Álgebra Lineal Numérica, Mínimos Cuadrados y Optimización

Lorenzo Héctor Juárez Valencia y Diana Assaely Léon Velasco
hect@xanum.uam.mx, assaely86@gmail.com
Departamento de Matemáticas,
Universidad Autónoma Metropolitana-Iztapalapa

18 de octubre de 2010

Índice general

Prólogo

Este documento contiene las notas del curso Algebra Lineal Numérica, Mínimos Cuadrados y Optimización, que se ofrecerá en el cuarto Coloquio del Departamento de Matemáticas de la UAM–I. Es un curso para estudiantes avanzados de licenciatura o del primer año de posgrado en matemática y áreas afines. El material contiene una introducción a los temas mencionados en el título. El objetivo del curso es presentar estos temas desde un enfoque unificador e integrador de conocimientos, con el objeto que el estudiante aprecie la interconexión de diferentes temas de la matemática aplicada. Se hace énfasis en el análisis y comprensión de los métodos y algoritmos, así como de los alcances y limitaciones de los mismos. Nuestro propósito es que el estudiante no solo aprenda a utilizar los métodos, sino que además sea capaz de elegir la mejor estrategia en la solución de un problema.

El primer capítulo presenta una introducción a la solución de sistemas de ecuaciones lineales por medio de los métodos directos básicos, a saber el método de Gauss, el método de factorización LU y el método de Choleski. El segundo capítulo es una introducción al estudio de los problemas de mínimos cuadrados lineales y su solución por medio de los algoritmos de ecuaciones normales y la factorización QR. En este capítulo se establece en forma clara la fuerte relación entre proyecciones ortogonales, problemas de mínimos cuadrados lineales y la solución de sistemas de ecuaciones lineales. En el tercer capítulo se realiza un estudio de la optimización de funciones cuadráticas por medio de métodos iterativos y se establece su conexión con el problema de mínimos cuadrados lineales. Al final, esperamos que el lector tenga claro que resolver un problema de mínimos cuadrados equivale a minimizar una función cuadrática y que ambos, a su vez, equivalen a resolver sistemas de ecuaciones lineales. Además la solución de cada uno de estos tres problemas se puede hacer por medio de métodos directos, como el método de Choleski o método de factorización QR, o bien por medio de métodos iterativos como el método de gradiente conjugado

Este documento concluye con la presentación y solución de dos problemas que involucran, como parte del problema, la solución de un problema de mínimos cuadrados en espacios de dimensión infinita. Uno de los problemas es un problema inverso: la ecuación de calor con retroceso en el tiempo. El otro problema es un problema de asimilación de datos en meteorología: la recuperación de un campo vectorial de viento de información incompleta. Estos problemas se reducen a dimensión finita mediante métodos de aproximación y permiten apreciar la importancia del algebra lineal numérica, la optimización y los mínimos cuadrados en la solución de los mismos.

Dr. L. Héctor Juárez y Mat. D. Assaely León México, D.F. a 17 de Octubre del 2010.

Capítulo 1

Solución de Sistemas de Ecuaciones Lineales

Uno de los problemas más frecuentemente encontrados en la computación científica es el de la solución de sistemas de ecuaciones algebraicas lineales. Este problema consiste en encontrar $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$ tal que $Ax = b$, donde la matriz $A \in \mathbb{R}^{m \times n}$ y el vector $b = (b_1, b_2, \ldots, b_m)^T \in \mathbb{R}^m$ son dados. Este problema aparece muy a menudo en muchas de las aplicaciones de la matemática, ciencias e ingeniería. Algunos ejemplos son el ajuste de datos, problemas de optimización, aproximación de ecuaciones diferenciales y de ecuaciones integrales. En este capítulo consideraremos sistemas de ecuaciones lineales cuadrados ($m = n$) que tengan solución única. Algunas de las condiciones más conocidas para que el sistema $Ax = b$ tenga solución única son:

1. A es una matriz no-singular (invertible)
2. La única solución de $Ax = \vec{0}$ es $x = \vec{0}$
3. $\det(A) \neq 0$.

1.1. Eliminación de Gauss

El método más conocido (y en muchos casos el más popular) para resolver sistemas de ecuaciones algebraicas lineales es el *método de eliminación de Gauss*. La idea básica de este método consiste en manipular las ecuaciones por medio de operaciones elementales para tranformar el sistema original en un sistema equivalente que sea más sencillo de resolver. Las *operaciones elementales* en la eliminación de Gauss son tres:

1. *Multiplicación de una ecuación por una constante no cero.*

2. *Sustracción del múltiplo de una ecuación de otra ecuación.*

3. *Intercambio de ecuaciones.*

Si alguna de estas operaciones se aplican a algún sistema de ecuaciones el sistema obtenido será *equivalente* al original. Lo mismo sucede cuando se realiza una cadena de estas operaciones. Nuestro objetivo es resolver el sistema $Ax = b$, donde $A = (a_{ij})$, $1 \leq i, j \leq n$, $b = (b_1, b_2, \ldots, b_n)^T$, que en forma explícita es:

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = & b_2 \\ \vdots \quad\quad \vdots \quad\quad\quad\quad \vdots & & \vdots \\ a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n & = & b_i \\ \vdots \quad\quad \vdots \quad\quad\quad\quad \vdots & & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n & = & b_n \end{array}$$

Si a este sistema le llamamos $A^{(1)}x = b^{(1)}$, y a los coeficientes a_{ij}, b_i los denotamos por $a_{ij}^{(1)}$, $b_i^{(1)}$, para indicar el estado original del sistema, entonces el *proceso de eliminación de Gauss* es como se muestra a continuación:

1^{er} Paso de eliminación. Si $a_{11}^{(1)} \neq 0$, podemos eliminar la incógnita x_1 a partir de la segunda ecuación. El paso típico es restar de la $i-$ésima ecuación $(i = 2, \ldots, n)$ la primera multiplicada por

$$m_{i1} = a_{i1}^{(1)}/a_{11}^{(1)} \quad i = 2, 3, \ldots, n$$

A m_{i1} se le denomina *multiplicador* asociado a la $i-$ésima ecuación en el primer paso de eliminación. Después de realizar esta operación la $i-$ésima ecuación tendrá nuevos coeficientes $a_{ij}^{(2)}$ y $b_i^{(2)}$ cuyos valores son:

$$a_{i1}^{(2)} = 0, \qquad a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i1}\, a_{1j}^{(1)}, \text{ para } j = 2, \ldots, n, \qquad b_i^{(2)} = b_i^{(1)} - m_{i1}\, b_1^{(1)}.$$

Haciendo lo anterior para cada renglón $i = 2, \ldots, n$, obtenemos el nuevo sistema $A^{(2)}x = b^{(2)}$:

$$\begin{array}{rcl} a_{11}^{(1)}x_1 + \quad a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n & = & b_1^{(1)} \\ a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n & = & b_2^{(2)} \\ \vdots \quad\quad\quad\quad \vdots & & \vdots \\ a_{n2}^{(2)}x_2 + \cdots + a_{nn}^{(2)}x_n & = & b_n^{(2)} \end{array}$$

Nota. Observese que si se va a resolver computacionalmente el problema, para almacenar los coefecientes a_{ij} y b_i, podemos escribir sobre los $a_{ij}^{(1)}$ los nuevos $a_{ij}^{(2)}$ justamente calculados. Podemos almacenar también los multiplicadores m_{i1} en donde teniamos los coeficientes $a_{i1}^{(1)}$, y recordando que todos los elementos debajo de la diagonal de la primera columna de $A^{(2)}$ son realmente cero. Más adelante veremos porqué es útil almacenar los multiplicadores.

2º paso de eliminación. En este paso el objetivo es eliminar la incógnita x_2 de la tercera ecuación hasta la última ecuación. Si $a_{22}^{(2)} \neq 0$, primero se calculan los multiplicadores

$$m_{i2} = a_{i2}^{(2)}/a_{22}^{(2)}, \quad i = 3, \ldots, n\,.$$

Los nuevos coeficientes $a_{ij}^{(3)}$ y $b_i^{(3)}$ de la i–ésima ecuación serán:

$$a_{i2}^{(3)} = 0 \qquad a_{ij}^{(3)} = a_{ij}^{(2)} - m_{i2}\, a_{2j}^{(2)}, \text{ para } j = 3, \ldots, n, \qquad b_i^{(3)} = b_i^{(2)} - m_{i2}\, b_2^{(2)}.$$

Haciendo lo anterior para cada renglón $i = 3, \ldots, n$, obtenemos el nuevo sistema $A^{(3)}x = b^{(3)}$ que es:

$$\begin{array}{rrrcl}
a_{11}^{(1)}x_1+ & a_{12}^{(1)}x_2+ & a_{13}^{(1)}x_3 + \cdots + a_{1n}^{(1)}x_n & = & b_1^{(1)} \\
 & a_{22}^{(2)}x_2+ & a_{23}^{(2)}x_3 + \cdots + a_{2n}^{(2)}x_n & = & b_2^{(2)} \\
 & & a_{33}^{(3)}x_3 + \cdots + a_{3n}^{(3)}x_n & = & b_3^{(3)} \\
 & & \vdots \qquad\qquad\quad \vdots & & \vdots \\
 & & a_{n3}^{(3)}x_3 + \cdots + a_{nn}^{(3)}x_n & = & b_n^{(3)}
\end{array}$$

Continuando de esta manera, y después de $n-1$ pasos de eliminación, obtenemos un *sistema triangular superior*

$$\begin{array}{rrrrcl}
a_{11}^{(1)}x_1+ & a_{12}^{(1)}x_2+ & a_{13}^{(1)}x_3 + \cdots + & a_{1n}^{(1)}x_n & = & b_1^{(1)} \\
 & a_{22}^{(2)}x_2+ & a_{23}^{(2)}x_3 + \cdots + & a_{2n}^{(2)}x_n & = & b_2^{(2)} \\
 & & a_{33}^{(3)}x_3 + \cdots + & a_{3n}^{(3)}x_n & = & b_3^{(3)} \\
 & & & \vdots & & \vdots \\
 & & & a_{nn}^{(n)}x_n & = & b_n^{(n)}
\end{array}$$

que denotaremos por $A^{(n)}x = b^{(n)}$. El proceso anterior se termina sin problemas siempre y cuando ninguno de los coeficientes $a_{11}^{(1)}$, a_{22}^{2}, …, $a_{nn}^{(n)}$, denominados ***pivotes***, sea cero. Cuando se realiza computacionalmente este procedimiento la matriz A se reescribe en forma sucesiva, en cada paso de eliminación, almacenando los nuevos coeficientes $a_{ij}^{(k)}$ y los correspondientes multiplicadores m_{ik} en los lugares asociados a las variables eliminadas. Lo mismo ocurre con el vectro b. Al término del proceso de eliminación obtenemos un *sistema*

triangular superior $Ux = b$ (donde $U = A^{(n)}$, $b = b^{(n)}$) el cual es equivalente al sistema original, es decir este nuevo sistema tiene exactamente la misma solución que el sistema original. Sin embargo, este nuevo sistema puede resolverse muy fácilmente por medio de la técnica de sustitución hacia atrás ó *sustitución regresiva*:

$$x_n = b_n/a_{nn}$$
$$x_i = (b_i - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii}, \quad i = n-1, n-2, \ldots, 1$$

en donde hemos suprimido los superíndices para simplificar la notación. Entonces, suponiendo que en el proceso de eliminación ninguno de los pivotes $a_{ii}^{(i)}$ es cero, el algoritmo de eliminación de Gauss para resolver el sistema $A\,x = b$ puede escribirse de la siguiente manera:

Algoritmo de eliminación de Gauss. Dados los coeficientes a_{ij} de la matriz A, y los coeficientes b_i de b

Para $k = 1, 2, \ldots, n-1$ /* *Pasos de eliminación* */
. Para $i = k+1, \ldots, n$
. . $m := a_{ik}/a_{kk}$ /* *Multiplicador asociado al renglón* i */
. . Para $j = k+1, \ldots, n$
. . . $a_{ij} := a_{ij} - m\,a_{kj}$
. . Fín
. . $b_i := b_i - mb_k$
. Fín
Fín

$x_n = b_n/a_{nn}$ /* *Empieza la sustitución regresiva* */
Para $i = n-1, \ldots, 1$
. $x_i := b_i$
. Para $j = i+1, \ldots, n$
. . $x_i := x_i - a_{ij}\,x_j$
. Fín
. $x_i := x_i/a_{ii}$
Fín

Ejemplo 1.1. *Dada la matriz A y el vector b*

$$A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 8 \\ 30 \\ 41 \end{bmatrix}$$

aplicar el método de eliminación de Gauss para calcular la solución del sistema $A\,x = b$. La solución exacta de este sistema es $x = (-1, 2, 1, 3)^T$.

Solución. En la práctica, para aplicar el método de eliminación de Gauss, es útil escribir solo los coeficientes de la matriz A en el lado izquierdo y los del vector b en el lado derecho, sin incluir las incógnitas. Entonces, el sistema inicial se puede escribir de la siguiente manera:

$$A^{(1)}x = b^{(1)}: \quad \left[\begin{array}{cccc|c} \mathbf{2} & 1 & 1 & 0 & 1 \\ 4 & 3 & 3 & 1 & 8 \\ 8 & 7 & 9 & 5 & 30 \\ 6 & 7 & 9 & 8 & 41 \end{array}\right]$$

El proceso de eliminación de Gauss se muestra a continuación:

$1^{\mathbf{er}}$ **paso de eliminación**.

