Rump. Eigenvalues, pseudospectrum and structured perturbations. *Linear Algebra Appl.*, 413(2-3):567–593, 2006.

[83] C. Schröder. A canonical form for palindromic pencils and palindromic factorizations. Technical report, MATHEON, DFG Research Center "Mathematics for key technologies"in Berlin, TU Berlin, 2006.

[84] I. Slapničar. *Accurate Symmetric Eigenreduction by a Jacobi Method*. PhD thesis, Fernuniversität Hagen, 1992.

[85] I. Slapničar. Componentwise analysis of direct factorizations of real symmetric and Hermitian matrices. *Linear Agebra Appl.*, 272:227–275, 1998.

[86] I. Slapničar. Highly accurate symmetric eigenvalue decomposition and hyperbolic SVD. *Linear Algebra Appl.*, 358:387–424, 2003.

[87] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, New York, 1990.

[88] F. de Terán, F. Dopico, and J. Moro. First order spectral perturbation theory of matrix pencils via the Kronecker form, 2006. In preparation.

[89] R. C. Thompson. Pencils of complex and real symmetric and skew matrices. *Linear Algebra Appl.*, 147:323–371, 1991.

[90] F. Tisseur. Backward error and condition of polynomial eigenvalue problems. *Linear Algebra Appl.*, 309(1-3):339–361, 2000.

[91] F. Tisseur. A chart of backward errors for singly and doubly structured eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 24(3):877–897, 2003.

[92] L.Ñ. Trefethen. Pseudospectra of linear operators. *SIAM Rev.*, 39:383–406, 1997.

[93] L.Ñ. Trefethen and M. Embree. *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*. Princeton University Press, Princeton, 2005.

[94] K. Veselić. Floating-point perturbations of Hermitian matrices. *Linear Algebra Appl.*, 195(1):81–116, 1993.

[95] K. Veselić. A Jacobi eigenreduction algorithm for definite matrix pairs. *Numer. Math.*, 64(2):241–269, 1993.

[96] M. I. Vishik and L. A. Ljusternik. Solution of some perturbation problems in the case of matrices and self-adjoint or non-selfadjoint differential equations. I. *Russian Math. Surveys*, 15(3):1–73, 1960.

[97] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, 1965.

[98] D. Xie, X. Hu, and L. Zhang. The solvability conditions for inverse eigenproblem of symmetric and anti-persymmetric matrices and its approximation. *Numer. Linear Algebra Appl.*, 10(3):223–234, 2003.