Métodos numéricos y alta precisión relativa para ciertas clases de matrices

Paula Nerea Corral Hernández

Trabajo de fin de grado de Matemáticas
Universidad de Zaragoza

Director del trabajo:
Juan Manuel Peña Ferrández
31 de mayo de 2024

Resumen

Esta memoria se sitúa dentro del análisis numérico, que es una rama de las matemáticas y la informática que se encarga del desarrollo, análisis y aplicación de métodos numéricos para resolver problemas matemáticos de los que no se han podido encontrar soluciones analíticas o en los que se necesita una aproximación numérica de las soluciones.

Este campo abarca una amplia gama de técnicas, como métodos de aproximación, interpolación, integración numérica, resolución de ecuaciones diferenciales, entre otros. En este trabajo nos centramos en conseguir métodos numéricos con alta precisión relativa para resolver problemas de álgebra lineal como valores propios, valores singulares, inversa de una matriz o la resolución de sistemas de ecuaciones lineales. La alta precisión relativa es muy deseable, pero solo se ha conseguido hasta ahora en muy pocos casos. En particular, para ciertas clases de matrices estructuradas.

En esta memoria, vamos a hacer uso de las matrices totalmente positivas (TP), que son matrices con todos sus menores no negativos, y que presentan propiedades que han permitido obtener algoritmos con alta precisión relativa para varias subclases de las mismas. Además, necesitamos un algoritmo alternativo a la eliminación gaussiana, como es la eliminación de Neville, que será la herramienta básica para conseguir la factorización bidiagonal de una matriz cualquiera (veremos que es única). Esta factorización bidiagonal nos va a proporcionar los parámetros de partida necesarios para construir algoritmos con alta precisión relativa. Para asegurar dicha alta precisión relativa usamos una condición suficiente que consiste en prohibir las restas de números de signos iguales, excepto cuando se trata de datos iniciales. Es decir, el algoritmo solo puede realizar multiplicaciones, divisiones, sumas de números reales con signos iguales. El problema con la eliminación de Neville aplicada a matrices TP es que involucra restas. Por tanto, puede hacer falta conseguir la factorización bidiagonal por algún procedimiento alternativo. En esta memoria, consideramos tanto matrices de Pascal y sus generalizaciones como matrices de q-enteros.

Dada la factorización bidiagonal de una matriz TP no singular es posible llevar a cabo los cálculos necesarios para los problemas algebraicos mencionados anteriormente de manera implícita, mediante la transformación de las entradas de su factorización bidiagonal de tal manera que no se requieran restas. Por lo tanto, el problema de realizar cálculos con alta precisión relativa con una matriz TP no singular se transforma en el problema de encontrar su factorización bidiagonal con alta precisión relativa.

En el primer capítulo introducimos algunos conceptos, notaciones y resultados básicos que son necesarios en el resto de la memoria. Empezamos considerando los errores que pueden aparecer en los cálculos computacionales y presentando la importancia de la alta precisión relativa. A continuación, introducimos notaciones matriciales y la definición y propiedades de las matrices totalmente positivas, que serán de gran utilidad en el resto del trabajo. También presentamos la eliminación de Neville que es un procedimiento utilizado para hacer ceros en las distintas columnas bajo la diagonal principal de la matriz de partida. Además, presentamos la factorización bidiagonal de dicha matriz, que se obtiene teóricamente a través de la eliminación de Neville. Finalmente, damos algunas operaciones que se pueden realizar con alta precisión relativa en matrices totalmente positivas.

En el segundo capítulo consideramos las matrices de Pascal que, en particular, son matrices totalmente positivas y simétricas, que presentan importantes aplicaciones en probabilidad, en combinatoria y en análisis numérico, entre otros campos. Estas matrices están mal condicionadas, siendo incluso peor condicionadas que las matrices de Vandermonde. Aún así, se pueden obtener algoritmos con alta precisión relativa para el cálculo de valores propios e inversas de matrices de Pascal, así como para resolver ciertos sistemas lineales cuyas matrices de coeficientes son matrices de Pascal. Comenzamos introduciendo

las definiciones de las matrices de Pascal de orden n y de las matrices de Pascal triangulares inferiores de orden n. Continuaremos viendo la factorización bidiagonal de estas matrices, que es extraordinariamente simple. Posteriormente, definimos las matrices de Pascal generalizadas y algunas notaciones que necesitaremos más adelante. Por último, describimos la factorización bidiagonal de este tipo de matrices.

En el último capítulo presentamos las matrices de q-enteros. Muchos cálculos algebraicos como el cálculo de valores propios, valores singulares e inversas de estas matrices se pueden realizar con alta precisión relativa. Comenzamos viendo qué es un q-entero y algunas de sus propiedades. También estudiamos qué aspecto tiene la factorización bidiagonal de las matrices de q-Pascal. Observamos que esta factorización bidiagonal (para $q \neq 0$) no es tan sencilla como la de las matrices de Pascal vista en el capítulo anterior. Finalmente, introducimos los números de q-Stirling y la factorización bidiagonal de las matrices con números de q-Stirling.

Abstract

This work falls within the field of numerical analysis, which is a branch of mathematics and computer science that deals with the development, analysis, and application of numerical methods. It is used to solve mathematical problems for which analytical solutions have not been found or where numerical approximations of solutions are needed.

This field encompasses a wide range of techniques among which are remarkable approximation methods, interpolation, numerical integration, and solving differential equations. In this work, we focus on achieving high relative accuracy (HRA) numerical methods to solve linear algebra problems such as eigenvalues, singular values, matrix inversion, or the solution of linear systems of equations. High relative accuracy is desirable, but it has only been achieved in very few cases so far. In particular, for certain classes of structured matrices.

In this work, we will use totally positive (TP) matrices, which are matrices with all their minors nonnegative and exhibit properties that have enabled the development of algorithms with HRA for several subclasses of TP matrices. Additionally, we need an alternative algorithm to Gaussian elimination, such as Neville elimination, which will be the basic tool for achieving the bidiagonal factorization of any matrix (we will see that it is unique). This bidiagonal factorization will provide us with the necessary starting parameters to construct algorithms with high relative accuracy. To ensure such HRA, we use a sufficient condition that consists of prohibiting subtractions of numbers of the same sign, except when dealing with initial data. That is, the algorithm can only perform multiplications, divisions, and additions of real numbers with equal signs. The problem with Neville elimination applied to TP matrices is that it involves subtractions. Therefore, an alternative procedure may be necessary to achieve the bidiagonal factorization. In this work, we consider both Pascal matrices and their generalizations as well as q-integers matrices.

Given the bidiagonal factorization of a nonsingular TP matrix, it is possible to perform the required calculations for the algebraic problems implicitly. This is achieved by transforming the entries of its bidiagonal factorization in a manner that avoids the need for subtractions. Therefore, the problem of performing calculations with HRA with a nonsingular TP matrix is transformed into the problem of finding its bidiagonal factorization with HRA.

In the first chapter, we introduce some concepts, notations, and basic results that are necessary throughout the work. We start by considering the errors that may arise in computational calculations and highlighting the importance of high relative accuracy. Next, we introduce matrix notations and the definition and properties of totally positive matrices, which will be very useful in the rest of the work. We also present Neville elimination, which is a procedure used to create zeros in different columns below the main diagonal of the starting matrix. Furthermore, we discuss the bidiagonal factorization of such a matrix, which is theoretically obtained through Neville elimination. Finally, we provide some operations that can be performed with high relative accuracy on totally positive matrices.

In the second chapter, we consider Pascal matrices, which are, in particular, totally positive and symmetric. They have significant applications in probability, combinatorics, and numerical analysis, among other fields. These matrices are ill-conditioned, even worse conditioned than Vandermonde matrices. Nonetheless, algorithms with high relative accuracy can be obtained for calculating eigenvalues and inverses of Pascal matrices, as well as for solving certain linear systems whose coefficient matrices are Pascal matrices. We begin by introducing the definitions of Pascal matrices of order n and lower triangular Pascal matrices of order n. We then proceed to discuss the bidiagonal factorization of these matrices, which is

remarkably simple. Subsequently, we define generalized Pascal matrices and some notations that we will need later on. Finally, we describe the bidiagonal factorization of this type of matrices.

In the last chapter, we introduce q-integers matrices. Many algebraic calculations such as eigenvalue computation, singular values, and inverses of these matrices can be performed with HRA. We start with the concept of q-integer and some of its properties. We also study the form of the bidiagonal factorization of q-Pascal matrices. Observing that this bidiagonal factorization (for $q \neq 0$) is not as simple as that of Pascal matrices seen in the previous chapter. Finally, we introduce q-Stirling numbers and the bidiagonal factorization of matrices with q-Stirling numbers.

Índice general

Capítulo 1

Conceptos y resultados básicos

1.1. Introducción

Este capítulo tiene como objetivo introducir conceptos, notaciones y resultados básicos para el resto de la memoria.

En el álgebra lineal numérica, la búsqueda de la precisión y la exactitud en los cálculos es esencial para garantizar la validez de los resultados obtenidos. Los errores inherentes a los métodos numéricos pueden surgir debido a la limitada precisión de las representaciones numéricas utilizadas en los algoritmos, lo que puede impactar significativamente en la fiabilidad de los resultados. Es por ello que la alta precisión relativa es una propiedad muy deseable para mitigar la propagación de errores y mejorar la fiabilidad de los cálculos.

El segundo apartado de este capítulo presenta los conceptos relacionados con los errores y la alta precisión relativa.

En el tercer apartado de este primer capítulo vamos a introducir notaciones matriciales y las matrices totalmente positivas, ya que estas estructuras matriciales especiales presentan propiedades que son utilizadas para mejorar la estabilidad numérica en diversos problemas.

En el cuarto apartado presentamos la eliminación de Neville y la factorización bidiagonal. Esta factorización bidiagonal nos proporciona los parámetros de partida para poder construir algoritmos con alta precisión relativa, con tal de que dichos parámetros también los obtengamos con alta precisión relativa.

Finalmente, en el último apartado introducimos algunas operaciones que se pueden realizar con alta precisión relativa en matrices totalmente positivas.

1.2. Errores y cálculos con alta precisión relativa

En este apartado, consideraremos los errores que pueden aparecer en los cálculos computacionales y presentaremos la importante noción de alta precisión relativa, que será muy importante en esta memoria.

Comenzaremos con definiciones y tipos de errores. Continuaremos presentando los errores backward y forward, y después consideraremos el condicionamiento. Finalmente, trataremos de la cancelación y de la alta precisión relativa.

1.2.1. Definiciones y tipos de errores

Comenzaremos con algunas definiciones básicas de errores.

Definición 1: Sea $\hat{x}$ una aproximación del número real x. El *error absoluto* cometido al hallar $\hat{x}$ es $E_{\text{abs}}(\hat{x}) = |x - \hat{x}|$, y su *error relativo* $E_{\text{rel}}(\hat{x}) = \frac{|x-\hat{x}|}{|x|}$ cuando $x \neq 0$.

Una definición equivalente del error relativo es $E_{\text{rel}}(\hat{x}) = |\rho|$ donde $\hat{x} = x(1+\rho)$. Cuando el signo es importante en el error absoluto, hablaremos simplemente del error $x - \hat{x}$.

El error relativo está relacionado con la noción de dígitos significativos correctos (o cifras significativas correctas). Los k dígitos significativos en un número son el primer dígito no nulo y los $k-1$ dígitos que le siguen. Es decir, una aproximación x a $\hat{x}$ tiene ρ dígitos significativos correctos si $\hat{x}$ y x redondean al mismo número con ρ dígitos significativos.