Pivote: $a_{11} = 2$.

Multiplicadores para el segundo, tercero y cuarto renglones:

$$m_{21} = a_{21}/a_{11} = 4/2 = 2$$
$$m_{31} = a_{31}/a_{11} = 8/2 = 4$$
$$m_{41} = a_{41}/a_{11} = 6/2 = 3$$

Entonces,

restamos del segundo renglón el primero multiplicado por $m_{21} = 2$,

restamos del tercer renglón el primero multiplicado por $m_{31} = 4$,

restamos del cuarto renglón el primero multiplicado por $m_{41} = 3$,

con lo cual obtenemos:

$$A^{(2)}x = b^{(2)}: \quad \left[\begin{array}{cccc|c} 2 & 1 & 1 & 0 & 1 \\ & \mathbf{1} & 1 & 1 & 6 \\ & 3 & 5 & 5 & 26 \\ & 4 & 6 & 8 & 38 \end{array}\right]$$

2º paso de eliminación

Pivote: $a_{22} = 1$.

Multiplicadores para el tercero y cuarto renglones:

$$m_{32} = a_{32}/a_{22} = 3/1 = 3$$
$$m_{42} = a_{42}/a_{22} = 4/1 = 4$$

Entonces

restamos del tercer renglón el segundo multiplicado por $m_{32} = 3$,

restamos del cuarto renglón el segundo multiplicado por $m_{42} = 4$,

y se obtiene:

$$A^{(3)}x = b^{(3)}: \quad \left[\begin{array}{cccc|c} 2 & 1 & 1 & 0 & 1 \\ & 1 & 1 & 1 & 6 \\ & & \mathbf{2} & 2 & 8 \\ & & 2 & 4 & 14 \end{array}\right]$$

3er paso de eliminación

Pivote : $a_{33} = 2$.

Multiplicador para el cuarto renglón

$$m_{43} = a_{43}/a_{33} = 2/2 = 1$$

Entonces, restando del cuarto renglón el tercero, pues $m_{43} = 1$, se obtiene

$$A^{(4)}x = b^{(4)}: \quad \left[\begin{array}{cccc|c} 2 & 1 & 1 & 0 & 1 \\ & 1 & 1 & 1 & 6 \\ & & 2 & 2 & 8 \\ & & & 2 & 6 \end{array}\right] \Rightarrow \begin{array}{rcl} 2x_1 + x_2 + x_3 & = & 1 \\ x_2 + x_3 + x_4 & = & 6 \\ 2x_3 + 2x_4 & = & 8 \\ 2x_4 & = & 6 \end{array}$$

Sustitución regresiva. En el sistema triangular superior obtenido hacemos sustitución regresiva para encontrar la solución.

$$x_4 = \frac{6}{2} = 3$$
$$x_3 = \frac{8 - 2x_4}{2} = \frac{8-6}{2} = 1$$
$$x_2 = \frac{6 - x_3 - x_4}{1} = \frac{6-1-3}{1} = 2$$
$$x_1 = \frac{1 - x_2 - x_3}{2} = \frac{1-2-1}{2} = -1$$

1.2. Factorización LU

En la sección anterior hemos visto como el proceso de eliminación de Gauss transforma un sistema lineal completo en un sistema triangular superior por medio de la aplicación de operaciones elementales de eliminación. Este proceso de eliminación se puede interpretar desde un punto de vista meramente matricial. Es decir, cada paso de eliminación se puede escribir en forma compacta por medio de la multiplicación de una matriz. Por ejemplo, para el sistema $Ax = b$ con

$$A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 8 \\ 30 \\ 41 \end{bmatrix},$$

el primer paso de eliminación se puede expresar multiplicando el sistema por una matriz triangular inferior. Esta matriz triangular inferior contiene unos en la diagonal y los multiplicadores con signo contrario en sus posiciones correspondientes. El resultado se muestra a continuación

$$L_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -4 & 0 & 1 & 0 \\ -3 & 0 & 0 & 1 \end{bmatrix} \quad \Longrightarrow \quad L_1 A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 3 & 5 & 5 \\ 0 & 4 & 6 & 8 \end{bmatrix}, \qquad L_1 b = \begin{bmatrix} 1 \\ 6 \\ 26 \\ 38 \end{bmatrix},$$

obteniendo la misma matriz y lado derecho que previamente obtuvimos al final del primer paso de eliminación. En forma análoga, el segundo paso de eliminación equivale a premultiplicar el sistema anterior por la matriz triangular inferior (con los multiplicadores correspondientes con signo contrario)

$$L_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -3 & 1 & 0 \\ 0 & -4 & 0 & 1 \end{bmatrix}$$

En este caso se obtiene

$$L_2 L_1 A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 2 & 4 \end{bmatrix}, \qquad L_2 L_1 b = \begin{bmatrix} 1 \\ 6 \\ 8 \\ 14 \end{bmatrix}.$$

Finalmente, el tercer paso de eliminación equivale a premultiplicar el último sistema por la matriz

$$L_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

obteniendo

$$L_3L_2L_1A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{bmatrix}, \qquad L_3L_2L_1b = \begin{bmatrix} 1 \\ 6 \\ 8 \\ 6 \end{bmatrix}.$$

Si denotamos esta última matriz triangular superior por U, y la matriz $L_3L_2L_1$ por L^{-1}, entonces esta claro que

$$A = LU$$

El cálculo de la matriz L es sencillo como veremos a continuación. Observese que $L = (L_3L_2L_1)^{-1} = L_1^{-1}L_2^{-1}L_3^{-1}$, y basta con calcular las inversas de las matrices L_1, L_2 y L_3. El cálculo de estas inversas es trivial. Por ejemplo

$$L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -4 & 0 & 1 & 0 \\ -3 & 0 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 0 & 1 & 0 \\ 3 & 0 & 0 & 1 \end{bmatrix}.$$

Análogamente las inversas de L_2 y L_3 se obtienen simplemente cambiando el signo de sus coeficientes debajo de la diagonal, y su producto es:

$$L = L_1^{-1}L_2^{-1}L_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 3 & 1 & 0 \\ 3 & 4 & 1 & 1 \end{bmatrix}$$

La cual es una matriz triangular inferior con unos en la diagonal, y con los multiplicadores debajo de la diagonal. A esta matriz L se le conoce como la ***matriz de multiplicadores***. Podemos generalizar el resultado anterior:

Factorización *LU. Si en el proceso de eliminación de Gauss ninguno de los pivotes* $a_{ii}^{(i)}$ *es cero, entonces la matriz A se puede factorizar en la forma* $A = LU$. *La matriz L es*

triangular inferior con unos en la diagonal y con los multiplicadores debajo de la diagonal. La matriz U es la matriz triangular superior que se obtiene al final del proceso de eliminación ($U = A^{(n)}$) y contiene los pivotes en la diagonal. Es decir,

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & & \ddots & \\ l_{n1} & \dots & l_{n(n-1)} & 1 \end{bmatrix} \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{nn}^{(n)} \end{bmatrix}$$

donde $l_{ij} = m_{ij}$ para $i > 1$ son los multiplicadores que se obtienen en el proceso de eliminación de Gauss.

Observación. Como $a_{ii}^{(i)} \neq 0$, entonces A es no singular y

$$detA = det(LU) = (detL)(detU) = (1)(\prod_{i=1}^{n} a_{ii}^{(i)}) = \text{producto de los pivotes.}$$

Solución del sistema $Ax = b$ utilizando la factorización LU

Sea el sistema $Ax = b$ con $A \in \mathbb{R}^{n\times n}$ invertible, $b \in \mathbb{R}^n$. Supongase que ya tenemos una factorización $A = LU$. Entonces, el sistema de ecuaciones también su puede escribir en la forma $LUx = b$. Si hacemos $Ux = y$, entonces $Ly = b$, y por lo tanto el sistema puede resolverse en dos pasos:

1. Se resuelve el sistema triangular inferior $Ly = b$ utilizando *sustitución hacia adelante ó progresiva*:

$$\begin{aligned} y_1 &= b_1, \\ y_i &= b_i - \sum_{j=1}^{i-1} l_{ij}\, y_j\,, \qquad i = 2, \dots, n. \end{aligned}$$

2. Una vez obtenido y del paso anterior, se resuelve el sistema triangular superior $Ux = y$ utilizando *sustitución hacia atrás ó regresiva*:

$$\begin{aligned} x_n &= y_n/a_{nn}, \\ x_i &= (y_i - \sum_{j=i+1}^{n} u_{ij}\, x_j)/u_{ii}\,, \qquad i = n-1, \dots, 1, \end{aligned}$$

en donde hemos denotado por u_{ij} a los coeficientes $a_{ij}^{(i)}$, $j \geq i$, de la matriz U.

Ejemplo 1.2. *Resolver el sistema de ecuaciones anterior utilizando factorización LU.*

Solución. Del proceso de elminación de Gauss obtenemos:

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 3 & 1 & 0 \\ 3 & 4 & 1 & 1 \end{bmatrix} \quad U = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{bmatrix} \quad b = \begin{bmatrix} 1 \\ 8 \\ 30 \\ 41 \end{bmatrix}$$

Entonces, $Ly = b$ es

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 3 & 1 & 0 \\ 3 & 4 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \\ 30 \\ 41 \end{bmatrix}$$

y la solución por sustitución progresiva es

$$\begin{aligned} y_1 &= 1 \\ y_2 &= 8 - 2y_1 = 8 - 2 = 6 \\ y_3 &= 30 - 4y_1 - 3y_2 = 30 - 4 - 18 = 8 \\ y_4 &= 41 - 3y_1 - 4y_2 - y_3 = 41 - 3 - 24 - 8 = 6 \end{aligned}$$

Luego $Ux = y$ es

$$\begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 6 \\ 8 \\ 6 \end{bmatrix}$$

y la solución por sustitución regresiva es

$$\begin{aligned} x_4 &= 6/2 = 3 \\ x_3 &= (8 - 2x_4)/2 = (8 - 6)/2 = 1 \\ x_2 &= (6 - x_3 - x_4)/1 = (6 - 3 - 1)/1 = 2 \\ x_1 &= (1 - x_2 - x_3 - 0x_4)/2 = (1 - 2 - 1)/2 = -1 \end{aligned}$$

Por lo tanto, la solución es la misma que la obtenida anteriormente.

1.3. Inestabilidad del método de eliminación de Gauss

Desafortunadamente el método de eliminación de Gauss, como se ha presentado hasta el momento, no es un buen método práctico de propósito general para resolver sistemas de ecuaciones lineales. Para muchos problemas el método no es estable y, en algunos casos, su inestabilidad está asociada a una dificultad muy simple: ***para algunas matrices el método no funciona debido que se corre el peligro de dividir por cero***.

Ejemplo 1.3. *La matriz*

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

es simétrica, tiene rango completo (es invertible) y es bién condicionada. Sin embargo, el método de eliminación de Gauss falla en el primer paso debido a que el primer pivote es cero. En este caso es claro que el problema se resuelve intercambiando las filas de la matriz.

Al introducir una pequeña perturbación en el primer coeficiente de la matriz, por ejemplo 10^{-20} obtenemos una nueva matriz

$$\tilde{A} = \begin{bmatrix} 10^{-20} & 1 \\ 1 & 1 \end{bmatrix},$$

muy cercana a la anterior. Esto nos permitirá entender un poco más sobre la inestabilidad del método de eliminación de Gauss. Es decir, veremos como esta perturbación aparentemente insignificante afecta el resultado final. Consideremos por ejemplo el lado derecho $b = (1, 0)^T$, entonces el sistema original $Ax = b$ tiene la solución exacta $x = (-1, 1)^T$. Mientras que el sistema perturbado $\tilde{A}x = b$ es

$$\begin{bmatrix} 10^{-20} & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Después del primer paso de elminación obtenemos el sistema perturbado equivalente

$$\begin{bmatrix} 10^{-20} & 1 \\ 0 & 1 - 10^{20} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -10^{20} \end{bmatrix}.$$

Sin embargo, en aritmética de punto flotante *IEEE* de doble precisión, el número $1 - 10^{20}$ no se puede representar en forma exacta y éste es redondeado al número de punto flotante más cercano, que es -10^{20}. Éste cambio ocasiona que la solución obtenida sea $x = (0, 1)^T$

en lugar de la solución exacta $x = (-1, 1)^T$. Por lo tanto, una perturbación insignificante en el primer coeficiente de la matriz origina que el resultado cambie drásticamente. Este fenómeno de inestabilidad se presenta muy frecuentemente cuando se aplica eliminación de Gauss, y ocurre cuando, en algún paso de eliminación, el pivote es cero ó es muy pequeño comparado con los demás coeficientes. Cuando esto fenómemo ocurre los multiplicadores correspondientes sérán muy grandes haciendo que la matriz L en la descomposición LU tenga coeficientes muy grandes comparados con aquellos de la matriz A.

1.4. Técnicas de pivoteo

Si bién no podemos eliminar la inestabilidad en el proceso de eliminación de Gauss (o en el método de factorización LU) completamente, si podemos controlarla permutando el orden de los renglones y columnas de la matriz del sistema de ecuaciones. A esta técnica se le conoce como ***pivoteo*** y ha sido usada desde la aparición de las computadoras (alrededor de 1950). El propósito del pivoteo es asegurar que los factores L y U no sean tan grandes comparados con la matriz A. Siempre que las cantidades que aparecen en la eliminación sean manejables, los errores de redondeo se mantendrán controlados y el algoritmo será estable.

1.4.1. Pivoteo completo

La idea es la siguiente: en el k–ésimo paso de eliminación debemos escoger un pivote de entre los coeficientes del subsistema con matriz

$$A(k:n, k:n) \equiv \begin{bmatrix} a_{k,k} & a_{k,k+1} & \cdots & a_{k,n} \\ a_{k+1,k} & a_{k+1,k+1} & \cdots & a_{k+1,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,k} & a_{n,k+1} & \cdots & a_{n,n} \end{bmatrix},$$

en donde, para simplificar la exposición, los superíndices se han suprimido. Utilizamos la notación que utiliza el ambiente *MATLAB* para denotar rangos de índices, es decir $k : n$ en $A(k : n, k : n)$ denota los subíndices $k, k + 1, \ldots, n$. En esta submatriz el pivote no necesariamente es $a_{k,k}$, como lo hemos considerado hasta ahora. Con el objeto de controlar el crecimiento de los coeficientes en las matrices de factorización L y U es conveniente escoger como pivote a aquel coeficiente que tiene valor absoluto máximo:

$$|a| = \max_{k \le i,j \le n} |a_{ij}| = |a_{lm}|.$$

Después se procede a hacer el intercambio del renglón k con el renglón l, y de la columna k con la columna m, y se continua con la eliminación en la forma usual, calculando los

multiplicadores y los nuevos coeficientes. Al final, los multiplicadores que se obtienen son tales que

$$m_{i\,k} = \frac{a_{i\,k}}{|a|} \leq 1\,, \quad i = k+1, \ldots, n$$

y, en consecuencia, ninguno de los coeficientes de la matriz L, al final del proceso de eliminación (ó factorización), será mayor a uno. A esta estrategia se le denomina *pivoteo completo*; sin embargo, esta estrategia en muy poco usada por dos razones:

1. En el paso k hay $(n-k+1)^2$ posibilidades para buscar el máximo, y el costo para seleccionar los pivotes en los $n-1$ pasos de eliminación implica $\mathcal{O}(n^3)$ operaciones, lo cual es excesivo.

2. Hay que darle seguimiento al intercambio de renglones y columas.

1.4.2. Pivoteo parcial

En la práctica, es posible encontrar pivotes tan útiles como los encontrados con pivoteo completo pero realizando un número mucho menor de operaciones de busqueda. El método más común se denomina *pivoteo parcial.* En esta estrategia se intercambia solamente dos renglones en cada paso de eliminación. Así, en el $k-$ésimo paso de eliminación se escoge como pivote el coeficiente en la primera columna de la submatriz con mayor valor absoluto:

$$|a| = \max_{k \leq i \leq n} |a_{ik}| = |a_{lk}|.$$

Posteriormente se intercambian los renglones k y l. En este caso hay $n-k+1$ posibilidades para el pivoteo en el $k-$ésimo paso, y por lo tanto el número de operaciones de busqueda en todo el proceso de eliminación es en total $\mathcal{O}(n^2)$ (en realidad $n(n-1)/2$).

Como es usual con otras operaciones en el álgebra lineal numérica, el intercambio de renglones puede expresarse por medio de un producto de matrices. Como vimos anteriormente un paso de eliminación corresponde a la multiplicación izquierda por una matriz triangular inferior L_k en el $k-$ésimo paso. El pivoteo parcial complica un poco más el proceso pues ahora es necesario multiplicar por una *matriz de permutación* P_k por la izquierda antes de cada eliminación.

Matrices de permutación

Una matriz de permutación es una matriz con ceros en todos lados excepto por un

coeficiente 1 en cada renglón y columna. Por ejemplo la matriz

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

tiene un sólo 1 en cada renglón y columna, y en todas las demas entradas tiene ceros. Cualquier matriz de permutación es el producto de matrices de permutación elemental. Una matriz de permutación elemental se obtiene de la matriz identidad permutando dos de sus renglones (o dos de sus columnas) solamente. Por ejemplo, las matrices

$$P_{12} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \text{y} \qquad P_{34} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

son matrices de permutación elementales que se obtienen de la matriz identidad en $\mathbb{R}^{4\times 4}$ permutando los renglones (ó columnas) 1 y 2, y permutando los renglones (ó columnas) 3 y 4 respectivamente. La matriz de permutación P dada un poco más arriba se puede expresar como el producto de estas dos matrices, pues

$$P = P_{12}P_{34} = P_{34}P_{12}.$$

Dada cualquier matriz $A \in \mathbb{R}^{4\times 4}$, el producto por la izquierda $P_{12}A$ intercambia los renglones 1 y 2 de la matriz A, y el producto por la derecha AP_{12} intercambia las columnas 1 y 2 de la matriz A:

$$P_{12}A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} a_{21} & a_{22} & a_{23} & a_{24} \\ a_{11} & a_{12} & a_{13} & a_{14} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

$$AP_{12} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{12} & a_{11} & a_{13} & a_{14} \\ a_{22} & a_{21} & a_{23} & a_{24} \\ a_{32} & a_{31} & a_{33} & a_{34} \\ a_{42} & a_{41} & a_{43} & a_{44} \end{bmatrix}$$

1.4.3. Factorización LU con pivoteo parcial

Tomando en cuenta el intercambio de renglones en cada paso para realizar el pivoteo parcial, encontramos que, para una matriz no–singular $A \in \mathbb{R}^{n\times n}$, al término de los $n-1$ pasos de eliminación se obtiene la siguiente factorizacón

$$L_{n-1}P_{n-1}\cdots L_2P_2L_1P_1A = U\,.$$

El siguiente ejemplo ilustra esta aseveración.

Ejemplo 1.4. *Encontrar la factorización LU con pivoteo parcial para la matriz*

$$A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix} = A^{(1)}$$

Solución.
1er paso de eliminación: Claramente el pivote debe ser 8 y hay que intercambiar los renglones 1 y 3

$$P_1A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix} = \begin{bmatrix} 8 & 7 & 9 & 5 \\ 4 & 3 & 3 & 1 \\ 2 & 1 & 1 & 0 \\ 6 & 7 & 9 & 8 \end{bmatrix}$$

los multiplicadores son: $m_{21} = 4/8 = 1/2$, $m_{31} = 2/8 = 1/4$, $m_{41} = 6/8 = 3/4$. Luego

$$L_1P_1A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1/2 & 1 & 0 & 0 \\ -1/4 & 0 & 1 & 0 \\ -3/4 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 8 & 7 & 9 & 5 \\ 4 & 3 & 3 & 1 \\ 2 & 1 & 1 & 0 \\ 6 & 7 & 9 & 8 \end{bmatrix} = \begin{bmatrix} 8 & 7 & 9 & 5 \\ 0 & -1/2 & -3/2 & -3/2 \\ 0 & -3/4 & -5/4 & -5/4 \\ 0 & 7/4 & 9/4 & 17/4 \end{bmatrix} = A^{(2)}$$

2o paso de eliminación: Ahora el pivote (para el subsistema 3×3) es $7/4$, debemos intercambiar los renglones 2 y 4

$$P_2L_1P_1A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}\begin{bmatrix} 8 & 7 & 9 & 5 \\ 0 & -1/2 & -3/2 & -3/2 \\ 0 & -3/4 & -5/4 & -5/4 \\ 0 & 7/4 & 9/4 & 17/4 \end{bmatrix} = \begin{bmatrix} 8 & 7 & 9 & 5 \\ 0 & 7/4 & 9/4 & 17/4 \\ 0 & -3/4 & -5/4 & -5/4 \\ 0 & -1/2 & -3/2 & -3/2 \end{bmatrix}$$

y los multiplicadores ahora son: $m_{32} = \frac{-3/4}{7/4} = -3/7$, $m_{42} = \frac{-1/2}{7/4} = -2/7$. Así que

$$L_2P_2L_1P_1A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 3/7 & 1 & 0 \\ 0 & 2/7 & 0 & 1 \end{bmatrix} \begin{bmatrix} 8 & 7 & 9 & 5 \\ 0 & 7/4 & 9/4 & 17/4 \\ 0 & -3/4 & -5/4 & -5/4 \\ 0 & -1/2 & -3/2 & -3/2 \end{bmatrix} = \begin{bmatrix} 8 & 7 & 9 & 5 \\ 0 & 7/4 & 9/4 & 17/4 \\ 0 & 0 & -2/7 & 4/7 \\ 0 & 0 & -6/7 & -2/7 \end{bmatrix}$$
$$= A^{(3)}$$

3er paso de eliminación: El pivote (para el subsistema 2×2) es -6/7. Entonces, intercambiamos los renglones 3 y 4

$$P_3L_2P_2L_1P_1A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 8 & 7 & 9 & 5 \\ 0 & 7/4 & 9/4 & 17/4 \\ 0 & 0 & -2/7 & 4/7 \\ 0 & 0 & -6/7 & -2/7 \end{bmatrix} = \begin{bmatrix} 8 & 7 & 9 & 5 \\ 0 & 7/4 & 9/4 & 17/4 \\ 0 & 0 & -6/7 & -2/7 \\ 0 & 0 & -2/7 & 4/7 \end{bmatrix}$$

El multiplicador es $m_{43} = \frac{-2/7}{-6/7} = 1/3$. Finalmente

$$L_3P_3L_2P_2L_1P_1A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1/3 & 1 \end{bmatrix} \begin{bmatrix} 8 & 7 & 9 & 5 \\ 0 & 7/4 & 9/4 & 17/4 \\ 0 & 0 & -6/7 & -2/7 \\ 0 & 0 & -2/7 & 4/7 \end{bmatrix} = \begin{bmatrix} 8 & 7 & 9 & 5 \\ 0 & 7/4 & 9/4 & 17/4 \\ 0 & 0 & -6/7 & -2/7 \\ 0 & 0 & 0 & 2/3 \end{bmatrix}$$
$$= U\,.$$

En el anterior ejemplo hemos encontrado entonces que

$$L_3P_3L_2P_2L_1P_1A = U.$$

Con un poco más de trabajo podemos reescribir esta última igualdad en forma más adecuada. Para ello, definimos

$$L_3' = L_3, \qquad L_2' = P_3L_2P_3^{-1}, \qquad L_1' = P_3P_2L_1P_2^{-1}P_3^{-1}.$$

Se puede verificar directamente que estas últimas matrices son triangulares inferiores y que $L_3'L_2'L_1'P_3P_2P_1 = L_3P_3L_2P_2L_1P_1$. Por lo tanto

$$L_3'L_2'L_1'P_3P_2P_1A = U\,.$$

Entonces, podemos escribir

$$PA = LU \quad \text{con} \quad P = P_3P_2P_1\,, \quad L = (L_3'L_2'L_1')^{-1}\,.$$

Un cálculo directo muestra que

$$P = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \qquad L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 3/4 & 1 & 0 & 0 \\ 1/2 & -2/7 & 1 & 0 \\ 1/4 & -3/7 & 1/3 & 1 \end{bmatrix}.$$

La matriz U ya se calculó al término del proceso de eliminación.

A la anterior factorización se le denomina *factorización LU* de la matriz A con *estrategia de pivoteo simple o parcial*. Por supuesto la factorización LU corresponde, estrictamente hablando, no a A sino a una permutación de la matriz A, a saber PA. Este algoritmo se muestra a continuación:

Algoritmo de factorización LU con pivoteo parcial

Dados los coeficientes a_{ij} de A y los coeficientes b_j de b

Para $k = 1, 2, \ldots, n-1$
. Encontrar $p \geq k$ tal que $|a_{pk}| = \text{máx}_{k \leq i \leq n} |a_{ik}|$
. Intercambiar los renglones p y k (si $p \neq k$)
. Si $|a_{kk}| = 0,$ escribir: "la matriz es singular". Parar y salir
. Si no, hacer para $i = k+1, \ldots, n$
. . $m := a_{ik}/a_{kk}$
. . para $j = k+1, \ldots, n$
. . . $a_{ij} := a_{ij} - m a_{kj}$
. . Fín
. . $b_i := b_i - m b_k$
. Fín
Fín

1.5. Método de Factorización de Cholesky

Para matrices simétricas y definidas positivas el proceso de eliminación de Gauss, y por tanto la factorización LU, puede realizarse en forma más eficiente. El método que se utiliza se denomina *factorización de Choleski*. Este algorimto opera en el lado izquierdo y derecho de la matriz explotando la simetría. Este algoritmo descompone las matrices simétricas y

definidas positivas en factores triangulares haciendo la mitad de las operaciones que las necesarias para matrices generales.

1.5.1. Matrices definidas positivas

Una matriz $A \in \mathbb{R}^{n\times n}$ se dice que es *definida positiva* si $x^T Ax > 0$ para todo $x \in \mathbb{R}^n$ con $x \neq \vec{0}$. Si A es una matriz definida positiva, algunas de sus propiedades importantes son:

1. A es no singular.

2. Los valores propios de A son todos reales y positivos.

3. El determinante de la matriz A y de cada uno de sus n menores.

$$A(1:k,1:k) \equiv \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix},$$

   $k = 1, \ldots, n$, es siempre mayor a cero.

4. Cualquier submatriz principal de A de la forma $A(1:k,1:k)$ ó de la forma

$$A(k:n,k:n) \equiv \begin{bmatrix} a_{kk} & \cdots & a_{kn} \\ \vdots & & \vdots \\ a_{nk} & \cdots & a_{nn} \end{bmatrix},$$

   $k = 1, \ldots, n$, es definida positiva.