Cuando x y $\hat{x}$ son vectores, la definición anterior se extiende de esta forma:

Definición 2: El *error relativo* cometido al calcular el vector $\hat{x}$, definido cuando $x \neq 0$, es

$$E_{\text{rel}}(\hat{x}) = \frac{\|\hat{x}-x\|}{\|x\|}.$$

Para las normas más usadas $\|x\|_\infty := \text{máx}_i|x_i|$, $\|x\|_1 := \sum_i|x_i|$, y $\|x\|_2 := \sqrt{x^\top x}$, la desigualdad

$$\frac{\|x-\hat{x}\|}{\|x\|} \leq \frac{1}{2}\cdot 10^{-\rho}$$

implica que las componentes $\hat{x}_i$ con $|\hat{x}_i| \approx \|x\|$ tienen alrededor de ρ dígitos correctos significativos pero, para componentes menores, la desigualdad simplemente acota el error.

Esto motiva la siguiente definición:

Definición 3: El error relativo componente a componente del vector $\hat{x}$, definido cuando $x_i \neq 0$, es $\text{máx}\,\frac{|x_i-\hat{x}_i|}{|x_i|}$.

Hay tres fuentes principales de errores en el cálculo numérico: de redondeo, incertidumbre en los datos y truncamiento.

Por lo general, los efectos de los errores en los datos son más fáciles de entender que los efectos de los errores de redondeo cometidos durante un cálculo, ya que los errores en los datos pueden analizarse utilizando teoría de perturbaciones para el problema en cuestión, mientras que los errores de redondeo intermedio requieren un análisis específico para el método dado.

1.2.2. Precisión, errores forward y errores backward

Cuando trabajamos computacionalmente con aritmética de precisión finita, la *precisión* (*precision* también en inglés) es la exactitud con la que se realizan las operaciones aritméticas básicas $+,-,*,/$ y en la aritmética de punto flotante se mide mediante la unidad de redondeo u.

Pero, hay un segundo sentido de la palabra precisión (que corresponde en inglés a *accuracy*) y que se refiere al error absoluto o relativo de una cantidad aproximada.

Suponemos que una aproximación $\hat{y}$ de $y = f(x)$ se calcula en una aritmética de precisión u, donde f es una función escalar real de variable real escalar. Al error (absoluto o relativo) obtenido al calcular $\hat{y}$ lo llamamos *error forward* (o progresivo).

En lugar de enfocarnos en el error relativo de $\hat{y}$ nos podemos preguntar: ¿para qué conjuntos de datos hemos resuelto realmente nuestro problema?, es decir, ¿para qué Δx se cumple $\hat{y} = f(x+\Delta x)$? En general, habrá muchos Δx, pero necesitamos encontrar el más pequeño. El valor de $|\Delta x|$, a veces dividido por $|x|$ (en el caso relativo), se llama *error backward* (o regresivo). Aunque nos interesan los errores forward, suele ser más fácil hallar los errores backward, que se relacionan con los forward como se indica en el apartado siguiente.

Un método para calcular $y = f(x)$ se dice que es estable backward si, para cada x, $\hat{y} = f(x+\Delta x)$ para algunos Δx pequeños. En general, un problema podrá ser resuelto a través de varios métodos de los cuales algunos serán estables backward y otros no. De hecho, la estabilidad backward depende del método utilizado.

1.2.3. Condicionamiento

La relación entre el error backward y forward viene dada por el condicionamiento del problema, es decir, la sensibilidad de la solución a perturbaciones en los datos.

Continuando con el ejemplo anterior $y = f(x)$ cuya aproximación satisface $\hat{y} = f(x+\Delta x)$. Asumiendo que f es dos veces diferenciable, llegamos a la definición del número de condición

$$\frac{\hat{y}-y}{y} = (\frac{xf'(x)}{f(x)})\frac{\Delta x}{x} + o((\Delta x)^2),$$

donde

$$c(x) = |\frac{xf'(x)}{x}|$$

es el número de condición de f. Si x o f es un vector, el número de condición se define de manera análoga usando normas.

En general, cuando en un problema tenemos bien definido el error forward, el error backward y el número de condición correspondientes, se da la relación:

$$\text{error forward} \leq \text{número de condición} \cdot \text{error backward}$$

Una forma de interpretarla es viendo que la solución calculada para un problema mal condicionado puede tener un error forward grande. Incluso si la solución calculada tiene un error backward pequeño, este error puede amplificarse por un factor tan grande como es el número de condición.

Como el mal condicionamiento es intrínseco al problema, si queremos evitarlo debemos reparametrizar el problema inicial.

1.2.4. Cancelación y alta precisión relativa

La cancelación se produce cuando se restan dos números casi iguales. A menudo, pero no siempre, es algo perjudicial.

Para obtener una mayor comprensión de este fenómeno consideramos la resta (en aritmética exacta) $\hat{x} = \hat{a} - \hat{b}$, donde $\hat{a} = a(1+\Delta a)$ y $\hat{b} = b(1+\Delta b)$. Los términos Δa y Δb son los errores relativos. Tomando $x = a - b$ tenemos

$$|\frac{x-\hat{x}}{x}| = |\frac{-a\Delta a + b\Delta b}{a-b}| \leq max(|\Delta a|, |\Delta b|)\frac{|a|+|b|}{|a-b|}.$$

La cota del error relativo para $\hat{x}$ es grande cuando $|a-b| \ll |a|+|b|$, es decir, cuando hay una gran cancelación en la resta. Este análisis muestra que la cancelación sustractiva puede amplificar considerablemente errores anteriores. Es importante darse cuenta de que la cancelación no es siempre algo perjudicial. En primer lugar, los números que se están restando pueden ser libres de errores, como cuando provienen de datos iniciales que se conocen exactamente. En segundo lugar, la cancelación puede ser un síntoma del mal condicionamiento intrínseco del problema y, por lo tanto, puede ser inevitable. Y, en tercer lugar, el efecto de la cancelación depende del papel que juegue el resultado en el cálculo restante.

Para obtener resultados con varias cifras significativas correctas, buscamos que el error de nuestro algoritmo satisfaga la siguiente relación:

$$\text{error forward relativo} \leq Ku, \quad \text{para alguna constante } K, \text{donde u es la unidad de redondeo.}$$

En este contexto, afirmamos que los cálculos se han llevado a cabo con *alta precisión relativa* (HRA, de *high relative accuracy*). Sin embargo, lamentablemente, no es posible lograr la HRA para todos los problemas. Un ejemplo sencillo que no puede realizarse con HRA es la evaluación de la expresión $x+y+z$ (véase [10]). Como hemos comentado antes, las cancelaciones pueden conducir a errores relativos grandes, aunque no siempre.

Por ejemplo, podemos realizar la resta de dos datos iniciales conocidos con precisión sin que se produzca una cancelación perjudicial. En cualquier caso, este fenómeno es algo que debemos tener en cuenta al diseñar un método con alta precisión relativa (HRA).

Existe una condición suficiente para asegurar la alta precisión relativa de un algoritmo (véase [11]). Cuando las operaciones realizadas en el algoritmo incluyen sumas de números del mismo signo, multiplicaciones, divisiones y restas de datos iniciales (entendiendo la resta como la diferencia entre dos cantidades del mismo signo). Es decir, se prohíben las restas, excepto cuando se trata de datos iniciales.

Para tener algoritmos que cumplan la condición anterior, suele ser necesaria una reparametrización del problema de partida. En el caso de las matrices especiales totalmente positivas que usaremos en esta memoria (y que introducimos en el apartado siguiente) la reparametrización vendrá dada por usar una factorización de las mismas (llamada factorización bidiagonal) en vez de las entradas de la matriz.

1.3. Notaciones matriciales y matrices totalmente positivas

Comenzamos definiendo las matrices totalmente positivas y estrictamente totalmente positivas.

Definición 4: Una matriz $A = (a_{ij})_{1\leq i,j\leq n}$ con todos los menores no negativos se llama *matriz totalmente positiva* (TP). Si todos los menores son estrictamente positivos la matriz se llama *estrictamente totalmente positiva* (STP).

Las matrices TP y STP también se llaman totalmente no negativas y totalmente positivas, respectivamente.

Estas clases de matrices tienen importantes aplicaciones en diversos campos como teoría de aproximación, diseño geométrico asistido por ordenador, sistemas mecánicos, combinatoria, estadística, economía, etc. (como se comenta en [3],[13],[14] y [21]).

Sea $Q_{k,n}$ el conjunto de sucesiones estrictamente crecientes de k números naturales menores o iguales que n. Denotamos como $A[\alpha|\beta]$ la submatriz $k\times k$ de A conteniendo las $\alpha_1,...,\alpha_k$ filas y las $\beta_1,...,\beta_k$ columnas (siendo $\alpha=(\alpha_1,...,\alpha_k), \beta=(\beta_1,...,\beta_k)$ dos sucesiones de $Q_{k,n}$). Si α=β denotaremos $A[\alpha]$:= $A[\alpha|\alpha]$ a la correspondiente submatriz principal.

El siguiente teorema muestra una fórmula clásica para los menores del producto de dos matrices.

Teorema 1.1. *Identidad de Cauchy-Binet para determinantes. Sean A, B matrices $n\times n$. Entonces:*

$$\det(AB)[\alpha|\beta] = \sum_{w\in Q_{k,n}} \det A[\alpha|w].\det B[w|\beta] \quad para \quad \alpha,\beta\in Q_{k,n}. \tag{1.1}$$

La demostración de este teorema se puede ver en la demostración de la fórmula (1.23) de [3].

Usando este mismo teorema, se puede deducir el siguiente resultado:

Corolario 1. *Si A y B son matrices TP $n\times n$, entonces AB es TP.*

Veamos ahora cómo son las inversas de las matrices TP. Definamos para ello la matriz $n\times n$ diagonal dada por

$$J_n := \begin{pmatrix} 1 & & & & & \\ & -1 & & & & \\ & & 1 & & & \\ & & & -1 & & \\ & & & & \ddots & \\ & & & & & (-1)^{n-1} \end{pmatrix}$$

El siguiente resultado es consecuencia del Teorema 3.3 de [3].

Teorema 1.2. *Si una matriz A $n\times n$ es TP no singular, entonces $J_nA^{-1}J_n$ es TP.*

Observemos que, si A es TP, por el teorema anterior A^{-1} tiene estructura de signos ajedrezada.

Otra propiedad interesante de las matrices TP no singulares es que sus menores principales son estrictamente positivos.

Teorema 1.3. *Si A es TP no singular entonces* $\det A[\alpha] > 0$ *para todo k y* $\alpha\in Q_{k,n}$.

Una demostración del teorema anterior se puede ver en el Corolario 3.8 de [3].

El Teorema 1.3 nos garantiza que, si A es TP no singular, entonces A admite una factorización LDU con L matriz triangular inferior con 1's en la diagonal principal, D matriz diagonal no singular y U matriz triangular superior con 1's en la diagonal principal. Además, se sabe que la factorización LDU es única.

Definición 5: Sea $D = (d_{ij})_{1\leq i,j\leq n}$ una matriz diagonal denotada como $D = diag(d_1,...,d_n)$ donde $d_i := d_{ii}$ para $i = 1,...,n$.

Usando esta notación, la matriz identidad $n \times n$ se expresa como $I_n = diag(1,...,1)$.

Denotamos por $E_i(x)$ con $i = 2,\ldots,n$ la matriz bidiagonal elemental inferior cuya entrada $(i,i-1)$ es x:

$$E_i(x) = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & x & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{pmatrix}. \tag{1.2}$$

La matriz $E_i^\top(x) = (E_i(x))^\top$ se llama matriz bidiagonal elemental superior. Las matrices $E_k(x)$ cumplen la propiedad:

$$E_i(x)E_j(y) = E_j(y)E_i(x), \tag{1.3}$$

salvo $|i-j| = 1$ con $xy \neq 0$.