5. Cada uno de los pivotes obtenidos en el poceso de eliminación de Gauss aplicado a la matriz A es mayor a cero.

Se deja al lector verificar las propiedades 3 y 5. Aquí verificaremos el resto de las propiedades.

Verificación de la propiedad 1. La propiedad 1 es consecuencia directa de la propiedad 3.

Verificación de la propiedad 2. Si λ es un valor propio de $A \in \mathbb{R}^{n\times n}$ y $x \in \mathbb{R}^n$ es el vector propio correspondiente, entonces $x \neq \vec{0}$ y

$$x^T Ax = x^T \lambda x = \lambda \|x\|_2^2$$

así que

$$\lambda = \frac{x^T A x}{\|x\|_2^2} > 0\,.$$

Verificación de la propiedad 4. Para todo vector $x = [x_1, \ldots, x_k]^T \in \mathbb{R}^k$ no cero se tiene

$$[x_1, \ldots, x_k] A(1:k, 1:k) \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} = [x_1, \ldots, x_k, 0, \ldots, 0] A \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix} > 0\,,$$

por ser A definida positiva. Por lo tanto la submatriz $A(1:k, 1:k)$ es definida positiva. En forma análoga se puede verificar que la submatriz $A(k:n, k:n)$ es definida positiva si A lo es.

1.5.2. Factorización de Choleski

Sea $A \in \mathbb{R}^{n\times n}$ una matriz cualquiera, simétrica y definida positiva. Nuestro propósito es descomponer esta matriz en factores triangulares explotando las propiedades de la matriz. La existencia de esta factorización viene dada por el siguiente teorema

Teorema 1.5. *Cualquier matriz $A \in \mathbb{R}^{n\times n}$ simétrica y definida positiva tiene una única factorización de Choleski $A = LL^T$, donde L es una matriz triangular inferior no singular.*

Nota: Dado que $L^T = U$ es una matriz triangular superior, también podemos escribir $A = U^T U$. Una demostración de este teorema puede encontrarse en [1]. Otras referencias excelentes para profundizar en el material de este capítulo son [2] y [4].

El algoritmo de Choleski

Cuando el algoritmo de Choleski se programa sólo se necesita almacenar la parte triangular superior de A ó bién la parte triangular inferior. Esta simplificación permite reducir el número de operaciones a la mitad para lograr la factorización $A = LL^T$. El algortimo se puede construir realizando comparación de los coeficientes:

Sean $A = (a_{ij})_{1\le i,\, j\le n}$ con $a_{ij} = a_{ji}$, y $L = (l_{ij})_{1\le i,\, j\le n}$ con $l_{ii} \neq 0$, $i = 1, \ldots, n$ y $l_{ij} = 0$ si $j > i$. Comparando los coeficientes en la ecuación matricial $A = LL^T$, se obtiene

$$a_{ii} = (i\text{-ésimo renglón de } L) \times (i\text{-ésima columna de } L^T) = \sum_{k=1}^{i} l_{ik} l_{ik} = \sum_{k=1}^{i-1} l_{ik}^2 + l_{ii}^2\,.$$

Por lo tanto

$$l_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2 \right)^{1/2} \qquad i = 1, \ldots, n.$$

Análogamente

$$a_{ij} = (i\text{-ésimo renglón de } L) \times (j\text{-ésima columna de } L^T) = \sum_{k=1}^{\min(i,j)} l_{ik} l_{jk}\,.$$

Considerando el caso $i > j$:

$$a_{ij} = \sum_{k=1}^{j} l_{ik} l_{jk} = \sum_{k=1}^{j-1} l_{ik} l_{jk} + l_{ij} l_{jj}\,,$$

entonces

$$l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk} \right) / l_{jj} \qquad i = j+1, \ldots, n.$$

Obervese que en este método no hay intercambio de renglones o pivoteo. A continuación se muestra el

Algoritmo de factorización de Choleski.
$l_{11} = \sqrt{a_{11}}$
Para $i = 2, \ldots, n$
· $l_{i1} = a_{i1}/l_{11}$
Fín
Para $j = 2, \ldots, n-1$
· $l_{jj} = \left(a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 \right)^{1/2}$
· Para $i = j+1, \ldots, n$
· · $l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk} \right) / l_{jj}$
· Fín
Fín
$l_{nn} = \left(a_{nn} - \sum_{k=1}^{n-1} l_{nk}^2 \right)^{1/2}$

Ejemplo 1.6. *Encontrar la factorización de Choleski de la siguiente matriz*

$$A = \begin{bmatrix} 4 & -2 & 0 & -4 \\ -2 & 10 & 3 & 2 \\ 0 & 3 & 2 & 3 \\ -4 & 2 & 3 & 29 \end{bmatrix}.$$

Solución. La matriz es simétrica, y puede verificarse (calculando sus valores propios) que es definida positiva. Entonces, aplicando el algoritmo anterior obtenemos

$l_{11} = \sqrt{a_{11}} = \sqrt{4} = 2$
.....
$l_{21} = a_{21}/l_{11} = -2/2 = -1$
$l_{31} = a_{31}/l_{11} = 0/2 = 0$
$l_{41} = a_{41}/l_{11} = -4/2 = -2$
.....
$l_{22} = (a_{22} - l_{21}^2)^{1/2} = (10 - (-1)^2)^{1/2} = 3$
$l_{32} = (a_{32} - l_{31}l_{21})/l_{22} = (3 - (0)(-1))/3 = 1$
$l_{42} = (a_{42} - l_{41}l_{21})/l_{22} = (2 - (-2)(-1))/3 = 0$
.....
$l_{33} = (a_{33} - l_{31}^2 - l_{32}^2)^{1/2} = (2 - 0^2 - 1^2)^{1/2} = 1$
$l_{43} = (a_{43} - l_{41}l_{31} - l_{42}l_{32})/l_{33} = (3 - (-2)(0) - (0)(1))/1 = 3$
.....
$l_{44} = (a_{44} - l_{41}^2 - l_{42}^2 - l_{43}^2)^{1/2} = (29 - (-2)^2 - 0^2 - 3^2)^{1/2} = 4$

Luego la factorización de Choleski es

$$\underbrace{\begin{bmatrix} 4 & -2 & 0 & -4 \\ -2 & 10 & 3 & 2 \\ 0 & 3 & 2 & 3 \\ -4 & 2 & 3 & 29 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 2 & 0 & 0 & 0 \\ -1 & 3 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ -2 & 0 & 3 & 4 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 2 & -1 & 0 & -2 \\ 0 & 3 & 1 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 4 \end{bmatrix}}_{L^T}.$$

El algoritmo de Choleski no utiliza ninguna técnica de pivoteo y siempre es estable. Podemos verificar esta aseveración estimando el factor de crecimiento ρ al hacer la factorización. Sabemos que

$$a_{ii} = \sum_{k=1}^{i} l_{ik}^2 \,.$$

y si suponemos que $|l_{ik}| \geq 1$, entonces

$$|l_{ik}| \leq |l_{ik}|^2 \leq a_{ii} \leq \max_{1 \leq i,j \leq n} |a_{ij}| \quad \forall\, i, k.$$

Esto implica que

$$\max |l_{ik}| \leq \max |a_{ij}| \,.$$

Por lo tanto

$$\rho = \frac{\text{máx}\,|l_{ik}|}{\text{máx}\,|a_{ij}|} \leq 1\,.$$

En consecuencia, el factor de crecimiento es $\mathcal{O}(1)$, y el algoritmo siempre es estable regresivo.

1.6. Ejercicios

1. Considerese la matriz

$$A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 9 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}$$

Encuentra el determinante de A utilizando la factorización $A = LU$ (sín pivoteo). Ahora calcula el mismo determinante utilizando la factorizacón $PA = LU$ (con pivoteo). Menciona la forma general de calcular el determinante de una matriz no singular utilizando eliminación de Gauss con pivoteo parcial.

2. Supongase que se aplica eliminación de Gauss con pivoteo parcial para resolver $Ax = b$, y sea $a = \text{máx}\,|a_{ij}|$. Al realizar el primer paso de eliminación, suponiendo que $a_{11}^{(1)}$ es el que tiene mayor valor absoluto de entre los posibles pivotes, se tiene

$$|a_{ij}^{(2)}| = |a_{ij}^{(1)} - m_{i1}a_{1j}^{(1)}| \leq |a_{ij}^{(1)}| + |m_{i1}|\,|a_{1j}^{(1)}|$$

Demuestra que $|a_{ij}^{(2)}| \leq 2\,a$ primero, y luego que $|a_{ij}^{(k)}| \leq 2^{k-1}a$ para $k = 1, \ldots, n-1$, inductivamente. Concluye que el máximo factor de crecimiento al hacer eliminación de Gauss con pivoteo parcial es de orden 2^{n-1}, es decir

$$\rho = \frac{\text{máx}\,|u_{ij}|}{\text{máx}\,|a_{ij}|} = \mathcal{O}(2^{n-1})$$

3. Experimenta resolviendo sistemas $Ax = b$ con matrices A de la forma

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \qquad n = 4,$$

utilizando eliminación sin pivoteo y eliminación con pivoteo parcial. ¿Para que valor de n los resultados son inservibles? ¿Cuál es el factor de crecimiento en cada caso?

4. Demuestra que si $A \in \mathbb{R}^{n\times n}$ es una matiz definida positiva, entonces

a) El determinante de la matriz A y de cada uno de sus n menores.

$$A(1:k,1:k) \equiv \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix},$$

$k = 1, \ldots, n$, es siempre mayor a cero.

b) Cada uno de los pivotes obtenidos en el proceso de elminación de Gauss aplicado a la matriz A es mayor a cero.

5. Dada la matriz $A = (a_{ij})_{1\leq i,j\leq n}$ simétrica y definida positiva verifica que

a) Si $a_{11} = 1$, al realizar eliminación sin pivoteo $A = L_1\, A^{(1)}$, donde

$$L_1 = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ \omega_2 & 1 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ \omega_n & 0 & \ldots & 1 \end{bmatrix} \quad A^{(1)} = \begin{bmatrix} 1 & \omega_2 & \ldots & \omega_n \\ 0 & & & \\ \vdots & & M_2 & \\ 0 & & & \end{bmatrix},$$

con $\omega_2 = a_{12}$, $\omega_3 = a_{13}$,..., $\omega_n = a_{1n}$ y $M_2 = A(2:n, 2:n) - \omega\omega^T$ es simétrica.

b) En el caso general en que $a_{11} > 0$ se tiene $A = L_1\, A_s^{(1)}\, L_1^T$ donde

$$L_1 = \begin{bmatrix} \sqrt{a_{11}} & 0 & \ldots & 0 \\ \omega_2/\sqrt{a_{11}} & 1 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ \omega_n/\sqrt{a_{11}} & 0 & \ldots & 1 \end{bmatrix} \quad A_s^{(1)} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & & & \\ \vdots & & M_2 & \\ 0 & & & \end{bmatrix},$$

siendo ω_2, $\omega_3, \ldots, \omega_n$ como en el inciso anterior y $M_2 = A(2{:}n, 2{:}n) - \omega\omega^T/a_{11}$.

6. Utiliza el algoritmo de Choleski para resolver $Ax = b$, donde

a)

$$A = \begin{bmatrix} 4 & 1 & 1 & 1 \\ 1 & 3 & -1 & 1 \\ 1 & -1 & 2 & 0 \\ 1 & 1 & 0 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 8 \\ -1 \\ 6 \\ 6 \end{bmatrix}$$

b)

$$A = \begin{bmatrix} 6 & 2 & 1 & -1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 4 & -1 \\ -1 & 0 & -1 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} 3 \\ -2 \\ -1 \\ 2 \end{bmatrix}$$

Verifica primero que la matriz A sea definida positiva.

7. Resuelve el siguiente sistema de n ecuaciones con n incógnitas tomando $n = 10$ y $n = 100$ y $b_i = \left(\frac{\pi}{n+1}\right)^2 \operatorname{sen}\left(\frac{\pi i}{n+1}\right)$, $i = 1, \ldots, n$. ¿Para que valor máximo de n se pueden obtener resultados en tu computadora?

$$\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & & \ddots & & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}$$

Capítulo 2

Mínimos Cuadrados y Sistemas Lineales

El término ***mínimos cuadrados*** describe un enfoque frecuentemente usado para resolver sistemas de ecuaciones sobredeterminados ó especificados inexactamente en algún sentido apropiado. En lugar de resolver las ecuaciones exactamente, se busca solamente minimizar la suma de los cuadrados de los residuales. Muchos de los problemas que aparecen en la ciencias y en las aplicaciones se pueden reducir a la solución de un problema de mínimos cuadrados, o bien contienen subproblemas de mínimos cuadrados. Asimismo, en la actualidad los métodos de mínimos cuadrados son de fundmantal importancia en la teoría y solución de los problemas inversos asi como de los problemas mal planteados en el sentido de Hadamard. Estos problemas usualmente no tienen solución ó bien la solución no es única y, en el mejor de los casos, la solución no es continua respecto de los datos. En una gran cantidad de estos problemas es necesario ***regularizar*** el problema para encontrar una solución. El enfoque generalmete es por medio de mínimos cuadrados en espacios de Hilbert.

En este capítulo abordaremos el problema de mínimos cuadrados lineales, estudiaremos el problema desde el enfoque de proyecciones ortogonales, presentaremos dos métodos de solución que involucran la solución de sistemas de ecuaciones lineales: el método de ecuaciones normales y el método de factorización QR.

2.1. Ajuste de curvas. Mínimos cuadrados lineales

Una fuente común que da origen a problemas de mínimos cuadrados es el ajuste de curvas a un conjunto de datos dados. Sea x una variable independiente y sea $y(x)$ una función

desconocida de x la cual queremos aproximar. Suponiendo que tenemos m observaciones

$$(x_1,\, y_1),\, (x_2,\, y_2), \ldots,\, (x_m,\, y_m)\,,$$

donde $y_i \approx y(x_i)$, $i = 1, 2, \ldots,\, m$, la idea es modelar $y(x)$ por medio de una combinación de n funciones base $\phi_1(x)$, $\phi_2(x), \ldots,\, \phi_n(x)$. En el caso lineal suponemos que la función que se ajusta a los datos es una *combinación lineal* de la forma

$$y(x) = c_1\phi_1(x) + c_2\phi_2(x) + \ldots + c_n\phi_n(x) \tag{2.1}$$

Entonces, los datos deben satisfacer de manera aproximada

$$y_i = c_1\phi_1(x_i) + c_2\phi_2(x_i) + \ldots + c_n\phi_n(x_i), \quad i = 1, 2, \ldots, m\,. \tag{2.2}$$

La última expresión constituye un sistema de m ecuaciones con n incógnitas c_1, c_2,..., c_n. En el ajuste de curvas el número de funciones base n es generalmente menor que el número de datos m, es decir, $m > n$. En forma matricial la condición (2.2) puede expresarse de la siguiente forma

$$\begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \ldots & \phi_n(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \ldots & \phi_n(x_2) \\ \vdots & & \ddots & \vdots \\ \phi_1(x_m) & \phi_2(x_m) & \ldots & \phi_n(x_m) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}. \tag{2.3}$$

A la matriz de este sistema $A = (a_{ij})$ con $a_{ij} = \phi_j(x_i)$ se le denomina *matriz de diseño*. Las funciones base $\phi_i(x)$ con $i = 1, \ldots, n$, pueden ser funciones no lineales de x, pero los coeficientes y parámetros c_j aparecen en el modelo en forma lineal cuando se trata de un ajuste lineal. Dependiendo del problema particular y el objeto de estudio, las funciones base $\phi_i(x)$ pueden escogerse de muchas maneras, e incluso pueden depender de ciertos parámetros. Algunas elecciones comunes pueden ser, entre otras: polinomios, $\phi_j(x) = x^{j-1}$; funciones racionales, $\phi_j(x) = x^{j-1}/(\alpha_0 + \alpha_1 x + \cdots + \alpha_{n-1}x^{n-1})$, con $\alpha_0, \ldots,\, \alpha_{n-1}$ parámetros dados; exponenciales, $\phi_j(x) = e^{-\lambda_j x}$, con parámetros de decaimiento λ_i.

Dado que $m > n$, el sistema $A\,c = y$ dado por (2.3) es sobredeterminado y, por lo tanto, tiene solución solo si el vector de datos y se encuentra en el espacio imagen de A, denotado por $Im(A)$. En general, y no se encuentra en $Im(A)$ y por lo tanto no es posible encontrar una solución c del sistema (2.3). Entonces el problema es buscar los coeficientes de la función (2.1) que mejor ajusten los datos. El enfoque de mínimos cuadrados consiste en buscar aquel vector de coeficientes c que minimice el residual $r = y - A\,c$. Si denotamos la norma Euclideana en $\mathbb{R}^m$ por $\|\cdot\|$, entonces el problema consiste en resolver

$$\min_{c\in\mathbb{R}^n} \|Ac - y\|^2. \tag{2.4}$$

Es decir, para encontrar el ajuste de mínimos cuadrados debemos encontrar el vector de coeficientes $c = (c_1, \ldots, c_n)^T$ que minimiza la suma de cuadrados:

$$\min_{c \in \mathbb{R}^n} \sum_{i=1}^{m} (c_1\, \phi_1(x_i) + c_2\, \phi_2(x_i) + \cdots + c_n\, \phi_n(x_i) - y_i)^2 \,. \tag{2.5}$$

2.2. Método de ecuaciones normales

Para resolver el problema de minimización (2.4) se puede utilizar la metodología clásica basada en condiciones necesarias y suficientes para encontrar mínimos de funciones multidimensionales. Antes de considerar esta metodología, utilizaremos un argumento geométrico–algebraico. Una de las ideas principales para generar algoritmos que resuelvan el anterior problema de minimización descansa en el concepto de ***proyección ortogonal***. Suponiendo que y no pertenece a $Im(A)$, denotamos por

$$P_A : \mathbb{R}^m \to Im(A)$$

a la proyección ortogonal que mapea $\mathbb{R}^m$ sobre $Im(A)$. Entonces, el valor de c que minimiza la norma de $r = y - A\,c$ es aquel que satisface $A\,c = P_A\, y$, ver Figura 2.1. En otras palabras,

Figura 2.1: El espacion imagen de A es ortogonal al residual.

el residual $r = y - A\,c$ debe ser ortogonal al espacio $Im(A)$.

Por otro lado, el espacio $Im(A)$ es generado por los vectores columna de A, es decir

$$Im(A) = gen\{a_1, a_2, \ldots, a_n\}, \tag{2.6}$$

donde a_i denota al i–ésimo vector columna de A. Esto, debido a que si $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$, entonces $A\,x = x_1\, a_1 + x_2\, a_2 + \cdots + x_n\, a_n$.

Proposición 2.1. *El residual $r = y - A\,c$ es ortogonal al espacio $Im(A)$ si y solo si*

$$A^T A\, c = A^T y. \tag{2.7}$$

Demostración. El residual r es ortogonal a $Im(a)$ si y solo si $a_i^T r = 0$ para toda $i = 1, \ldots, n$. Por lo tanto r es ortogonal a $Im(a)$ si y solo si $A^T(y - A\,c) = \vec{0}$. De aquí se sigue que r es ortogonal a $Im(A)$ si y solo si se satisface (2.7).

Al sistema de ecuaciones (2.7) se le conoce como el ***sistema de ecuaciones normales*** para el problema de mínimos cuadrados. El método de ecuaciones normales consiste en resolver este sistema de ecuaciones. Basta entonces con que la matriz de las ecuaciones normales A^TA sea no singular para asegurar que el sistema tiene solución. El siguiente resultado indica las condiciones bajo las cuales esta matriz es invertible.

Teorema 2.2. *Si la matriz A tiene rango completo, entonces A^TA es una matriz cuadrada no singular, simétrica y definida positiva.*

Demostración. Dado que A es de $m \times n$, entonces A^T es de orden $n \times m$ y, por lo tanto, A^TA es una matriz cuadrada y simétrica de orden $n \times n$. Para demostrar que "A de rango completo implica A^TA no singular", basta con demostrar que si A^TA es singular entonces es de rango deficiente:

$$\begin{aligned} A^TA \text{ singular } &\Rightarrow A^TAx = \vec{0} \text{ para algún vector } x \neq \vec{0} \text{ en } \mathbb{R}^n \\ &\Rightarrow x^TA^TAx = 0, \quad x \neq \vec{0} \\ &\Rightarrow \|Ax\|_2^2 = 0, \quad x \neq \vec{0} \\ &\Rightarrow Ax = \vec{0}, \quad x \neq \vec{0} \\ &\Rightarrow A \text{ es singular} \\ &\Rightarrow A \text{ tiene rango deficiente.} \end{aligned}$$

Ahora bién, dado $x \neq \vec{0}$ en $\mathbb{R}^n$, entonces, como A^TA es no singular, se satisface $x^TA^TA\,x = \|A\,x\|_2^2 > 0$. Se concluye que A^TA debe ser definida positiva. □

Del resultado anterior se deduce que si la matriz A es de rango completo, entonces la solución del sistema de ecuaciones normales (2.7) es única e igual a

$$c = (A^TA)^{-1}A^Ty. \tag{2.8}$$

A la matriz $A^\dagger \equiv (A^TA)^{-1}A^T$ se le denomina la *seudoinversa* de A ó inversa generalizada de Moore–Penrose. Además, como $c = A^\dagger y$ entonces $A\,c = AA^\dagger y$ y, por lo tanto, la matriz $P_A = A\,A^\dagger = A(A^TA)^{-1}A^T$ es la *matriz de proyección* sobre el espacio $Im(A)$. Finalmente, es claro que dado $f(c) = \|Ac - y\|^2$, el vector c dado por (2.8) satisface $f'(c) = 2\,A^T(A\,c - y) = 0$ y, dado que la matriz Hessiana de f, $2\,A^TA$ es definida positiva, entonces $c = A^\dagger y$ es un minimo global estricto de f.

Desde el punto de vista práctico, la solución del problema de mínimos cuadrados no se encuentra invirtiendo la matriz de ecuaciones normales, a menos que la matriz del sistema

sea de orden muy pequeño para que pueda utilizarse aritmética exacta. Sin embargo, para la mayoría de los problemas prácticos esto no es posible y es necesario utilizar algoritmos computacionales estables y eficientes. Dado que, el caso no degenerado, la matriz del sistema de ecuaciones normales es simétrica y definida positiva, entonces es posible utilizar métodos directos, y el más adecuado para sistemas de ecuaciones con este tipo de matrices es el método de Choleski. El método de Choleski para resolver el sistema de ecuaciones normales (2.7) es un método de factorización que consiste en expresar la matriz de ecuaciones normales como el producto de un matriz triangular inferior por su transpuesta, para posteriormente resolver el sistema en dos pasos: sustitución progresiva y sustitución regresiva. El algoritmo se detalla a continuación.

Algoritmo de ecuaciones normales. Dada la matriz de diseño A de $m \times n$ de rango completo (ver (2.3)) y el vector de datos $y \in \mathbb{R}^m$

1. Calcular $A^T A$ y el vector $A^T y$.
2. Calcular la factorización de Choleski $A^T A = LL^T$.
3. Resolver para $z \in \mathbb{R}^n$ el sistema triangular inferior $Lz = A^T y$ (sustitución progresiva)
4. Resolver para $c \in \mathbb{R}^n$ el sistema triangular superior $L^T c = z$ (sustitución regresiva)

Ejemplo 2.3. *El NIST (National Institute of Standards and Technology) es una rama del Departamento de Comercio de los EU, que se responsabiliza de establecer estandares nacionales e internacionales. El NIST mantiene conjuntos de datos de referencia, para su uso en la calibración y certificación de software en estadística. En su página Web*

http://www.itl.nist.gov/div898/strd

en la liga Dataset Archives*, y después bajo* Linear Regresion*, se encuentra el conjunto de datos* Filip*, que consiste de 82 observaciones de una variable y para diferentes valores de x. El problema es modelar este conjunto de datos por medio de un polinomio de grado 10. Este conjunto de datos es controversial: algunos paquetes pueden reproducir valores cercanos a los coeficientes que el NIST ha declarado como lo* valores certificados*; otros paquetes dan advertencias ó mensajes de error de que el problema es mal condicionado.*

Aplicamos el método de ecuaciones normales para resolver el problema. Tenemos $m = 82$ datos (x_i, y_i) y debemos encontrar $n = 11$ coeficientes c_j para el polinomio de ajuste de grado 10. La matriz de diseño tiene coeficientes $a_{ij} = x_i^{j-1}$, $i = 1, \ldots, 82$, $j = 1, \ldots, 11$.

Para dar una idea de la complejidad de esta matriz observese que el coeficiente con mínimo valor absoluto es 1 y el coeficiente con mayor valor absoluto es $2.726901792451598\times10^9$. Esta matriz es *mal condicionada*, siendo su númeor de condición $\kappa(A) \approx \mathcal{O}(10^{15})$.

El número de condición de una matriz A se define como $\kappa(A) = \sigma_{\text{máx}}(A)/\sigma_{\text{mín}}(A)$, en donde $\sigma_{\text{máx}}(A), \sigma_{\text{mín}}(A)$ son los valores singulares de A. Una matriz de rango deficiente tiene número de condición infinito. Para una matriz cuadrada un número de condición se expresa por medio de $\kappa(A) = ||A||\,||A^{-1}||$ y coincide con la definición anterior cuando se usa la norma euclideana. En el caso de matrices normales el anterior cociente coincide con el cociente de la maginutd del máximo y mínimo valores propios. Para una matriz cuadrada un número de condición muy grande indica que es una matriz cercana a una matriz singular, por lo que perturbaciónes por errores de redondeo serán amplificadas cuando se realizan operaciones donde interviene la matriz A. De hecho, con aritmética de precisión finita, las ecuaciones normales pueden llegar a ser singulares y, en consecuencia $(A^TA)^{-1}$ no existir a pesar de que A sea de rango completo.