También se cumple que

$$E_i(x)^{-1} = E_i(-x). \tag{1.4}$$

1.4. Eliminación de Neville y factorización bidiagonal

La *eliminación de Neville* (EN) es un procedimiento utilizado para hacer ceros en las distintas columnas bajo la diagonal principal de la matriz A. Es considerado un procedimiento alternativo a la eliminación Gaussiana en el que para hacer un cero en una fila añadimos a cada una de ellas un múltiplo de la anterior. Sin embargo, en la eliminación de Gauss se utiliza un mismo pivote para toda la columna. Este proceso es muy útil cuando trabajamos con ciertas clases de matrices como las TP.

Dada una matriz no singular $A = (a_{ij})_{1\leq i,j\leq n}$, la EN consiste en $n-1$ etapas llegando a la siguiente sucesión de matrices:

$$A := A^{(1)} \to \tilde{A}^{(1)} \to A^{(2)} \to \tilde{A}^{(2)} \to \ldots \to A^{(n)} \to \tilde{A}^{(n)} = U,$$

donde U es una matriz triangular superior.

La matriz $\tilde{A}^{(k)} = (\tilde{a}_{ij})^{(k)}_{1\leq i,j\leq n}$ se obtiene a partir de la matriz $A^{(k)} = (a_{ij})^{(k)}_{1\leq i,j\leq n}$ mediante una permutación de filas que traslada hacia abajo las filas con una entrada de ceros en la columna k-ésima por debajo de la diagonal principal.

Para matrices TP, siempre se puede realizar la eliminación de Neville sin cambios de filas. En el caso en el que no sea necesaria una permutación de filas en el paso k-ésimo, se obtiene que $\tilde{A}^{(k)} = A^{(k)}$. Por tanto, $A^{(k+1)} = (a_{ij})^{(k+1)}_{1\leq i,j\leq n}$ se obtiene a partir de $\tilde{A}^{(k)} = (\tilde{a}_{ij})^{(k)}_{1\leq i,j\leq n}$ usando la fórmula:

$$\tilde{a}_{ij}^{(k+1)} := \begin{cases} \tilde{a}_{ij}^{(k)} - \frac{\tilde{a}_{ik}^{(k)}}{\tilde{a}_{i-1,k}^{(k)}}\tilde{a}_{i-1,j}^{(k)} & \text{si } k \leq j < i \leq n \text{ y } \tilde{a}_{i-1,k}^{(k)} \neq 0, \\ \tilde{a}_{ij}^{(k)}, & \text{en otro caso.} \end{cases}$$

$\forall k = 1,\ldots,n-1$.

El pivote (i,j) de la eliminación de Neville se define de la siguiente manera: $p_{ij} := \tilde{a}_{ij}^{(j)}$ $1 \leq j \leq i \leq n$.

Si todos los pivotes son distintos de cero, se puede dar una expresión directa para calcularlos que usa entradas o cocientes de menores de la matriz (véase Lema 2.6 de [15]):

$$p_{i1} = \tilde{a}_{i1}, \quad 1 \leq i < n$$
$$p_{ij} = \frac{\det(A[i-j+1,\ldots,i|1,\ldots,j])}{\det(A[i-j+1,\ldots,i-1|1,\ldots,j-1])} \quad 1 \leq j \leq i \leq n.$$

Si $i = j$ se dice que p_{ii} es el pivote diagonal.
Se define también el multiplicador (i, j) de la eliminación de Neville, con $1 \leq j \leq i \leq n$ así:

$$m_{ij} := \begin{cases} \tilde{a}_{ij}^{(j)} = \frac{p_{ij}}{p_{i-1,j}}, & \text{si} \quad \tilde{a}_{i-1,j}^{(j)} \neq 0 \\ 0, & \text{si} \quad \tilde{a}_{i-1,j}^{(j)} = 0 \end{cases}$$

Los multiplicadores satisfacen que

$$m_{ij} = 0 \Rightarrow m_{hj} = 0 \quad \forall h > i.$$

Además, entre pivotes y multiplicadores se da la relación

$$p_{ij} = 0 \Leftrightarrow m_{ij} = 0.$$

La eliminación de Neville completa de una matriz A consiste en realizar la eliminación de Neville de A para obtener U, y a continuación proceder con la eliminación de Neville de $U^\top$, la traspuesta de U. El elemento pivote (i, j) (multiplicador) de la eliminación de Neville completa de A es el de la eliminación de Neville de A si $i \geq j$ y el elemento pivote (j, i) (multiplicador) de la eliminación de $U^\top$ si $j \geq i$.

En el caso de las matrices TP no singulares, estas pueden ser expresadas como un producto de matrices bidiagonales no negativas. El siguiente resultado corresponde al Teorema 4.2 de la página 120 de [16].

Teorema 1.4. *Sea $A = (a_{ij})_{1 \leq i,j \leq n}$ una matriz TP no singular. Entonces, A admite una factorización de la forma*

$$A = F_{n-1}F_{n-2}\ldots F_1 D G_1 \ldots G_{n-2}G_{n-1} \tag{1.5}$$

donde D es la matriz diagonal $diag(p_{11}\ldots p_{nn})$ con elementos diagonales mayores que cero y F_i, G_i son matrices bidiagonales no negativas dadas por

$$F_i = \begin{pmatrix} 1 & & & & & & & \\ 0 & 1 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & 0 & 1 & & & & \\ & & & m_{i+1,1} & 1 & & & \\ & & & & \ddots & \ddots & & \\ & & & & & m_{n,n-i} & 1 \end{pmatrix}, \tag{1.6}$$

$$G_i = \begin{pmatrix} 1 & 0 & & & & & & \\ & 1 & \ddots & & & & & \\ & & \ddots & 0 & & & & \\ & & & 1 & \tilde{m}_{i+1,1} & & & \\ & & & & 1 & \ddots & & \\ & & & & & \ddots & \tilde{m}_{n,n-i} \\ & & & & & & 1 \end{pmatrix}, \tag{1.7}$$

$\forall i \in \{1,\ldots,n-1\}$. *Además, las entradas m_{ij} y $\tilde{m}_{ij}$ satisfacen*

$$m_{ij} = 0 \Rightarrow m_{hj} = 0 \quad \forall h > i, \tag{1.8}$$

$$\tilde{m}_{ij} = 0 \Rightarrow \tilde{m}_{hj} = 0 \quad \forall h > i, \tag{1.9}$$

y la factorización es única.

En esta factorización, las entradas m_{ij} y p_{ij} son los multiplicadores y pivotes diagonales respectivamente correspondientes a la eliminación de Neville de A y las entradas $\tilde{m}_{ij}$ son los multiplicadores de la EN de $A^\top$ (que coinciden con los de la EN de $U^\top$).

El siguiente resultado muestra que la descomposición bidiagonal también caracteriza las matrices TP no singulares.

Teorema 1.5. *Una matrix $n \times n$ A no singular es TP si y solo si puede ser factorizada de la siguiente forma: D una matriz diagonal con entradas positivas, F_i, G_i dadas por* (1.6) *y* (1.7) *y las entradas m_{ij} y $\tilde{m}_{ij}$ números no negativos.*

Recordemos que un algoritmo puede ejecutarse con alta precisión relativa si solo incluye productos, divisiones, sumas de números del mismo signo y restas de datos iniciales. Siendo A una matriz bidiagonal TP no singular, Plamen Koev ([19]) diseña algoritmos eficientes para calcular, con alta precisión relativa, los valores propios, los valores singulares y la inversa de A, así como la solución a sistemas de ecuaciones lineales $Ax = b$ cuando b presenta signos alternados (véase apartado 1.5). Su punto de partida es la factorización bidiagonal (1.5) obtenida con HRA.

La notación $\mathscr{BD}(A)$ se introduce para referirnos a la factorización bidiagonal descrita en el Teorema 1.4

$$(\mathscr{BD}(A))_{ij} = \begin{cases} m_{ij}, & \text{si } i > j, \\ \tilde{m}_{ij}, & \text{si } i < j, \\ p_{ii}, & \text{si } i = j. \end{cases} \tag{1.10}$$

Observemos que si A es una matriz TP, $A^\top$ también es una matriz TP. Transponiendo la fórmula (1.5) del Teorema 1.4 obtenemos la factorización bidiagonal única de $A^\top$:

$$A^\top = G_{n-1}^\top \ldots G_1^\top D F_1^\top \ldots F_{n-1}^\top$$

donde F_i y G_i, $i \in \{1, \ldots, n-1\}$, son las matrices bidiagonales inferiores y superiores triangulares no negativas dadas por (1.6) y (1.7) respectivamente.

También se satisface que

$$\mathscr{BD}(A^\top) = \mathscr{BD}(A)^\top.$$

El siguiente resultado, que corresponde al Corolario 2 de [8], proporciona la factorización bidiagonal de la inversa de una matriz TP triangular inferior A en términos de $\mathscr{BD}(A)$, siempre que los multiplicadores de la EN de A no sean nulos.

Teorema 1.6. *Sea $A = (a_{ij})_{1 \le i,j \le n}$ una matriz triangular inferior TP tal que*

$$(\mathscr{BD}(A))_{ij} = \begin{cases} m_{ij} > 0, & si\ i > j, \\ 0, & si\ i < j, \\ 1, & si\ i = j. \end{cases} \tag{1.11}$$

Entonces la factorización bidiagonal de su inversa está dada por

$$(\mathscr{BD}(A^{-1}))_{ij} = \begin{cases} -m_{i,i-j}, & si\ i > j, \\ 0, & si\ i < j, \\ 1, & si\ i = j. \end{cases} \tag{1.12}$$

Demostración. Como, en la factorización única (1.5), D y G_i para $i = 1, \cdots, n$ son iguales a la matriz identidad $n \times n$ I_n, podemos usar (1.5) y (1.11) para factorizar la matriz A de esta forma:

$$A = F_{n-1} \cdots F_1 = \{E_n(m_{n,1})\}\{E_{n-1}(m_{n-1,1})E_n(m_{n,2})\} \cdots \{E_2(m_{2,1}) \cdots E_n(m_{n,n-1})\}.$$

Como consecuencia directa, A^{-1} puede ser escrita así:

$$\begin{aligned} A^{-1} =& \{E_n(-m_{n,n-1}) \cdots E_2(-m_{2,1})\}\{E_n(-m_{n,n-2}) \cdots E_3(-m_{3,1})\} \cdots \\ & \{E_n(-m_{n,2})E_{n-1}(-m_{n-1,1})\}\{E_n(-m_{n,1})\}. \end{aligned} \tag{1.13}$$

Usando (1.3) podemos reescribir (1.13) haciendo una permutación de las matrices $E_i(x)$:

$$\begin{aligned} A^{-1} =& \{E_n(-m_{n,n-1}) \cdots E_3(-m_{3,2})\}\{E_n(-m_{n,n-2}) \cdots E_4(-m_{4,2})\} \cdots \\ & \{E_n(-m_{n,2})\}\{E_2(-m_{2,1})E_3(-m_{3,1}) \cdots E_{n-1}(-m_{n-1,1})E_n(-m_{n,1})\}. \end{aligned} \tag{1.14}$$

La matriz $E_2(-m_{2,1}) \cdots E_n(-m_{n,1})$ será el primer factor de $\mathscr{BD}(A^{-1})$. Continuando con esta argumentación podemos seguir reordenando las matrices $E_i(x)$ de (1.13) hasta obtener (1.12).

□

Un resultado análogo es cierto para matrices TP triangulares superiores.