Volviendo a nuestro problema, la matriz de ecuaciones normales A^TA tiene coeficientes mínimo y máximo en valor absoluto 82 y 5.1×10^{19}, respectivamente. A pesar de ser de orden 11×11 tendrá un numéro de condición proporcional a $\kappa(A)^2$, lo cual nos da una idea de los problemas numéricos asociados a esta matriz. A pesar de ello aplicamos el método de ecuaciones normales para resolver el problema de mínimos cuadrados. La Figura 2.2 muestra la gráfica de los datos junto con las gráficas del polinomio certificado por el NIST y el obtenido por medio de medio del método de ecuaciones normales (con la variante de Choleski). Se observa que a pesar de que nuestra solución difiere de la certificada, sobre

Figura 2.2: Gráfica de los datos *Filip* junto con el ajuste certificado del NIST y el ajusto obtenido por el método de ecuaciones normales.

todo en en intervalo $(-6, -3)$ donde presenta oscilaciones pronunciadas, ésta se ajusta a los datos en forma aceptable. Sin embargo, es claro que la solución certificada produce menos oscilaciones y el ajuste aparentemente es mucho mejor. Para tener una mejor idea de la diferencia entre ambas soluciones, en el Cuadro 2.1 mostramos los valores de los coeficientes. Sorprendentemente los valores son muy diferentes, de hecho la diferencia relativa entre ambos, cuando se utiliza la norma euclideana es $||c_{nist} - c|| \times 100/||c_{nist}|| = 118\,\%$. Cabe aclarar que los valores certificados fueron calculados utilizando cálculos de precisión múltiple (con presición hasta de 500 dígitos) utilizando el paquete de subrutinas de Bayley en *FORTRAN* (http://crd.lbl.gov/˜dhbailey/)

Coeficientes	NIST ($\times 10^3$)	Ecuaciones Normales ($\times 10^2$)
c_1	-1.467489614229800	3.508390286748969
c_2	-2.772179591933420	5.458039227995218
c_3	-2.316371081608930	3.674467707503117
c_4	-1.127973940983720	1.398316266894402
c_5	-0.354478233703349	0.330592755393584
c_6	-0.075124201739376	0.050170726293308
c_7	-0.010875318035534	0.004860547121394
c_8	-0.001062214985889	0.000287297102007
c_9	-0.000067019115459	0.000009255550238
c_{10}	-0.000002467810783	0.000000120446348
c_{11}	-0.000000040296253	0.000000000044876

Cuadro 2.1: Comparación de los coeficientes del polinomio de ajuste obtenidos con el método de ecuaciones normales con los certificados por el NIST.

La diferencia entre los valores encontrados y los certificados en esencia se debe a las propiedades del método utilizado, pues la matriz de ecuaciones normales es muy mal condicionada y, por tanto la utilización de arimética finita de doble presición, estandar en *MATLAB*, puede provocar que la matriz de ecuaciones normales sea singular. Esto nos indica que debemos usar con cautela el método de ecuaciones normales. A pesar de ello, el método de ecuaciones normales aparece en muchos libros de estadística y métodos numéricos sin advertir al lector sobre sus propiedades.

Finalizamos esta sección mostrando la comparación de algunas valores estadísticos entre los valores obtenidos y los certificados por el NIST. Las definiciones son:

- Suma residual de cuadrados (Residual sum os squares): $RSS = (y - A\,c)^T y$.
- Media residual cuadrada (Residual mean square): $RMS = RSS/(m - n)$.
- Desviación estandar residual (Residual standard deviation): $RSD = \sqrt{RMS}$.
- R–cuadrado (R–squared): $RS = 1 - RSS/(y^T y - m\,\bar{y}^2)$.
- Regresión de suma de cuadrados (regression sum of squares): $SSReg = (A\,c)^T y - m\,\bar{y}^2)$.
- Regresión media cuadrada (Regression mean square): $MSReg = SSReg/(n - 1)$.
- Stadística F (F Statistic): $MSReg/RMS$

Estadística	NIST	Ec. Normales
RSD	0.003348010513245	0.003040024083178
R–cuadrada	0.996727416185620	0.997301818251669

Cuadro 2.2: Comparación de la desviasión estandar residual.

Estadística	NIST	Ec. Normales
SSReg	0.242391619837339	0.242531307223274
MSREg	0.024239161983734	0.024253130722327
RSS	$7.95851382172941\times10^{-4}$	$6.56163996267408\times10^{-4}$
RMS	$11.2091743968020\times10^{-6}$	$9.2417464263015\times10^{-6}$
F Stad.	2162.43954511489	2624.30168537239

Cuadro 2.3: Tabla de análisis de varianza.

Podría parecer que las estadísticas para el resultado utilizando el método de ecuaciones normales son mejores. Sin embargo, calculando la norma del residual $||A\,c-y)||$ encontramos en valor 0.028210838148692 para el resultado del NIST y el valor 0.041705784558465 (casi el doble) para el obtenido por medio de ecuaciones normales.

2.3. Ortogonalización de Gram-Schmidt

El otro método principal para resolver el problema de mínimos cuadrados es el método de factorización QR. Este es un método clásico, moderno, y popular desde 1960, citeGolub. De hecho, actualmente se considera que este método representa una de las ideas algorítmicas más importante en el álgebra lineal numérica.

2.3.1. Factorización reducida

Considere la matriz $A \in \mathbb{R}^{m\times n}$ con $m \geq n$ y sean $a_i \in \mathbb{R}^m$, $i = 1, \ldots, n$, los vectores columna de A. Es decir,

$$A = [\quad a_1 \quad | \quad a_2 \quad | \quad \cdots \quad | \quad a_n \quad]\,.$$

Los espacios sucesivos generados por los vectores columna de A tienen la siguiente propiedad

$$gen\{a_1\} \subseteq gen\{a_1, a_2\} \subseteq \ldots \subseteq gen\{a_1, \ldots, a_n\}\,.$$

La idea detrás de la factorización QR es construir una sucesión de vectores ortonormales $q_1, q_2, \ldots \in \mathbb{R}^m$ que generen estos espacios sucesivos. Supongamos que A es de rango

completo, entonces sus vectores columna son linealmente independientes y queremos que

$$gen\{q_1, \ldots, q_i\} = gen\{a_1, \ldots, a_i\}, \quad i = 1, 2, \ldots, n,$$

con $\|q_i\| = 1$ y $q_i^T q_j = \delta_{ij}$. Para contruir este conjunto de vectores podemos utilizar el método de Gram-Schmidt:

$$\begin{aligned}
q_1 &= \frac{v_1}{\|v_1\|} \quad \text{con} \quad v_1 = a_1, \\
q_2 &= \frac{v_2}{\|v_2\|} \quad \text{con} \quad v_2 = a_2 - (q_1^T a_2) q_1, \\
q_3 &= \frac{v_3}{\|v_3\|} \quad \text{con} \quad v_3 = a_3 - (q_1^T a_3) q_1 - (q_2^T a_3) q_2.
\end{aligned}$$

En general, en el j-ésimo paso, suponiendo conocidos q_1, q_2,..., q_{j-1}, un vector q_j ortonormal a ellos esta dado por

$$q_j = \frac{v_j}{\|v_j\|} \quad \text{con} \quad v_j = a_j - (q_1^T a_j) q_1 - (q_2^T a_j) q_2 - \ldots - (q_{j-1}^T a_j) q_{j-1} = a_j - \sum_{i=1}^{j-1} (q_i^T a_j) q_i \, .$$

Si definimos $r_{ij} \equiv q_i^T a_j$, y el escalar $r_{jj} = \|v_j\|$, entonces

$$\begin{aligned}
q_1 &= \frac{a_1}{r_{11}} \, , \\
q_2 &= \frac{a_2 - r_{12}\, q_1}{r_{22}} \, , \\
&\vdots \\
q_n &= \frac{a_n - \sum_{i=1}^{n-1} r_{in}\, q_i}{r_{nn}} \, .
\end{aligned}$$

Por lo tanto,

$$\begin{aligned}
a_1 &= r_{11}\, q_1 \, , \\
a_2 &= r_{12}\, q_1 + r_{22}\, q_2 \, , \\
&\vdots \\
a_n &= r_{1n}\, q_1 + r_{2n}\, q_2 + \cdots + r_{nn}\, q_n \, .
\end{aligned}$$

Este conjunto de ecuaciones tienen la siguiente representación matricial

$$\begin{aligned}
A &= \underbrace{\left[\begin{array}{c|c|c|c} a_1 & a_2 & \cdots & a_n \end{array} \right]}_{\text{matriz } m \times n} \\
&= \underbrace{\left[\begin{array}{c|c|c|c} q_1 & q_2 & \cdots & q_n \end{array} \right]}_{\text{matriz } m \times n} \underbrace{\begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ 0 & r_{22} & \ldots & r_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & r_{nn} \end{bmatrix}}_{\text{matriz } n \times n} \\
&= \qquad\qquad \hat{Q} \qquad\qquad\qquad \hat{R}
\end{aligned}$$

El siguiente algoritmo construye la factorización $\hat{Q}\hat{R}$ encontrada:

Algoritmo de Gram-Schmidt clásico

Para $j = 1, \ldots, n$

· $v_j = a_j$

· Para $i = 1, \ldots, j-1$

· · $r_{ij} = q_i^T a_j$

· · $v_j = v_j - r_{ij}\, q_i$

· Fín

· $r_{jj} = \|v_j\|$

· $q_j = v_j / r_{jj}$

Fín

Teorema 2.4. *Si $A \in \mathbb{R}^{m\times n}$ con $m \geq n$ es de rango completo, existe una única factorización reducida QR, $A = \hat{Q}\hat{R}$ con $r_{ii} > 0$, $i = 1, \ldots, n$. (Ver [2])*

2.3.2. Factorización completa

Una factorización completa QR de $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ va más allá agregando $m - n$ columnas ortonormales a $\hat{Q}$, y agregando $m - n$ renglones de ceros a $\hat{R}$ de manera tal que obtenemos una matriz ortogonal $Q \in \mathbb{R}^{m\times m}$ y otra matriz $R \in \mathbb{R}^{m\times n}$ triangular superior. Esquemáticamente

En la factorización completa las columnas q_j para $j = n+1, \ldots, m$, son ortogonales a $Im(A)$. Obsérvese que la matriz completa Q tiene la propiedad

$$Q^T Q = \begin{bmatrix} q_1^T \\ q_2^T \\ \vdots \\ q_m^T \end{bmatrix} [\ q_1 \mid q_2 \mid \cdots \mid q_m\ ] = I$$

debido a que $q_i^T q_j = \delta_{ij}$. Por tanto $Q^{-1} = Q^T$. A las matrices con esta propiedad se les denomina ***matrices ortogonales***.

Teorema 2.5. *Cualquier matriz $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ tiene una factorización completa QR, $A = QR$ con $Q \in \mathbb{R}^{m\times m}$ matriz ortogonal y $R \in \mathbb{R}^{m\times n}$ matriz triangular superior (ver Trefethen–Bau).*

Habiendo obtenido una factorización completa QR de $A \in \mathbb{R}^{m\times n}$, un problema sobredeterminado de la forma $Ax = b$ con $b \in \mathbb{R}^m$ se puede expresar en la forma $QRx = b$, que a su vez es equivalente al sistema triangular superior

$$Rx = Q^T b \quad (\text{pues } Q^{-1} = Q^T)\,.$$

Este sistema en forma expandida es

$$\begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & r_{nn} \\ 0 & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \\ f_{n+1} \\ \vdots \\ f_m \end{bmatrix}$$

con $f_i = (Q^T b)_i,\ i = 1, \dots, m$. Esta claro que este sistema se puede resolver utilizando sustitución regresiva, y que las últimas $m - n$ componentes de $f = Q^T b$ no intervienen en la solución. De hecho estas últimas componentes de $Q^T b$ están relacionadas con el residual $r = b - Ax$, pues $Q^T r = Q^T b - Rx = (0, \dots, 0, f_{n+1}, \dots, f_m)^T$ y, en consecuencia, $||r|| = ||z||$, donde $z = (f_{n+1}, \dots, f_m)^T$.

Obsérvese que en el caso que A sea una matriz cuadrada $(m = n)$ no singular este algoritmo es útil para resolver sistemas lineales $Ax = b$. Sin embargo no es el método estándar porque requiere el doble de operaciones que el método de eliminación de Gauss o el método de factorización LU.

En la práctica las fórmulas de Gram-Schmidt no se aplican para producir una factorización QR debido a que la sucesión de operaciones resulta numéricamente inestable (sensible a errores de redondeo). Esta inestabilidad se produce debido a las sustracciones y a que los vectores q_j no son estrictamente ortogonales debido a errores de redondeo. Se pueden utilizar métodos de estabilización, cambiando el orden en que se realizan las operaciones. Sin embargo, hay un método más efectivo, estable por supuesto, para encontrar la factorización QR. El nuevo método hace uso de las propiedades de las ***proyecciones ortogonales***. Por tal motivo hacemos un paréntesis en nuestra discusión para estudiar dichas propiedades.

2.4. Proyecciones en $\mathbb{R}^n$

Una proyección en $\mathbb{R}^n$ es una matriz cuadrada P de $n \times n$ tal que $P^2 = P$. El ***espacio imagen*** de P se define por

$$Im(P) = \{v \in \mathbb{R}^n \mid v = Px, \text{ para algún } x \in \mathbb{R}^n\}.$$

El ***espacio nulo*** de P se define por

$$Nul(P) = \{x \in \mathbb{R}^n \mid Px = \vec{0}\}.$$

Ejemplo 2.6. *Verificar que la siguiente matriz cuadrada es una proyección y encontrar su espacio imagen y su espacio nulo.*

$$P = \frac{1}{5}\begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 4 \end{bmatrix}$$

Solución. P es una proyección en $\mathbb{R}^3$, pues $P^2 = P$. El espacio imagen de P es el conjunto de vectores $v \in \mathbb{R}^n$ de la forma

$$v = P\,x = \frac{1}{5}\begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \frac{1}{5}(x_1 + 2x_3)\begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} = c\begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}, \quad c \in \mathbb{R}.$$

Es decir, $Im(P)$ es la línea recta determinada por el vector $(1, 0, 2)^T$. El espacio nulo es el conjunto de vectores $x = (x_1, x_2, x_3)^T$ que satisfacen $x_1 + 2x_3 = 0$. Es decir, $Nul(P)$ es el plano que pasa por el origen con vector normal $(1, 0, 2)^T$. La Figura 2.3 (izquierda) ilustra geométricamente el espacio imagen de P, el cual es una línea recta en el plano $x_1 - x_3$ de $\mathbb{R}^3$. El espacio nulo de P es el plano perpendicular a dicha recta. En dicha figura sólo se ilustra la intersección de este plano con el plano $x_1 - x_3$.

2.4.1. Algunas propiedades de las proyecciones

1. Si $v \in Im(P)$, entonces $Pv = v$.
   Demostración. $v \in Im(P) \Rightarrow v = Px$ para algún $x \in \mathbb{R}^n \Rightarrow Pv = P^2v = Px = v$.

2. Dado $v \in \mathbb{R}^n$, v se puede escribir como $v = v_1 + v_2$ con $v_1 \in Im(P)$ y $v_2 \in Nul(P)$.
   Demostración. Sea $v_1 = Pv$, entonces $v_2 = v - Pv \in Nul(P)$ pues $Pv_2 = P(v - Pv) = Pv - P^2v = \vec{0}$.

3. Si P es una proyección en $\mathbb{R}^n$, entonces $I - P$ también es una proyección.
   Demostración. $(I-P)^2 = I - 2IP + P^2 = I - 2P + P = I - P$.

   A la proyección $I - P$ se le llama ***proyección complementaria*** de P.

4. Si P es una proyección en $\mathbb{R}^n$, entonces

   - $Im(P) = Nul(I - P)$: P proyecta sobre el espacio nulo de $I - P$.
   - $Im(I - P) = Nul(P)$: $I - P$ proyecta sobre el espacio nulo de P.

   Demostración. $v \in Im(P) \Rightarrow v = Px$, con $x \in \mathbb{R}^n \Rightarrow (I-P)v = v - Pv = Px - P^2x = \vec{0} \Rightarrow v \in Nul(I-P) \Rightarrow Im(P) \subseteq Nul(I-P)$. Por otro lado $v \in Nul(I-P)$ $\Rightarrow v - Pv = \vec{0} \Rightarrow v = Pv \in Im(P) \Rightarrow Nul(I-P) \subseteq Im(P)$. Con esto se concluye que $Im(P) = Nul(I-P)$. En forma análoga se verifica que $Im(I-P) = Nul(P)$.

Ejemplo 2.7. *Retomando la proyección del ejemplo anterior, encontramos que su proyección complementaria es*

$$I - P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{1}{5}\begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 4 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 4 & 0 & -2 \\ 0 & 5 & 0 \\ -2 & 0 & 1 \end{bmatrix}$$

En la Figura 2.3 (derecha) se ilustran estas proyecciones. En dicha figura se tiene que $Im(I-P) = Nul(P)$ es el plano en $\mathbb{R}^3$ con normal $(1,0,2)^T$, y $Nul(I-P) = Im(P)$ es la línea determinada por el vector $\vec{a} = (1,0,2)^T$.

Figura 2.3: Ilustración del espacio imagen y el espacio nulo (izquierda), y de la proyección complementaria (derecha).

De la discusión anterior concluimos que

5. Una proyección P en $\mathbb{R}^n$ separa el espacio completo en dos subespacios S_1 y S_2 con $S_1 \cap S_2 = \{\vec{0}\}$, y $S_1 + S_2 = \mathbb{R}^n$. Es decir, dado $v \in \mathbb{R}^n$, $v = v_1 + v_2$ con $v_1 = Pv$ y $v_2 = (I-P)v$, ó bien $v_1 = (I-P)v$ y $v_2 = Pv$.

2.4.2. Proyecciones ortogonales

Los tipos de proyecciones más importantes en el álgebra lineal numérica y en las aplicaciones son las denominadas ***proyecciones ortogonales***. Se dice que una *proyección P es ortogonal* si $P^T = P$. De hecho, una proyección ortogonal separa el espacio completo $\mathbb{R}^n$

en dos subespacios ortogonales $S_1 \perp S_2$ con $S_1 \cap S_2 = \{\vec{0}\}$, $S_1 + S_2 = \mathbb{R}^n$. Dado $v \in \mathbb{R}^n$, $v = v_1 + v_2$ con $v_1 = Pv$ y $v_2 = (I - P)v$, y si P es ortogonal, entonces

$$v_1^T v_2 = (Pv)^T (I - P)v = v^T P^T (I - P)v = v^T P(I - P)v = v^T (P - P^2)v = \vec{0}$$

En resumen

6. Si P en una proyección ortogonal, entonces Pv y $(I - P)v$ son ortogonales para toda $v \in \mathbb{R}^n$.

Ejemplo 2.8. *La proyección del Ejemplo 2.6 es una proyección ortogonal.*

Ejemplo 2.9. ***Proyección ortogonal de rango uno****. Dado cualquier vector unitario $q \in \mathbb{R}^n$, la matriz de rango 1 dada por $P_q \equiv qq^T$ es una proyección ortogonal.*

Demostración. Primero, obsérvese que P_q es una matriz de rango 1, pues cada una de sus columnas es un múltiplo del vector $q = (q_1, \ldots, q_n)^T$:

$$P_q = \begin{bmatrix} q_1q_1 & q_1q_2 & \cdots & q_1q_n \\ q_2q_1 & q_2q_2 & \cdots & q_2q_n \\ \vdots & & \ddots & \vdots \\ q_nq_1 & q_nq_2 & \cdots & q_nq_n \end{bmatrix}.$$

P_q es una proyección, pues

$$(P_q)^2 = (qq^T)(qq^T) = q(q^Tq)q^T = q\|q\|^2q^T = qq^T = P_q\,.$$

P_q es ortogonal, pues

$$(P_q)^T = (qq^T)^T = (q^T)^Tq^T = qq^T = P_q\,.$$

Su proyección complementaria es

$$(P_q)^T = I - P_q = I - qq^T\,.$$

Ejemplo 2.10. *Para cualquier vector $a \in \mathbb{R}^n$, $a \neq \vec{0}$, existe una proyección ortogonal de rango uno dada por*

$$P_a = \frac{aa^T}{a^Ta}\,.$$

Demostración. Dado $a \in \mathbb{R}^n$, $a \neq \vec{0}$, $q = a/\|a\|$ es un vector unitario y

$$qq^T = \frac{a}{\|a\|} \cdot \frac{a^T}{\|a\|} = \frac{aa^T}{\|a\|^2} = \frac{aa^T}{a^T a} = P_a = P_a^T \,.$$

$P_a v$ produce la proyección del vector $v \in \mathbb{R}^n$ sobre la línea definida por el vector $a \in \mathbb{R}^n$. Su proyección complementaria

$$P_a^{\perp} = I - P_a = I - \frac{aa^T}{a^T a},$$

es de rango $n-1$ y proyecta el vector $v \in \mathbb{R}^n$ sobre el hiperplano perpendicular al vector a. Por lo tanto

- P_a proyecta v sobre la línea determinada por a:

$$Im(P_a) = \{x \in \mathbb{R}^n \mid x = \alpha\, a,\ \alpha \in \mathbb{R}\}\,.$$

- $P_a^{\perp}$ proyecta sobre el hiperplano con vector normal $a = (a_1, a_2, \ldots, a_n)^T$:

$$Im(P_a^{\perp}) = \{x \in \mathbb{R}^n \mid a_1 x_1 + a_2 x_2 + \ldots + a_n x_n = 0\}\,.$$

La proyección del Ejemplo 2.6 es un caso particular de una proyección de rango uno con $a = (1, 0, 2)^T$, pues

$$\frac{aa^T}{a^T a} = \frac{1}{5}\begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}\begin{bmatrix} 1 & 0 & 2 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 4 \end{bmatrix}.$$

Ejemplo 2.11. ***Proyección ortogonal sobre la imagen de una matriz*** $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$. *Si $A \in \mathbb{R}^{m\times n}$ con $m \geq n$ es de rango completo, sabemos que $A^T A$ es una matriz de $n \times n$, simétrica y definida positiva. La inversa generalizada derecha de A, se define como la matriz $A^{\dagger} = (A^T A)^{-1} A^T$ de orden $n \times m$. Esta matriz define una proyección ortogonal en el espacio $\mathbb{R}^m$ al premultiplicarse por A, pues $P_A = AA^{\dagger} = A(A^T A)^{-1} A^T \in \mathbb{R}^{m\times m}$, y claramente $P_A^2 = P_A$ y $P_A^T = P_A$. Cualquier vector $v \in \mathbb{R}^m$ es proyectado sobre el espacio imagen de la matriz A por la proyección P_A. Es decir, el espacio generado por las columnas de A es igual a $Im(P_A)$.*

Nota. Obsérvese que $P_A = A(A^T A)^{-1} A^T$ es la generalización multidimensional de la proyección de rango 1, $Pa = aa^T/a^T a$ con $a \in \mathbb{R}^m$, si observamos que a se puede ver como una de las columnas de la matriz A ó bien como una matriz de $m \times 1$.

2.5. Método de factorización QR

Uno de los métodos más utilizados para encontrar una factorización QR de una matriz dada A es el proceso de *triangularización de Housholder* introducida por Alston Householder, ver [5], debido a que es numéricamente más estable que el método de Gram-Schmidt. Este método consiste en un proceso de "***triangularización ortogonal***", en donde se construye una matriz triangular por una sucesión de operaciones matriciales semejante al proceso de eliminación de Gauss. Solo que en este caso se multiplica por matrices ortogonales Q_k, de tal manera que al final del proceso

$$Q_n \cdots Q_2 Q_1 A$$

resulta triangular superior. Cada matriz Q_k se escoge para introducir ceros debajo de la diagonal en la k-ésima columna. Por ejemplo, para una matriz A de 5×3, las operaciones Q_k se aplican como se muestra a continuación:

$$\overbrace{\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \end{bmatrix}}^{A^{(0)}} \underbrace{\overrightarrow{Q_1}}_{A} \overbrace{\begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & x & x \\ 0 & x & x \\ 0 & x & x \end{bmatrix}}^{A^{(1)}} \underbrace{\overrightarrow{Q_2}}_{Q_1 A} \overbrace{\begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & x \\ 0 & 0 & x \\ 0 & 0 & x \end{bmatrix}}^{A^{(2)}} \underbrace{\overrightarrow{Q_3}}_{Q_2 Q_1 A} \overbrace{\begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & x \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}^{A^{(3)}} = R\,,$$
$$Q_3 Q_2 Q_1 A$$

donde las x indican coeficientes no cero en general.

¿Cómo se construyen tales matrices ortogonales Q_k?

2.5.1. Tranformaciones o reflexiones de Householder

Si $A \in \mathbb{R}^{m \times n}$ con $m \geq n$, cada Q_k se escoge como una matriz ortogonal de la forma

$$Q_k = \begin{bmatrix} I & 0 \\ 0 & H \end{bmatrix},$$

donde I es la matriz identidad de orden $(k-1)\times(k-1)$ y H es una matriz ortogonal de orden $(m-k+1)\times(m-k+1)$. La multiplicación por H debe introducir ceros debajo de la diagonal

en la k–ésima columna. Esquemáticamente, esta operación se muestra a continuación

$$\begin{bmatrix} I & 0 \\ 0 & H \end{bmatrix}_{m\times m} \begin{bmatrix} a_{11} & \dots & a_{1k-1} & a_{1k} & \dots & a_{1n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & a_{k-1k-1} & a_{k-1k} & \dots & a_{k-1n} \\ 0 & \dots & 0 & a_{kk} & \dots & a_{kn} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & a_{mk} & \dots & a_{mn} \end{bmatrix}_{m\times n}$$

La matriz de la izquierda es la matriz ortogonal Q_k de orden $m \times m$, mientras que la de la derecha es la matriz $A^{(k-1)}$ de orden $m \times n$. El resultado de la anterior multiplicación es la matriz $A^{(k)} \in \mathbb{R}^{m\times n}$, la cual tiene los mismos bloques que $A^{(k-1)}$, excepto el bloque diagonal inferior que debe cambiar por

$$H \cdot \begin{bmatrix} a_{kk} & \dots & a_{kn} \\ \vdots & \ddots & \vdots \\ a_{mk} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} a'_{kk} & a'_{kk+1} & \dots & a'_{kn} \\ 0 & a'_{k+1k+1} & \dots & a'_{k+1n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a'_{mk+1} & \dots & a'_{mn} \end{bmatrix}$$