Corolario 2. *Sea A una matriz TP triangular superior tal que*

$$(\mathscr{BD}(A))_{ij} = \begin{cases} 0, & \text{si } i > j, \\ \tilde{m}_{j,i} > 0, & \text{si } i < j, \\ 1, & \text{si } i = j. \end{cases} \tag{1.15}$$

Entonces la factorización bidiagonal de su invera está dada por

$$(\mathscr{BD}(A^{-1}))_{ij} = \begin{cases} 0, & \text{si } i > j, \\ -\tilde{m}_{i-j,i}, & \text{si } i < j, \\ 1, & \text{si } i = j. \end{cases} \tag{1.16}$$

Demostración. Aplicar el Teorema 1.6 a $A^\top$.

□

El ejemplo siguiente muestra que la positividad estricta de los multiplicadores en el Teorema 1.6 (o análogamente en el Corolario 2) es necesaria, es decir, que si la matriz triangular tiene multiplicadores nulos, entonces el teorema no es cierto.

Ejemplo 1. *Sea $A = (a_{ij})_{1 \le i,j \le n}$ la matriz triangular inferior:*

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 3 & 1 \end{pmatrix}.$$

Aplicando la EN y usando (1.11) *podemos ver que su factorización bidiagonal es:*

$$\mathscr{BD}(A) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 1 \end{pmatrix},$$

lo que significa que

$$A = E_4(1)E_3(1)E_2(1)E_4(2), \tag{1.17}$$

y que hay dos multiplicadores nulos.

A partir de (1.17) *se deduce que*

$$A^{-1} = (E_4(1)E_3(1)E_2(1)E_4(2))^{-1} = E_4(-2)E_2(-1)E_3(-1)E_4(-1), \tag{1.18}$$

o usando la notación (1.10)*,*

$$\mathscr{BD}(A^{-1}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -2 & -1 & 1 \end{pmatrix}.$$

Por lo tanto, es necesario requerir que los multiplicadores no sean nulos, ya que si no $\mathscr{BD}(A^{-1})$ *no satisface* (1.12)*.*

1.5. Operaciones con alta precisión relativa para matrices totalmente positivas

Consideramos el problema de realizar cálculos precisos con matrices totalmente positivas no singulares $(n \times n)$. Estas matrices tienen la propiedad de tener una representación única como productos de matrices bidiagonales positivas. Dada esa representación, se puede calcular de forma precisa y eficiente la matriz inversa, la factorización LDU, los valores propios y la factorización de valores singulares (SVD) de una matriz totalmente positiva con alta precisión relativa con los algoritmos de [19].

Por precisión entendemos que cada cantidad debe calcularse con alta precisión relativa, con un signo y dígitos principales correctos. Por eficiencia nos referimos a realizar estos cálculos en un tiempo máximo de $O(n^3)$. Presentamos así algoritmos precisos y eficientes que realizan algunos de estos cálculos.

Recordemos en primer lugar la fuente de grandes errores relativos en los algoritmos de matrices convencionales. La precisión relativa en estos algoritmos se pierde debido a la cancelación sustractiva en la resta de cantidades aproximadamente del mismo signo. Por el contrario, la precisión relativa se conserva en la multiplicación, división, suma e incluso en las raíces cuadradas y, si acaso, en restas de datos iniciales.

Veamos algunos cálculos precisos de matrices conocida $\mathscr{BD}(A)$. Las entradas de $\mathscr{BD}(A)$ nos permiten calcular con precisión las entradas de la descomposición LDU, las entradas de la matriz inversa y la resolución de algunos sistemas de ecuaciones lineales. Además, aparte de estos cálculos, dada $\mathscr{BD}(A)$, se pueden realizar de manera precisa y eficiente muchas operaciones con A como matriz.

Supongamos que partimos de la factorización (1.5) descrita en el Teorema 1.4. Recordemos que los m_{ij} y los $\tilde{m}_{ij}$ son los multiplicadores de la eliminación de Neville de A y $A^\top$, respectivamente. Vamos a expresar esta factorización en términos de las matrices bidiagonales elementales $E_i(x)$ descritas en (1.2).

En [19], para expresar esta factorización (1.5), se introduce la notación $\prod_1^{k=n-1}$, que indica que el producto comienza en $k = n-1$ y el índice se reduce gradualmente hasta llegar a 1. Utilizando esta notación, podemos expresar el proceso de dicha factorización como sigue:

$$F_i = \prod_{j=i+1}^{n} E_j(m_{j,j-i}), \quad G_i = \prod_{i+1}^{j=n} E_j^T(\tilde{m}_{j,j-i}).$$

Sea ahora A una matriz cuadrada no singular $n \times n$. Vamos a calcular su factorización LDU con HRA sustituyendo en la fórmula (1.5) los factores F_i y G_i por los productos anteriores:

$$A = \left(\prod_{1}^{i=n-1} \prod_{j=i+1}^{n} E_j(m_{j,j-i}) \right) \cdot D \cdot \left(\prod_{i=1}^{n-1} \prod_{i+1}^{j=n} E_j^T(\tilde{m}_{j,j-i}) \right). \tag{1.19}$$

Teniendo en cuenta que las matrices $E_i(x)$ son triangulares inferiores y la unicidad de la descomposición LDU de una matriz no singular, llegamos a

$$L = \prod_{1}^{i=n-1} \prod_{j=i+1}^{n} E_j(m_{j,j-i}), \quad U = \prod_{i=1}^{n-1} \prod_{i+1}^{j=n} E_j^T(\tilde{m}_{j,j-i}).$$

Usando que, por el Teorema 1.4, los multiplicadores m_{ij}, $\tilde{m}_{ij}$ son no negativos concluimos que L y U se pueden calcular con HRA.

También podemos calcular la matriz inversa con alta precisión relativa. Invirtiendo la fórmula (1.19) y usando la propiedad (1.4) obtenemos:

$$A^{-1} = \left(\prod_{1}^{i=n-1} \prod_{j=i+1}^{n} E_j^T(-\tilde{m}_{j,j-i}) \right) \cdot D^{-1} \cdot \left(\prod_{i=1}^{n-1} \prod_{i+1}^{j=n} E_j(-m_{j,j-i}) \right). \tag{1.20}$$

Podemos obtener A^{-1} multiplicando su expresión en un tiempo de $O(n^3)$. Cada entrada de A^{-1} se calculará con alta precisión relativa ya que las multiplicaciones de las matrices de (1.20) no conllevan restas. Tengamos en cuenta para ello que el producto $E_i(x)M$ (con M una matriz) da lugar a sumar a la fila i-ésima de M la fila anterior multiplicada por x.

Veamos ahora la resolución de un sistema matricial $Ax = b$ con A TP no singular y suponiendo que las componentes del vector b tienen signos alternados (es decir, $\text{sign} b_i = (-1)^i$ o $\text{sign} b_i = (-1)^{i-1}$), la podemos realizar con HRA.

Podemos usar (1.20) para calcular la solución de $Ax = b$ en un tiempo de $O(n^2)$ multiplicando la expresión

$$x = A^{-1}b = \left(\prod_{1}^{i=n-1} \prod_{j=i+1}^{n} E_j^T(-\tilde{m}_{j,j-i}) \right) D^{-1} \left(\prod_{i=1}^{n-1} \prod_{i+1}^{j=n} E_j(-m_{j,j-i}) \right) b \tag{1.21}$$

de derecha a izquierda. Entonces (1.21) no conlleva cancelación sustractiva, y cada componente de x se calcula con alta precisión relativa. Observemos que, por el Teorema 1.4, todos los multiplicadores m_{ij} y $\tilde{m}_{ij}$ son no negativos, y como b tiene los signos alternados, todos los vectores intermedios que vamos obteniendo siguen teniendo los signos alternados y no se realiza ninguna resta, por lo que podemos garantizar la HRA.

Como hemos mencionado anteriormente, con la factorización bidiagonal de una matriz también podemos realizar otros cálculos importantes como los valores singulares o los valores propios entre otros. Para ello, como se ve en [19], es esencial realizar estas operaciones como una combinación de las siguientes transformaciones elementales $EETS$ (*elementary elimination transformations*):

1. EET1: restar un múltiplo de una fila (o columna) a la siguiente con el objetivo de hacer un cero de tal manera que la matriz transformada sigue siendo TP.

2. EET2: añadir un múltiplo de una fila (o columna) a la anterior.

3. EET3: añadir un múltiplo de una fila (o columna) a la siguiente.

4. EET4: multiplicar por una matriz diagonal positiva.

Observemos que llevar a cabo cualquiera de estas operaciones con una matriz TP resulta en otra matriz TP ([16]). El enfoque consistirá en no realizar las operaciones directamente sobre la matriz, sino en aplicarlas de manera implícita, transformando los parámetros de la descomposición bidiagonal y realizando los cálculos necesarios para evitar restas.

En [19] se puede encontrar cómo se llevan a cabo las EETs. La $EET1$ es la más sencilla, y realizar esta operación equivale a sustituir la componente correspondiente de $\mathscr{BD}(A)$ por cero. Las $EET2$ y $EET3$ requieren un cuidado especial.

A continuación, nos vamos a centrar en las operaciones $EET3$ y $EET4$. Sea la matriz TP C obtenida a partir de la matriz TP A de dimensiones $n \times n$ aplicando una EET a A. Veamos cómo, dada $\mathscr{BD}(A)$, se puede calcular la factorización $\mathscr{BD}(C)$ sin realizar restas.

La operación $EET3$ consistía en añadir un múltiplo de una fila (o columna) a la siguiente: sea A una matriz TP y C se obtiene a partir de A sumando un múltiplo de la fila $i-1$ de A a la fila i:

$$C = E_i(x)A, \quad x > 0.$$

Queremos ver cómo calcular con precisión $\mathscr{BD}(C)$, con x y $\mathscr{BD}(A)$ conocidos. El siguiente lema muestra cómo calcular la factorización bidiagonal del producto de dos matrices bidiagonales inferiores. Es la base del teorema posterior.

Lema 1. *Sean B y C matrices bidiagonales inferiores tales que sus elementos en la diagonal principal son todos unos, los elementos de la subdiagonal son no negativos ($b_i \geq 0$ y $c_i \geq 0$ para $i = 1,\ldots,n-1$) y también se cumple que $b_i = 0$ cuando $c_{i-1} = 0$. Entonces existen matrices bidiagonales B' y C' con elementos extradiagonales $b'_i \geq 0$ y $c'_i \geq 0$ con $i = 1,\ldots,n-1$ tales que $B'C' = BC$ y $b'_1 = 0$. Además, se pueden calcular b'_i y c'_i sin realizar restas en como mucho $4n$ operaciones elementales.*

Demostración. Comparamos las entradas a ambos lados de $B'C' = BC$,

$$\begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ & b'_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & b'_{n-1} & 1 \end{pmatrix} \begin{pmatrix} 1 & & & & \\ c'_1 & 1 & & & \\ & c'_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & c'_{n-1} & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ b_1 & 1 & & & \\ & b_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & b_{n-1} & 1 \end{pmatrix} \begin{pmatrix} 1 & & & & \\ c_1 & 1 & & & \\ & c_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & c_{n-1} & 1 \end{pmatrix},$$

y obtenemos que

$$\begin{cases} c'_1 &= b_1 + c_1, \\ b'_i &= \frac{b_i c_{i-1}}{c'_{i-1}}, \\ c'_i &= b_i + c_i - b'_i, \end{cases} \tag{1.22}$$

para $i = 2,3,\ldots,\text{mín}\{j | b_j = 0\}$, y $b'_i = b_i$, $c'_i = c_i$, en otro caso. Para evitar la resta de (1.22) definimos las variables auxiliares $d_i := b_i - b'_i$. De esta forma $d_1 = b_1 - b'_1 = b_1$ y

$$d_i = b_i - b'_i = b_i - \frac{b_i c_{i-1}}{c'_{i-1}} = \frac{b_i}{c'_{i-1}}(c'_{i-1} - c_{i-1}) = \frac{b_i}{c'_{i-1}}(b_{i-1} - b'_{i-1}) = \frac{b_i d_{i-1}}{c'_{i-1}}, \quad i = 2,\cdots,n-1.$$

La versión libre de restas (y por tanto, precisa) de (1.22) es

$$\begin{cases} c'_i &= c_i + d_i, \\ b'_i &= \frac{b_i c_{i-1}}{c'_{i-1}}, \\ d'_i &= \frac{b_i d_{i-1}}{c'_{i-1}}. \end{cases} \tag{1.23}$$

Este cálculo evidentemente no cuesta más de $4n$ operaciones aritméticas. Como $c'_i = 0$ implica que $b'_{i+1} = 0$, el producto $B'C'$ es $\mathscr{BD}(BC)$, ya que cumple la condición del Teorema 1.4. □

Implementamos el procedimiento del Lema 1 en el algoritmo 1 libre de restas. Reescribimos d_i y d_{i-1}, y los arrays b y c por b' y c', respectivamente. La cantidad $e = \frac{b_{i+1}}{c'_i}$ se calcula solo una vez y se utiliza para actualizar ambas, ahorrando así una división.

Algorithm 1 EET3

function DQD2(b,c)

$t \leftarrow c_1$

$c_1 \leftarrow b_1 + c_1$

$d \leftarrow b_1$

$b_1 \leftarrow 0$

$i \leftarrow 1$

while $(i < length(b))$ ***and*** $(b_{i+1} > 0)$ **do**

$e \leftarrow \frac{b_{i+1}}{c_i}$

$d \leftarrow e \cdot d$

$b_{i+1} \leftarrow e \cdot t$

$t \leftarrow c_{i+1}$

$c_{i+1} \leftarrow c_{i+1} + d$

$i \leftarrow i+1$

return b, c, i

end function

El algoritmo 1 usará el vector de parámetros b de B y el c de C.

El siguiente resultado ya prueba que la $EET3$ se puede realizar eficientemente y con alta precisión relativa.

Teorema 1.7. *Sea A una matriz TP $n \times n$ no singular. Dado $x > 0$ y $\mathscr{BD}(A)$, la factorización $\mathscr{BD}(E_i(x)A)$ puede ser calculada sin usar ninguna resta en como máximo $4n$ operaciones aritméticas.*

Demostración. Sea $\mathscr{BD}(A)$. Por el Teorema 1.4 está dada por

$$A = F_{n-1}F_{n-2}\cdots F_1 D G_1 \cdots G_{n-2}G_{n-1},$$

y sea $F = F_{n-1}F_{n-2}\cdots F_1$. La matriz $E_i(x)F$ es una matriz triangular inferior unitaria (con 1's en la diagonal) TP (por el Corolario 1) ya que es producto de matrices TP. Por tanto, de nuevo por el Teorema 1.4, posee factorización bidiagonal $\mathscr{BD}(E_i(x)\cdot F)$ y cumplirá la expresión:

$$E_i(x)F = L_{n-1}L_{n-2}\cdots L_1,$$

donde las matrices L_i son matrices bidiagonales triangulares inferiores. Entonces $\mathscr{BD}(E_i(x)A)$ satisfará, por la unicidad de la factorización bidiagonal (Teorema 1.4), que:

$$E_i(x)A = L_{n-1}L_{n-2}\cdots L_1 D G_1 \cdots G_{n-2}G_{n-1},$$

con lo que nos bastará hallar la factorización bidiagonal de $E_i(x)F$ para obtener la de $E_i(x)A$. Usando el Lema 1, vamos propagando $E_i(x)$ a través de los factores F_i del modo siguiente:

$$\begin{aligned} E_i(x)F &= E_i(x)F_{n-1}F_{n-2}\cdots F_1 \\ &= L_{n-1}E_{i_1}(x_1)F_{n-2}\cdots F_1 \\ &= L_{n-1}L_{n-2}E_{i_2}(x_2)\cdots F_1 \\ &\cdots \\ &= L_{n-1}L_{n-2}\cdots L_1, \end{aligned}$$

(denotando $E_{i_0}(x_0) := E_i(x)$). Empezamos con $k = 1$ y repetimos el siguiente proceso. Aplicamos el algoritmo 1 a las submatrices principales restantes de $E_{i_{k-1}}(x_{k-1})$ y de F_{n-k} que consisten en filas y columnas desde $i_{k-1} - 1$ hasta n. La única componente no nula de $E_{i_{k-1}}(x_{k-1})$ desaparece y obtenemos una nueva matriz $\overline{L}_{n-k} = E_{i_{k-1}}(x_{k-1})L_{n-k}$. Denotemos con $f_j^{(k)}$ al parámetro j-ésimo de la matriz F_j.

Si se cumplen algunas de estas condiciones:

1. $k = n-1$, o

2. no se introdujeron elementos no nulos en $\overline{L}_{n-k}$ que no estuvieran en L_{n-k}, o

3. se introdujo un elemento no nulo $\overline{l}_j^{(n-k)}$ en $\overline{L}_{n-k}$, pero $f_{j-1}^{(n-k-1)} \neq 0$,

entonces estableceremos $\overline{L}_{n-k} = L_{n-k}$ y habríamos terminado de propagar $E_i(x)$. En otro caso (un valor no nulo $\overline{l}_j^{(n-k)}$ se introduce en $\overline{L}_{n-k}$, y $f_{j-1}^{(n-k-1)} = 0$, con $k < n-1$), tenemos entonces que $\overline{L}_{n-k} = L_{n-k}E_{i_k}(x_k)$, donde L_{n-k} tiene la misma estructura de elementos no nulos que F_{n-k}. Fijamos $i_k = j$, $x_k = f_j^{(k)}$, aumentamos k en una unidad y repetimos el mismo proceso.

El cálculo de $\mathscr{BD}(E_i(x)A)$ se realiza sin utilizar restas. A lo sumo se modifican $2n-3$ entradas en $\mathscr{BD}(A)$ con no más de dos operaciones aritméticas por entrada (como se ve en el algoritmo 1 o en el Lema 1). El coste total, por lo tanto, no supera 4n.

□

El teorema anterior también se puede utilizar para conocer la $\mathscr{BD}(AB)$ a partir de $\mathscr{BD}(A)$ y $\mathscr{BD}(B)$ sin usar restas (véase [19]).

Por último, veamos $EET4$ que consistía en multiplicar por una matriz bidiagonal positiva. El producto de una matriz diagonal $F = \mathrm{diag}(f_1,\ldots,f_n)$, $f_i > 0$, $i = 1,2,\ldots,n$, y una matriz TP A de $n \times n$ es TP. Ahora mostramos cómo calcular $\mathscr{BD}(F)$, dado F y $\mathscr{BD}(A)$. Propagamos F a través de los factores F_i en $\mathscr{BD}(A)$ utilizando

$$\begin{pmatrix} f_1 & & & \\ & f_2 & & \\ & & \ddots & \\ & & & f_m \end{pmatrix} \begin{pmatrix} 1 & & & \\ c_1 & 1 & & \\ & \ddots & \ddots & \\ & & c_{m-1} & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & \\ b_1 & 1 & & \\ & \ddots & \ddots & \\ & & b_{m-1} & 1 \end{pmatrix} \begin{pmatrix} f_1 & & & \\ & f_2 & & \\ & & \ddots & \\ & & & f_m \end{pmatrix},$$

donde $b_i = \frac{c_i f_{i+1}}{f_i}$, $i = 1,2,\cdots,n-1$.

Dada $B = \mathscr{BD}(A)$ y el vector $(f_1, f_2, \cdots, f_n)$, el siguiente algoritmo calcula la factorización bidiagonal $\mathscr{BD}(\mathrm{diag}(f_1, f_2, \cdots, f_n)\cdot A)$ usando solo multiplicaciones y divisiones en un tiempo máximo $2n^2$.

Algorithm 2 EET4

function TNDIAGONALSCALE(f, B)

$b_{11} \leftarrow b_{11} \cdot f_1$ **for** $i = 2$ *to* m **do**

$i \leq n$

$b_{ii} \leftarrow b_{ii} \cdot f_i$

$b_{i,1:\text{mín}(i-1,n)} \leftarrow b_{i,1:\text{mín}(i-1,n)} \cdot \frac{f_i}{f_{i-1}}$

return B

end function

Capítulo 2

Matrices de Pascal, generalizaciones y alta precisión relativa

2.1. Introducción

En este capítulo hablaremos de las matrices de Pascal, que son matrices totalmente positivas que presentan importantes aplicaciones en el campo del diseño e imagen así como en probabilidad, combinatoria, análisis numérico e ingeniería eléctrica (véase [3], [14] y [21]).

Se sabe que las matrices de Pascal están mal condicionadas (véase ([2]) siendo incluso peor condicionadas que las matrices de Vandermonde. A pesar de este hecho, mostraremos que se pueden obtener algoritmos con alta precisión relativa (HRA) para el cálculo de valores propios e inversas de matrices de Pascal, así como para resolver ciertos sistemas lineales cuyas matrices de coeficientes son matrices de Pascal.

Para ello, por un lado necesitaremos una factorización bidiagonal exclusiva para las matrices de Pascal (estrechamente relacionada con el procedimiento de la eliminación de Neville). Por otro lado, algoritmos HRA para matrices TP (ya comentados en el último apartado del capítulo anterior).

En este capítulo comenzaremos viendo en el apartado 2.2 la factorización bidiagonal de las matrices de Pascal, que es extraordinariamente simple y permite de manera trivial garantizar la alta precisión relativa de los cálculos algebraicos mencionados en el capítulo anterior. Después, en el apartado 2.3, definiremos algunas matrices de Pascal generalizadas y algunas notaciones que necesitaremos más adelante. Por último, en el apartado 2.4, describiremos la factorización bidiagonal de este tipo de matrices.

2.2. Matrices de Pascal y su factorización bidiagonal

Comenzamos introduciendo las definiciones básicas de las matrices de Pascal.

Definición 6: Una *matriz de Pascal* de orden n es la matriz simétrica

$$P = (p_{ij})_{1\leq i,j\leq n}; \quad p_{ij} := \binom{i+j-2}{j-1}. \tag{2.1}$$

Definición 7: Una *matriz triangular inferior de Pascal* de orden n es la matriz triangular inferior

$$P_L = (q_{ij})_{1\leq i,j\leq n}; \quad q_{ij} := \binom{i-1}{j-1}. \tag{2.2}$$

Esta matriz P_L es el factor de la factorización de Cholesky de la matriz de Pascal P:

$$P = P_L P_L^\top. \tag{2.3}$$

El siguiente Lema (que corresponde al Lema 1 de ([2])) nos proporciona una factorización bidiagonal de la matriz triangular inferior de Pascal. Observemos que, en este caso, no obtenemos la factorización bidiagonal mediante la eliminación de Neville.

Lema 2. *La matriz triangular inferior de Pascal dada por* (2.2) *satisface*

$$P_L = \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & 0 & 1 & \\ & & & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & 1 & 1 & \\ & & & 1 & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & & & & \\ 1 & 1 & & & \\ & \ddots & \ddots & & \\ & & 1 & 1 & \\ & & & 1 & 1 \end{pmatrix} \tag{2.4}$$

Demostración. Sea F_i (con $1 \leq i \leq n-1$) la matriz $n \times n$

$$F_i = \begin{pmatrix} 1 & & & & & & \\ 0 & 1 & & & & & \\ & \ddots & \ddots & & & & \\ & & 0 & 1 & & & \\ & & & 1 & 1 & & \\ & & & & \ddots & \ddots & \\ & & & & & 1 & 1 \end{pmatrix} \xleftarrow{(n-i+1)\text{- ésima fila}}$$

y sea ${P_L}^{(k)} = (q_{ij})^{(k)}_{1 \leq i,j \leq n}$ la matriz ${P_L}^{(k)} = F_1 F_2 ... F_k$. Vamos a probar por inducción sobre k que

$$q^{(k)}_{ij} = q^{(k)}_{i-1,j-1} + q^{(k)}_{i-1,j}, \quad 1 \leq i \leq n, \quad n-k \leq j \leq n \tag{2.5}$$

definiendo $q^{(k)}_{i0} = q^{(k)}_{0j} = 0$, con $1 \leq i, j \leq n$ y $q^{(k)}_{00} = 1$. Como ${P_L}^{(k)}$ es una matriz triangular inferior con $1's$ en la diagonal, se cumple (2.5) $\forall k$ con $i \leq j$, y en particular para la última columna de ${P_L}^{(k)}$.

Por lo tanto, solo tenemos que probar que (2.5) es válido para $i > j$ con $j = n-k, n-k+1, ..., n-1$.

Para $k = 1$ es obvio. Suponiendo que (2.5) se cumple para $k-1$ vamos a probar que se satiface para todo k.

Notemos que ${P_L}^{(k)}$ se puede obtener a partir de ${P_L}^{(k-1)}$ añadiendo a cada una de las columnas $n-k, n-k+1, ..., n-1$ la siguiente.

Si $n-k \leq j \leq n-1$,

$$q^{(k)}_{ij} = q^{(k-1)}_{ij} + q^{(k)}_{i,j+1}, \quad \forall i \tag{2.6}$$

y por la hipótesis de inducción, para $i < n$, se tiene

$$q^{(k)}_{ij} = q^{(k-1)}_{i+1,j+1}, \quad n-k \leq j \leq n-1. \tag{2.7}$$

Si $n-k \leq j \leq n-1$ y $i \leq n$, a partir de (2.6), (2.7) se llega a (2.5).

Teniendo en cuenta (2.5) para $k = n-1$ vamos a probar que

$$q^{(n-1)}_{ij} = (F_1 F_2 \ldots F_{n-1})_{ij} = \binom{i-1}{j-1}, \quad \text{si } i \geq j,$$

es decir q_{ij} de (2.2).

Lo probaremos por inducción en las filas i de ${P_L}^{(n-1)}$ para $i = 1, ..., n$. Esta condición se satisface claramente para la primera fila. Suponiendo que es válida para $1, ..., i-1$ vamos a probar que también se cumple para i:

Si $1 < j < i$, a partir de (2.5) tenemos por hipótesis de inducción

$$q^{(n-1)}_{ij} = \binom{i-2}{j-2} + \binom{i-2}{j-1} = \binom{i-1}{j-1}.$$

Si $j = 1$, entonces tenemos por hipótesis de inducción, $q^{(n-1)}_{i1} = q^{(n-1)}_{i-1,1} = \binom{i-2}{0} = 1 = \binom{i-1}{0}$. Por último, si $i = j$, $q^{(n-1)}_{ii} = 1$ y con esto se termina la demostración. □

Si juntamos las expresiones de las factorizaciones (2.3) y (2.4) obtenemos la factorización bidiagonal de la matriz P que llamaremos

$$\mathscr{BD}(P) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}.$$

En este caso, evidentemente, la factorización bidiagonal de la matriz de Pascal se obtiene con HRA al ser todas sus entradas 1's. Además, todas las operaciones comentadas en el Capítulo 1 se van a poder realizar con alta precisión relativa ya que tenemos $\mathscr{BD}(A)$ con HRA. En particular, los valores propios y la inversa de P. Por otro lado, esta factorización también puede utilizarse como prueba de la total positividad de una matriz de Pascal.

Corolario 3. *Una matriz de Pascal P es TP.*

Demostración. Por la fórmula (2.4) y el Corolario 1 del Capítulo 1, P_L es TP ya que es producto de matrices TP puesto que las matrices bidiagonales no negativas son claramente TP. Por la fórmula (2.3) y de nuevo por el Corolario 1 tenemos que P también es TP por ser producto de matrices TP. □

2.3. Matrices de Pascal generalizadas

Comenzamos este apartado con una primera generalización de las matrices triangulares de Pascal y de las simétricas de Pascal ([7], [4] y [23]).

Definición 8: Para cualquier número real x, la *matriz triangular de Pascal generalizada* de primer tipo, $P_n[x]$ se define como la $(n+1)\times(n+1)$ matriz triangular inferior con 1's en la diagonal principal y

$$(P_n[x])_{ij} := x^{i-j}\binom{i-1}{j-1}, \quad 1 \le j \le i \le n+1$$

y la *matriz simétrica generalizada de Pascal* $(n+1)\times(n+1)$ $R_n[x]$ viene dada por

$$(R_n[x])_{ij} := x^{i+j-2}\binom{i+j-2}{j-1}, \quad 1 \le i,j \le n+1.$$

La definición anterior también se puede generalizar de la siguiente manera involucrando dos variables (véase definición 3 de [7], [4] y [23]).

Definición 9: Para $x, y \in \mathbb{R}$ se define la $(n+1)\times(n+1)$ matriz $R_n[x,y]$

$$(R_n[x,y])_{ij} := x^{j-1}y^{i-1}\binom{i+j-2}{j-1}, \quad 1 \le i,j \le n+1.$$

Observar que $R_n[x] = R_n[x,x]$, por lo que $P_n[1]$ es la matriz triangular inferior de Pascal y $R_n[1]$ es la matriz simétrica de Pascal.

Otra posible extensión de la definición 8 es la siguiente:

Definición 10: Sean x y λ dos números reales y n un entero no negativo. Definimos la notación $x^{n|\lambda}$ así:

$$x^{n|\lambda} := \begin{cases} x(x+\lambda)\dots(x+(n-1)\lambda), & \text{si } n > 0, \\ 1, & \text{si } n = 0. \end{cases} \tag{2.8}$$

La matriz de Pascal triangular inferior generalizada $P_{n,\lambda}$ viene dada por

$$(P_{n,\lambda}[x])_{ij} := x^{(i-j)|\lambda}\binom{i-1}{j-1}, \quad 1 \le j \le i \le n+1 \tag{2.9}$$

donde n es un número natural y λ y x son números reales.

Observamos que con el caso particular $\lambda = 0$ llegamos a la matriz de Pascal generalizada de primer tipo $P_{n,0}[x] = P_n[x]$.

La definición anterior de $P_{n,\lambda}[x]$ se puede generalizar al caso de dos variables x e y de la siguiente forma:

Definición 11: Sea $P_{n,\lambda}[x,y]$ la matriz dada por

$$(P_{n,\lambda}[x,y])_{i,j} := x^{(i-1)\lambda} y^{(j-1)\lambda} \binom{i-1}{j-1}. \tag{2.10}$$

Observemos que $P_n[x,y] := P_{n,0}[x,y]$.

2.4. Factorizaciones bidiagonales de las matrices de Pascal generalizadas

El resultado siguiente proporciona la factorización bidiagonal de la matriz de Pascal generalizada $P_{n,\lambda}[x]$.

Teorema 2.1. *Sean $x, \lambda \in \mathbb{R}$ y $n \in \mathbb{N}$ y $P_{n,\lambda}[x]$ la $(n+1)\times(n+1)$ matriz triangular inferior dada por* (2.9).

(I) Si $x \neq k\lambda$ para $k = -n+1, ..., 0, ..., n-1$ se tiene que

$$(\mathscr{BD}(P_{n,\lambda}[x]))_{ij} = \begin{cases} 1, & si\ i = j, \\ x+(i-2j)\lambda, & si\ i > j, \\ 0, & si\ i < j. \end{cases} \tag{2.11}$$

(II) Si $x = k\lambda$ para algunos $k \in \{0, \ldots, n-1\}$, se tiene que

$$(\mathscr{BD}(P_{n,\lambda}[x]))_{ij} = \begin{cases} 1, & si\ i = j, \\ x+(i-2j)\lambda, & si\ i > j, j \leq k, \\ 0, & en\ otro\ caso\ . \end{cases} \tag{2.12}$$

(III) Si $x = -k\lambda$ para algunos $k \in \{0, \ldots, n-1\}$, se tiene que

$$(\mathscr{BD}(P_{n,\lambda}[x]))_{ij} = \begin{cases} 1, & si\ i = j, \\ x+(i-2j)\lambda, & si\ 0 \leq i-j \leq k, \\ 0, & en\ otro\ caso\ . \end{cases} \tag{2.13}$$

Demostración. Supongamos en primer lugar que $x \neq k\lambda$ para $k = -n+1, ..., 0, ..., n-1$. Utilizamos el primer paso de la eliminación de Neville de $A = (a_{ij})_{1 \leq i,j \leq n+1}$, donde $a_{ij} := (P_{n,\lambda}[x])_{i,j}$ para $i, j = 1, ..., n+1$:

$$a_{ij}^{(2)} = a_{ij} - \frac{a_{i1}}{a_{i-1,1}} a_{i-1,j} = a_{ij} - (x+(i-2)\lambda) a_{i-1,j}, \quad i > j \geq 1.$$

Aplicando (2.8) a la fórmula anterior obtenemos

$$a_{ij}^{(2)} = x^{(i-j)|\lambda} \binom{i-1}{j-1} - (x+(i-2)\lambda) x^{(i-j-1)|\lambda} \binom{i-2}{j-1}.$$

Por la fórmula (2.9), tenemos que

$$a_{ij}^{(2)} = ((x+i-j-1)\lambda) \binom{i-1}{j-1} - (x+(i-2)\lambda) \binom{i-2}{j-2}) x^{(i-j-1)|\lambda} =$$

$$\left(x\binom{i-2}{j-2}+\frac{(i-j-1)(i-1)!}{(j-1)(i-j)!}\lambda-\frac{(i-2)(i-2)!}{(j-1)!(i-j-1)!}\lambda\right)x^{(i-j-1)\lambda}.$$

Después de varios cálculos deducimos que

$$a_{ij}^{(2)}=\left(x\binom{i-2}{j-2}-\lambda\binom{i-2}{j-2}\right)x^{(i-j-1)\lambda}=\binom{i-2}{j-2}(x-\lambda)^{(i-j)\lambda}.$$

Observamos que $a_{ij}^{(2)}=(P_{n,\lambda}[x])_{ij}^{(2)}=(P_{n,\lambda}[x-\lambda])_{i-1,j-1}$ para $i>j\geq 2$ y, por tanto, se tiene que la matriz $(P_{n,\lambda}[x])^{(2)}[2,\ldots,n+1]=(P_{n,\lambda}[x-\lambda])[1,\ldots,n]$.

A partir de ahí, se deduce que

$$(P_{n,\lambda}[x])_{ij}^{(k+1)}=(P_{n,\lambda}[x-k\lambda])_{i-k,j-k}\quad \text{para } i>j\geq k+1$$

y los multiplicadores del k-ésimo paso de la eliminación de Neville $P_{n,\lambda}[x]$ vienen dados por la expresión $x-(k-1)\lambda+(i-k-1)\lambda$ para $i=k+1,\ldots,n+1$, y con esto demostramos (2.11). Asumimos ahora que $x=k\lambda$ para algún $k\in\{0,\ldots,n-1\}$. Siguiendo la demostración anterior vemos que $(P_{n,\lambda}[x]_{ij}^{(k+1)})=(P_{n,\lambda}[0])_{i-k,j-k}$ y la EN acaba en el paso $k+1$. Por tanto, esto prueba (II).

Por último, si $x=-k\lambda$ para algún $k\in\{0,\ldots,n-1\}$, $x^{(i-j)|\lambda}=0$ para $i-j>k$. Entonces, las $n-k$ subdiagonales inferiores ya son cero y los multiplicadores asociados también lo son ya que el procedimiento de eliminación no se realiza en esas entradas. Con esto obtenemos que se satisface (III). □

Es inmediato ver que la matriz $P_{n,\lambda}[x,y]$ puede ser expresada como el producto de $P_{n,\lambda}[x]$ y de una matriz diagonal:

$$P_{n,\lambda}[x,y]=P_{n,\lambda}[x]diag(1,y^{\frac{1}{\lambda}},...,y^{\frac{n}{\lambda}}),\tag{2.14}$$

con lo que también tenemos una factorización bidiagonal de $P_{n,\lambda}$[x,y].

Observemos por (2.11), (2.12), (2.13) y (2.14) que las factorizaciones bidiagonales de $P_{n,\lambda}$[x] y $P_{n,\lambda}[x,y]$ pueden involucrar restas y por tanto, no está garantizada de antemano la HRA. De hecho, no todas las matrices de Pascal generalizadas son TP. El siguiente resultado, correspondiente al Corolario 7 del artículo [7], caracteriza cuándo lo son.

Corolario 4. Sea $P_{n,\lambda}[x]$ dada por (2.9) con $x,\lambda\in\mathbb{R}$ y $n\in\mathbb{N}$. Entonces, $P_{n,\lambda}[x]$ es una matriz TP si y solo una de las siguientes condiciones se cumplen:

(I) $x\geq(n-1)|\lambda|$.

(II) $x=k|\lambda|$ para $k=0,\cdots,n-1$.

Demostración. Por el Teorema 2.1 sabemos que $P_{n,\lambda}[x]$ admite una factorización como producto de matrices bidiagonales. Si (I) o (II) se cumplen, entonces todas las matrices bidiagonales serán no negativas y TP. Por tanto, dicho producto será también TP (véase el Corolario 1 del Capítulo 1). Para el recíproco, si $P_{n,\lambda}[x]$ es TP, dado que también es no singular, admite una factorización bidiagonal única por el Teorema 1.4. Además, esta factorización bidiagonal estará dada por el Teorema 2.1 y los m_{ij}'s serán no negativos. Entonces, o bien (I) o bien (II) se cumplen. □

Podemos dar una generalización de $P_{n,\lambda}[x,y]$ en términos de la sucesión arbitraria $\mathbf{a}=(a_n)_{n\geq 0}$

$$(P_{n,\lambda}[x,y,\mathbf{a}])_{i,j}:=a_{j-1}x^{(i-1)|\lambda}y^{(j-1)|\lambda}\binom{i-1}{j-1},$$

y así también deducimos

$$P_{n,\lambda}[x,y,\mathbf{a}]=P_{n,\lambda}[x]\text{diag}(a_0,a_1y^{\frac{1}{\lambda}},\ldots,a_ny^{\frac{n}{\lambda}}).\tag{2.15}$$

Observamos que la matriz $P_{n,\lambda}[x,y] = P_{n,\lambda}[x,y,\mathbf{1}]$, donde $\mathbf{1}$ es la sucesión formada por 1's. Usando (2.15) y el teorema anterior podemos deducir una expresión para la factorización bidiagonal de $P_{n,\lambda}[x,y,\mathbf{a}]$, $\mathscr{BD}(P_{n,\lambda}[x,y,\mathbf{a}])$. Por ejemplo, si $x \neq k\lambda$ para $k = -n+1,...,0,...,n-1$, su factorización bidiagonal viene dada por

$$(\mathscr{BD}(P_{n,\lambda}[x,y,\mathbf{a}]))_{ij} = \begin{cases} a_{j-1}y^{(j-1)|\lambda}, & \text{si } i = j, \\ x+(i-2j)\lambda, & \text{si } i > j, \\ 0, & \text{si } i < j. \end{cases}$$

Capítulo 3

Alta precisión relativa para matrices de q-enteros

3.1. Introducción

El cálculo cuántico (véase [18]) utiliza q-enteros y coeficientes q-binomiales entre otros conceptos extinguidos. Esto propicia el uso de matrices de q-enteros. Muchos cálculos algebraicos (cálculo de valores propios, valores singulares e inversas) de estas matrices pueden realizarse con HRA.

En este capítulo comenzaremos viendo en el apartado 3.2 qué es un q-entero y algunas de sus propiedades. Posteriormente, en el apartado 3.3, daremos la factorización bidiagonal de las matrices de q-Pascal. Observamos así que esta factorización bidiagonal (para $q \neq 0$) no es tan sencilla como la de las matrices de Pascal vista en el capítulo anterior.

Finalmente, en el apartado 3.4, introduciremos los números de q-Stirling y daremos la factorización de las matrices con números de q-Stirling.

3.2. q-Enteros y sus propiedades

Dado un número real positivo q y un número real r, definimos el *q-entero* [r] como

$$[r] := \begin{cases} 1+q+\ldots+q^{r-1} = \frac{1-q^r}{1-q}, & \text{si } q \neq 1, \\ r, & \text{si } q = 1, \end{cases}$$

el *q-factorial* [r]! como

$$[r]! := \begin{cases} [r][r-1]\ldots[1], & \text{si } q \neq 1, \\ r!, & \text{si } q = 1, \end{cases}$$

el *factorial q-desplazado* como

$$(a;q)_0 := 1, \quad (a;q)_n := \prod_{k=1}^{n}(1-aq^{k-1}),$$

con $n \in \mathbb{N}$, $a \in \mathbb{R}$, $q \in (0,1)$, y el *coeficiente q-binomial* $\binom{i}{j}$ como

$$\binom{i}{j} := \frac{[i]!}{[j]![i-j]!}.$$

Los coeficientes q-binomiales cumplen las siguientes relaciones de recurrencia

$$\binom{i}{j} = \binom{i-1}{j-1} + q^j \binom{i-1}{j}, \tag{3.1}$$

$$\binom{i}{j} = q^{i-j}\binom{i-1}{j-1} + \binom{i-1}{j}, \tag{3.2}$$

y además satisfacen la *q-análoga identidad de Vandermonde* (véase (12) de [8]):

$$\binom{m+n}{k} = \sum_{j=0}^{k} q^{(k-j)(m-j)}\binom{m}{j}\binom{n}{k-j}.$$

Se define también la *matriz triangular inferior de coeficientes q-binomiales*, $P_{L,q}$, cuyas entradas no nulas vienen dadas por

$$(P_{L,q})_{i,j} = \binom{i-1}{j-1}, \quad 1 \le j \le i \le n+1, \tag{3.3}$$

y su equivalente *triangular superior* $P_{U,q} := P_{L,q}^{\top}$.

Definimos por último la *matriz simétrica de q-Pascal* P_q como la matriz simétrica de coeficientes q-binomiales :

$$(P_q)_{ij} = \binom{i+j-2}{i-1}, \quad 1 \le i, j \le n+1. \tag{3.4}$$

3.3. Factorización bidiagonal de las matrices de q-Pascal

Este primer resultado da una factorización bidiagonal de $P_{L,q}$ con HRA. Además, también muestra que es una matriz TP. En consecuencia, se pueden realizar con HRA cálculos algebraicos mencionados en el Capítulo 1, como la obtención de valores propios o de la inversa.

Teorema 3.1. *Sea $P_{L,q}$ la matriz $(n+1)\times(n+1)$ dada por* (3.3)*. Entonces $P_{L,q}$ es TP y la factorización bidiagonal de $P_{L,q}$ viene dada por*

$$(\mathscr{B}\mathscr{D}(P_{L,q}))_{ij} = \begin{cases} 1, & \text{si } i = j, \\ q^{j-1}, & \text{si } i > j, \\ 0, & \text{en otro caso}, \end{cases} \tag{3.5}$$

que puede ser calculada con HRA.

Demostración. Vemos que los pivotes de la factorización de la EN de $P_{L,q}$ vienen dados por

$$p_{ij} = q^{(i-j)(j-1)}, \quad 1 \le j \le i \le n+1, \tag{3.6}$$

y los multiplicadores por

$$m_{ij} = q^{j-1}, \quad 1 \le j < i \le n+1. \tag{3.7}$$

Sea $A := P_{L,q}$ y sea $A^{(k)} = (a_{ij})_{1\le i,j\le n+1}^{(k)}$ la matriz obtenida después de aplicarle los $k-1$ pasos de la EN a A para $k = 2, ..., n+1$.

Primero probamos por inducción sobre $k \in \{2, \ldots, n+1\}$ que

$$a_{ij}^{(k)} = q^{(i-j)(k-1)}\binom{i-k}{j-k}, \quad k \le j \le i \le n+1. \tag{3.8}$$

Para $k = 2$, usando el primer paso de la EN y (3.2), tenemos que

$$a_{ij}^{(2)} = a_{ij} - \frac{a_{i1}}{a_{i-1,1}}a_{i-1,j} = a_{ij} - a_{i-1,j} = \binom{i-1}{j-1} - \binom{i-2}{j-1} = \binom{i-2}{j-2}q^{i-1},$$

para $2 \le j \le i \le n+1$.

Por tanto, ahora suponemos que (3.8) se cumple para algunos $k \in \{2,\dots,n\}$ y tomamos el k-ésimo paso de la EN para probar que (3.8) se cumple para $k+1$:

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{i-1,k}} a_{i-1,j}^{(k)}, \quad k+1 \le j \le i \le n+1.$$

Por hipótesis de inducción tenemos:

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{q^{(i-k)(k-1)}\binom{i-k}{k-k}}{q^{(i-1-k)(k-1)}\binom{i-1-k}{k-k}} a_{i-1,j}^{(k)} =$$

$$= q^{(i-j)(k-1)}\binom{i-k}{j-k} - q^{k-1}q^{(i-1-j)(k-1)}\binom{i-1-k}{j-k} =$$

$$= q^{(i-j)(k-1)}\left(\binom{i-k}{j-k} - \binom{i-1-k}{j-k}\right).$$

Aplicando (3.2) deducimos que

$$a_{ij}^{(k+1)} = q^{(i-j)k}\binom{i-(k+1)}{j-k},$$

y por tanto, (3.8) se cumple para $k+1$.

Por último, concluimos que el pivote $p_{ij} = a_{ij}^{(j)}$ viene dado por (3.8) para $k = j$ y así (3.6) se cumple.

En consecuencia, como $m_{ij} = \frac{p_{ij}}{p_{i-1,j}}$ para $i > j$ y (3.6) se cumple, tenemos que (3.7) también se cumple.

Así, $\mathscr{BD}(P_{L,q})$ puede ser calculada mediante un algoritmo libre de restas usando (3.5) y por tanto, se puede calcular con HRA. Además, por (3.5) $P_{L,q}$ se puede escribir como un producto de matrices bidiagonales no negativas (y por tanto, TP) y entonces, por el Corolario 1 del capítulo 1, $P_{L,q}$ es TP. □

Notemos que la matriz de Pascal $(n+1)\times(n+1)$ de coeficientes q-binominales $P_q = (\binom{i+j-2}{j-1})_{1\le i,j\le n+1}$ puede ser expresada como $P_q = P_{L,q}P_{U,q} = P_{L,q}P_{L,q}^{\top}$. Esta factorización se usa para deducir la factorización bidiagonal de P a partir de la de P_L. La HRA también está asegurada.

Finalmente, el Teorema 1.6 puede ser usado para deducir la factorización bidiagonal de $P_{L,q}^{-1}$. Podemos ver que

$$(\mathscr{BD}(P_{L,q}^{-1}))_{ij} = \begin{cases} 1, & \text{si } i = j, \\ -q^{i-j-1}, & \text{si } i > j, \\ 0, & \text{en otro caso.} \end{cases}$$

3.4. Matrices con números de q-Stirling

Sabemos que muchas matrices relevantes en combinatoria son matrices totalmente positivas (TP) (véase [5] y [6]), es decir, todos sus menores son no negativos (véase [3] y [14]). Como hemos visto anteriormente, bajo ciertas condiciones, muchos cálculos con matrices TP pueden realizarse con alta precisión relativa.

Dentro de la combinatoria, los números de Stirling han aparecido en muchas aplicaciones (véase [1] y [20]).

Los *números de Stirling de primer tipo* son los coeficientes $s(n,k)$ de la expansión:

$$(x)_n = \sum_{k=0}^{n} s(n,k)x^k$$

donde $(x)_n$ (*símbolo de Pochhammer*) denota el factorial descendente,

$$(x)_n := x(x-1)(x-2)\cdots(x-n+1)$$

y

$$s(n,k) = s(n-1,k-1) - (n-1)s(n-1,k).$$

Nótese que $(x)_0 := 1$ porque es un producto vacío.

Los *números de Stirling de primer tipo sin signo*, $c(n,k)$, se definen como:

$$c(n,k) := |s(n,k)| = (-1)^{n-k}s(n,k).$$

Es decir, estos números satisfacen la siguiente fórmula de recurrencia (véase [1] y [20]):

$$c(n,k) = s(n-1,k-1) + (n-1)s(n-1,k).$$

Los *números de Stirling de segundo tipo* $S(n,k)$ cuentan el número de formas de dividir un conjunto de n elementos en k partes:

$$S(n,k) := \text{card}\left(\{B|\text{card}(B) = k\ , B \subset \mathbb{N}_n\}\right)$$

donde el conjunto $\mathbb{N}_n = [1,n] \cap \mathbb{N}$ es el conjunto de los primeros n enteros.

Estos números cumplen la siguiente fórmula de recurrencia (véase [1] y [20]):

$$S(n,k) = S(n-1,k-1) + kS(n-1,k).$$

Estas tres definiciones son casos particulares de las más generales que vamos a usar en este apartado, en el que introduciremos números de q-Stirling. Obtendremos la factorización bidiagonal de las matrices S_q con números de q-Stirling de primer tipo. Comenzaremos con las matrices C_q que contienen números de q-Stirling de primer tipo sin signo. Dado que las matrices B_q con números de q-Stirling de segundo tipo son las inversas de las matrices con números de q-Stirling de primer tipo, usaremos los resultados de la factorización bidiagonal de la inversa de una matriz triangular TP (vistos en el apartado 1.4 del Capítulo 1) antes de introducir la factorización bidiagonal de las matrices B_q.

Pasamos a definir los números de q-Stirling de primer y segundo tipo, junto con las matrices formadas por ellos.

Definición 12: Los *números de q-Stirling de primer tipo*, $S_q = (s_{ij})_{1\le i,j\le n+1}$, satisfacen la siguiente relación

$$s_{ij} = s_{i-1,j-1} - [i-1]s_{i-1,j}, \tag{3.9}$$

con $s_{00} = 1$, $s_{i0} = 0$ para $i > 0$ y $s_{0j} = 0$ para $j > 0$.

Definición 13: Los *números de q-Stirling sin signo de primer tipo*, $C_q = (c_{ij})_{1\le i,j\le n+1}$, cumplen la siguiente relación:

$$c_{ij} = c_{i-1,j-1} + [i-1]c_{i-1,j}, \tag{3.10}$$

con $c_{00} = 1$, $c_{i0} = 0$ para $i > 0$ y $c_{0j} = 0$ para $j > 0$.

Las entradas de S_q son iguales en valor absoluto a las de $C_q = (c_{ij})_{1\le i,j\le n+1}$ dadas por (3.10). La diferencia radica en su patrón de signos: S_q presenta un patrón de tablero de ajedrez con signos alternantes, mientras que $C_q \ge 0$.

Definición 14: Los *números de q-Stirling de segundo tipo*, $B_q = (c_{ij})_{1\le i,j\le n+1}$, satisfacen la siguiente relación

$$b_{ij} = b_{i-1,j-1} + [j]b_{i-1,j}, \tag{3.11}$$

con $b_{00} = 1$, $b_{i0} = 0$ para $i > 0$ y $b_{0j} = 0$ para $j > 0$.

Haciendo $q = 1$ se puede comprobar que obtenemos los números de Stirling de primer y segundo tipo definidos anteriormente, respectivamente.

Vamos a deducir ahora la factorización bidiagonal de la matriz C_q. En particular, este teorema también sirve como prueba de que C_q es una matriz TP.

Teorema 3.2. *Sea $C_q = (c_{ij})_{1\leq i,j\leq n+1}$ la matriz cuya entrada (i,j) es el número de q-Stirling de primer tipo c_{ij} dado por* (3.10)*. Entonces C_q es TP y*

$$\mathscr{BD}(C_q) = \begin{cases} 1, & si\ i = j, \\ [i-j], & si\ i > j, \\ 0, & en\ otro\ caso. \end{cases} \tag{3.12}$$

se puede calcular con HRA.

Demostración. Usando (3.10), realizamos el primer paso de la EN de C_q:

$$c_{ij}^{(2)} = c_{ij} - \frac{c_{i1}}{c_{i-1,1}} c_{i-1,j} = c_{ij} - [i-1]c_{i-1,j} = c_{i-1,j-1}, \quad 2 \leq j \leq i \leq n+1.$$

Observemos que $p_{11} = 1$ y $m_{i1} = [i-1]$ para $i > 1$. Además, la matriz obtenida después de aplicar un paso de la eliminación de Neville cumple que $C_q^{(2)}[2,\cdots,n+1] = C_q[1,\cdots,n]$. Entonces, deducimos que $p_{jj} = 1$ y $m_{ij} = [i-j]$ para $i > j \geq 2$. Observamos ahora que la factorización bidiagonal de C_q (que es única) corresponde a (1.5) con D y todas las matrices G_i iguales a la matriz identidad y las matrices F_i dadas por (1.6) y $m_{ij} = [i-j]$ para $i > j \geq 2$. Por tanto, C_q es un producto de matrices bidiagonales no negativas (y por tanto, TP) y entonces, por el Teorema 3.1 de [3], C_q es TP. Observemos que C_q se puede calcular con HRA. □

Usando (3.9) en lugar de (3.10), la misma demostración del teorema anterior nos lleva a

$$\mathscr{BD}(S_q) = \begin{cases} 1, & \text{si } i = j, \\ -[i-j], & \text{si } i > j, \\ 0, & \text{en otro caso.} \end{cases} \tag{3.13}$$

A pesar de que S_q no es una matriz TP, está estrechamente relacionada con esta clase de matrices, ya que es la inversa de la matriz B_q.

El siguiente teorema muestra que las dos matrices de números de q-Stirling son inversas entre sí.

Teorema 3.3. *Las dos matrices con números de q-Stirling son inversas entre sí:*

$$\sum_k s_{ik} b_{kj} = \delta_{ij}$$

donde $\delta_{ij} := 1$ si $i = j$ y $\delta_{ij} := 0$ si $i \neq j$.

El corolario siguiente nos da la factorización bidiagonal de las matrices con números de q-Stirling de segundo tipo.

Corolario 5. *Sea $B_q = (b_{ij})_{1\leq i,j\leq n+1}$ la matriz cuya entrada (i,j) es el número de q-Stirling de segundo tipo b_{ij} dado por* (3.11)*. Entonces B_q es TP y*

$$\mathscr{BD}(B_q) = \begin{cases} 1, & \text{si } i = j, \\ [j], & \text{si } i > j, \\ 0, & \textit{en otro caso}. \end{cases}$$

Demostración. Por el Teorema 3.3, tenemos que S_q es la inversa de B_q. Por (3.13), conocemos la factorización bidiagonal de S_q. Por último, utilizando el Teorema 1.6 que nos dice cuál es la forma de la factorización bidiagonal de la matriz inversa, deducimos la factorización bidiagonal de B_q.

□

Bibliografía

[1] M. ABRAMOWITZ AND I. STEGUN, Stirling Numbers of the First Kind, en: *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables,* Dover, 1972, p. 824.

[2] P. ALONSO, J. DELGADO, R. GALLEGO, AND J. M. PEÑA, Conditioning and accurate computations with Pascal matrices. J. Comput. Appl. Math., 252: 21–26, 2013.

[3] T. ANDO, Totally positive matrices. *Linear Algebra Appl.*, 90: 165–219, 1987.

[4] M. BAYAT AND H. TEIMOORI, The linear algebra of the generalized Pascal functional matrix, *Linear Algebra Appl.*, 295: 81–89, 1999.

[5] F. BRENTI, Combinatorics and total positivity, *J. Combin. Theory*, 175–218, 1995.

[6] F. BRENTI, The applications of total positivity to combinatorics, en: *Total positivity and its applications*, Mathematics and Its Applications, **Vol 359.** Kluwer Academic Publishers, Dordrecht, 1996, pp. 451–473.

[7] J. DELGADO, H. ORERA, AND J. M. PEÑA, Accurate bidiagonal decomposition and computations with generalized Pascal matrices *J. Comput. Appl. Math.*, 391, 113443, 2021.

[8] J. DELGADO, H. ORERA, AND J. M. PEÑA, High relative accuracy with matrices of q-integers, *Numer Linear Algebra Appl*., 28: e2383, 2021.

[9] J. DELGADO AND J. M. PEÑA, Fast and accurate algorithms for Jacobi–Stirling matrices, *Applied Mathematics and Computation*, 236: 253–259, 2014.

[10] J. DEMMEL, I. DUMITRIU, O. HOLTZ, AND P. KOEV, Accurate and efficient expression evaluation and linear algebra. *Acta Numer.,* 17: 87–145, 2008.

[11] J. DEMMEL, M. GU, D. EISENSTAT, I. SLAPNIČAR, K. VESELIĆ, AND Z.DRMAČ, Computing the singular value decomposition with high relative accuracy. *Linear Algebra Appl,* 299: 21–80, 1999.

[12] J. DEMMEL AND P. KOEV, The accurate and efficient solution of a totally positive generalized Vandermonde linear system, *SIAM J. Matrix Anal. Appl.* 27: 142–152, 2005.

[13] S. M. FALLAT AND C. R. JOHNSON, *Totally nonnegative matrices*, Princeton Series in Applied Mathematics. **Vol 135.** Princeton, NJ: Princeton University, 2011.

[14] M. GASCA AND C. A. MICCHELLI, *Total positivity and Its Applications*, Mathematics and Its Applications, **Vol 359.** Kluwer Academic Publishers, 1996.

[15] M. GASCA AND J. M. PEÑA Total positivity and Neville elimination. , *Linear Algebra Appl.*, 165: 25–44, 1992.

[16] M. GASCA AND J. M. PEÑA, On factorizations of totally positive matrices, en: *Total positivity and its applications*, Mathematics and Its Applications, **Vol 359.** Kluwer Acad. Publ., Dordrecht, 1996, pp. 109–130.

[17] N. HIGHAM, *Accuracy and Stability of Numerical Algorithms*, SIAM, 2002.

[18] V. KAC AND P. CHEUNG, *Quantum calculus*, Springer, New York, 2002.

[19] P. KOEV, Accurate computations with totally nonnegative matrices, *SIAM J. Matrix Anal. Appl.*, 29: 731–751, 2007.

[20] F. MIKSA, Stirling numbers of the first kind, en: *Mathematical Tables and Other Aids to Computation*, **Vol 10.** 1956, pp. 37–38.

[21] A. PINKUS, *Totally positive matrices*, Tracts in Mathematics, **Vol 181.** Cambridge University Press, 2010.

[22] H. RONG AND C. DELI, Computing singular value decompositions of parameterized matrices with total nonpositivity to high relative accuracy, *J. Sci. Comput*. 71: 682–711, 2017.

[23] Z. ZHANG, The linear algebra of the generalized Pascal matrix, *Linear Algebra Appl.*, 250: 51–60, 1997.