Por supuesto, aquí el paso fundamental es la multiplicación de H por la primera columna de la submatriz de la derecha:

$$H \begin{bmatrix} a_{kk} \\ a_{k+1k} \\ \vdots \\ a_{mk} \end{bmatrix} = \begin{bmatrix} a'_{kk} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

donde $a'_{kk} = \|x_k\|$, con $x_k = (a_{kk}, \dots, a_{mk})^T$. Es decir, el algoritmo de Householder escoge H como una matriz particular llamada ***reflexión de Householder***. Esta matriz introduce los ceros correctos en la k–ésima columna:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{m-k+1} \end{bmatrix} \quad \longrightarrow \quad Hx = \begin{bmatrix} \|x\| \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \|x\| e_1$$

con $e_1 = (1, 0, \dots, 0)^T$ de tamaño $m - k + 1$. Para obtener la reflexión de Householder H, observamos en la Figura 2.4 (izquierda) que ésta refleja $\mathbb{R}^{m-k+1}$ a través del hiperplano H^+

ortogonal a $v = \|x\|e_1 - x$. Cuando la reflexión se aplica a cada punto en algún lado de H^+, el resultado es la imagen reflejada en el otro lado de H^+. En particular x es enviado a $\|x\|e_1$. Por otro lado, para cualquier $y \in \mathbb{R}^{m-k+1}$, el vector

$$P\,y = \left(I - \frac{vv^T}{v^T v}\right) y$$

es la proyección ortogonal de y sobre el espacio H^+ (perpendicular a v). Para reflejar y a través de H^+, debemos continuar al doble de la distancia en la misma dirección. La reflexión H debe ser entonces (ver Figura 2.4 derecha)

$$H = I - 2\frac{vv^T}{v^T v} \qquad \text{con} \qquad v = \|x\|e_1 - x.$$

Figura 2.4: Reflexión de Householder $H = I - 2vv^T/v^T v$ para un vector dado x

Ejemplo 2.12. *Dado el vector $x = (3, 4, 0)^T$, encontrar la reflexión de Householder H tal que $Hx = \|x\|e_1 = (5, 0, 0)^T$.*

Solución.

$$v = \|x\|e_1 - x = \begin{bmatrix} 5 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 3 \\ 4 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ -4 \\ 0 \end{bmatrix},$$

$$vv^T = \begin{bmatrix} 2 \\ -4 \\ 0 \end{bmatrix} \begin{bmatrix} 2 & -4 & 0 \end{bmatrix} = \begin{bmatrix} 4 & -8 & 0 \\ -8 & 16 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{y} \quad v^T v = 20.$$

Luego

$$H = I - 2\frac{vv^T}{v^T v} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{1}{10}\begin{bmatrix} 4 & -8 & 0 \\ -8 & 16 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 3/5 & 4/5 & 0 \\ 4/5 & -3/5 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Se puede verificar que H efectivamente satisface $Hx = \|x\|e_1 = (5,\, 0,\, 0)^T$. En la Figura 2.5 se ilustra el resultado de este ejemplo. El plano H^+ tiene ecuación $v_1x_1 + v_2x_2 + v_3x_3 = 0$, es decir, $2x_1 - 4x_2 = 0$ con x_3 arbitrario. Obsérvese además que

$$HH^T = \begin{bmatrix} 3/5 & 4/5 & 0 \\ 4/5 & -3/5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3/5 & 4/5 & 0 \\ 4/5 & -3/5 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I$$

Figura 2.5: Reflexión de Householder del ejemplo 2.12.

En general, se puede verificar fácilmente que

$$HH^T = \left[I - 2\frac{vv^T}{v^Tv}\right]\left[I - 2\frac{vv^T}{v^Tv}\right]^T = I\,.$$

de modo que H es siempre una matriz ortogonal. Por lo tanto, las matrices Q_k en el proceso de triangularización de Householder son ortogonales, pues

$$Q_kQ_k^T = \begin{bmatrix} I & 0 \\ 0 & H \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & H^T \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & HH^T \end{bmatrix} = \begin{bmatrix} I_{k-1} & 0 \\ 0 & I_{m-k+1} \end{bmatrix} = I_m\,.$$

2.5.2. La mejor de las dos reflexiones

En la exposición anterior se ha simplificado el procedimiento. El vector x en realidad puede reflejarse ya sea a $\|x\|e_1$ ó bien a $-\|x\|e_1$, dando lugar a dos reflexiones, una a través del hiperplano H^+, y otra a través del hiperplano H^-, como se muestra en la Figura 2.6.

Figura 2.6: Las dos reflexiones de Householder

Desde el punto de vista matemático, cualquiera de las dos elecciones es satisfactoria. Sin embargo, para asegurar estabilidad en el algoritmo numérico se debe escoger aquella reflexión para la cual x se encuentre más alejado de su punto reflejado $\|x\|e_1$ ó $-\|x\|e_1$. Para lograr esto escogemos $v = -signo(x_1)\|x\|e_1 - x$, o bien podemos quitar el signo "$-$", y escoger $v = signo(x_1)\|x\|e_1 + x$. La función *signo* se define como:

$$signo(x_1) = \begin{cases} 1 \text{ si } x_1 \geq 0\,, \\ -1 \text{ si } x_1 < 0\,. \end{cases}$$

Vale la pena recordar que x_1 es la primera coordenada de x. Las Figura 2.7 muestra ambos casos. La razón por la cual esta elección para v es conveniente es que si el ángulo entre H^+ y e_1 es muy pequeño, entonces el vector $v = \|x\|e_1 - x$ será más pequeño que x ó $\|x\|e_1$, y

Figura 2.7: Gráfica de la izquierda: $x_1 > 0 \Rightarrow signo(x_1) = +1$. Gráfica de la derecha: $x_1 < 0 \Rightarrow signo(x_1) = -1$.

el cálculo de v representa la sustracción de cantidades casi iguales, con lo cual se obtienen errores de cancelación. Por otro lado, si escogemos el signo de tal forma que en lugar de restar sumemos, cortamos el efecto de cancelación y $\|v\|$ nunca será más pequeño que $\|x\|$, con lo cual aseguramos estabilidad en los cálculos.

2.5.3. El algoritmo QR

Tomando en cuenta la discusión anterior, a continuación incluimos el procedimiento de factorización QR de la matriz A por medio de reflexiones de Householder. Asimismo, también incluimos el método de sustitución regresiva para encontrar la solución del problema de ajuste de mínimos cuadrados (2.4). En el procedimiento hemos utilizado la notación *MATLAB* para denotar matrices y submatrices asi como vectores, aunque el código puede programarse en cualquier otro lenguaje ó ambiente de programación. Cabe aclarar que en este algoritmo no se almacenan las matrices de reflexión Q_k, ni tampoco los vectores de reflexión v_k, solamente se calcula al matriz triangular superior R y se almacena en el mismo espacio de memoria que ocupa A. El vector $b = Q^T y$ se calcula simulataneamente al cálculo de R.

Algortimo de triangularización de Housholder

% Factorización QR y cálculo de $b = Q^T y$
Para $k = 1, \ldots, n$
. $x = A(k : m,\, k)$
. $v = signo(x_1)\|x\|e_1 + x$
. $v = v/\|v\|_2$
. $A(k : m,\, k : n) = A(k : m,\, k : n) - 2v\,(v^T\, A(k : m,\, k : m))$
. $b(k : m) = b(k : m) - 2v(v^T\, b(k : m))$
Fín

% Sustitución regresiva
$x(n) = b(n)/A(n, n)$
Para $k = n - 1 : -1 : 1$
$\vdots$ $x(k) = (b(k) - A(k, k + 1 : n) \cdot b(k + 1 : n))/A(k, k)$
Fín

Ejemplo 2.13. *Consideremos de nuevo el problema en el Ejemplo 1. Resovemos directamente el sistema sobredeterminado* $A\,c = y$ *por medio del método de factorización* QR*. Se utilizó arimética de punto flotante de doble presición. El Cuadro (2.4) muestra los coeficientes obtenidos junto con los certificados por el* NIST*. En esta ocasión los resultados son muy buenos. No presentamos las gráficas porque estan son indistinguibles a simple vista y la gráfica de la curva de ajuste certificada por el* NIST *se muestra en la Figura 2.2.*

En esta ocasión diferencia relativa entre ambos, cuando se utiliza la norma euclideana, es $||c_{nist} - c|| \times 100/||c_{nist}|| = 2.2\times10^{-6}\,\%$*. Los Cuadros 2.5 y 2.6 muestran la comparación de las estadísticas mencionadas en el ejemplo 1 pero ahora con los resultados del algoritmo* QR*. Finalmente, el residual* $\|Ac - y\|$ *en la norma euclideana es 0.028210838142565 para el resultado obtenido con el algoritmo* QR*, mientras que para la solución certificada por el* NIST *es 0.028210838148692, los cuales coinciden hasta la onceava cifra decimal.*

Coeficientes	NIST ($\times10^3$)	Alg. QR
c_1	-1.467489614229800	-1.467489582388863
c_2	-2.772179591933420	-2.772179531420028
c_3	-2.316371081608930	-2.316371030602281
c_4	-1.127973940983720	-1.127973915874232
c_5	-0.354478233703349	-0.354478225708109
c_6	-0.075124201739376	-0.075124200018508
c_7	-0.010875318035534	-0.010875317781920
c_8	-0.001062214985889	-0.001062214960612
c_9	-0.000067019115459	-0.000067019113828
c_{10}	-0.000002467810783	-0.000002467810721
c_{11}	-0.000000040296253	-0.000000040296251

Cuadro 2.4: Comparación de los coeficientes del polinomio de ajuste.

Estadística	NIST	Alg. QR
RSD	0.003348010513245	0.003347929499856
R–cuadrada	0.996727416185620	0.996727574560212

Cuadro 2.5: Comparación de la desviasión estandar residual.

En conclusión, los resultados obtenidos en este problema muestran la superioridad del método de factorización QR sobre el método de ecuaciones normales para la solución de

Estadística	NIST	Alg. QR
SSReg	0.242391619837339	0.242391658352084
MSREg	0.024239161983734	0.024239165835208
RSS	$7.95851382172941\times10^{-4}$	$7.958128674566620\times10^{-4}$
RMS	$11.2091743968020\times10^{-6}$	$11.20863193600932\times10^{-6}$
F Stad.	2162.43954511489	2162.544543668758

Cuadro 2.6: Tabla de análisis de varianza.

problemas de mínimos cuadrados.

2.6. Ejercicios

1. Sean los puntos $(x_1, y_1), \ldots, (x_m, y_m)$ y $p(x) = c_1 + c_2x + \cdots + c_nx^{n-1}$ un polinomio de grado $n < m$. Este polinomio será un ajuste de mínimos cuadrados si minimiza la suma de cuadrados $\sum_{i=1}^{m}(p(x_i)-y_i)^2$. Verifica que el problema de ajuste es equivalente a resolver el problema $\text{mín}_{c\in\mathbb{R}^n} \|Ac-y\|^2$, donde $y = (y_1, y_2, \ldots, y_m)^T$ y A es la matriz de diseño con $a_{ij} = x_i^{j-1}$ para $1 \leq i \leq m$, $1 \leq j \leq n$.

2. Sea A una matriz de orden $m\times n$ con $m \geq n$. Demuestra que si A es de rango deficiente, entonces A^TA es singular. Construye un ejemplo con una matriz A de 3×2. Ahora demuestra que si A es una matriz de rango completo, entonces A^TA es una matriz no singular, simétrica y definida positiva. Construye un ejemplo.

3. Suponiendo que A es de rango completo, demuestra que el vector $c^* \in R^n$ es la solución del problema de mínimos cuadrados sí y solo sí resuelve el sistema de ecuaciones normales $A^TA\,c^* = A^Ty$. Sugerencia: Considera la función $f : \mathbb{R}^n \mapsto \mathbb{R}$ definida por $f(c) = ||A\,c-y||^2$, y demuestra que c^* satisface las condiciones necesarias y suficientes de segundo orden para ser el único mínimo global de f. Recuerda que $||x||^2 = x^Tx$.

4. Utiliza el método de ecuaciones normales para encontrar el polinomio de interpolación de segundo grado que mejor se ajuste a los datos $(-2, 1)$, $(-1, 0)$, $(0, 1)$, $(1, 0)$, $(2, 3)$. Grafica los datos y el polinomio en la misma figura. Ahora encuentra el polinomio de ajuste de grado tres. Por último realiza el ajuste con un polinomio de grado cuatro. Grafica los polinomios. Comenta los resultados.

5. Dados los vectores $a_1, a_2, \ldots, a_n \in \mathbb{R}^n$ linealmente independientes, el proceso de ortogonalización de Gram–Schmidt produce una sucesión de vectores $q_1, q_2, \ldots, q_n$. De-

muestra que los vectores así generados son ortonormales, es decir $q_k^T q_j = \delta_{kj}$ para $1 \leq k, j \leq n$. Utiliza este algoritmo para encontrar los vectores ortonormales asociados a los vectores $a_1 = (2,0,1)^T$, $a_2 = (0,2,1)^T$, $a_3 = (1,1,3)^T$. Con el resultado, construye la factorización reducida $\hat{Q}\hat{R}$ para la matriz A cuyos vectores columna son los vectores a_1, a_2, a_3. Verifica que $\hat{Q}$ es una matriz ortogonal.

6. Dado el vector no nulo $a \in \mathbb{R}^n$, demuestra que $P_a = a\,a^T/a^T a$ y $P_a^\perp = I - P_a$ son proyecciones ortogonales complementarias. Sea $a = (1,1,1)^T$ y $v = (-1,0,3)^T$. Encuentra P_a y $P_a^\perp$ y verifica que $P_a\,v$ es ortogonal a $P_a^\perp\,v$. Verifica que P_a es de rango uno y que $P_a^\perp$ es de rango dos. Encuentra la ecuación del plano sobre el cual proyecta $P_a^\perp$.

7. Verifica que la matriz $A = \begin{bmatrix} 1 & 1 \\ 2 & 0 \\ -1 & 1 \end{bmatrix}$ es de rango completo.

a) Encuentra su inversa generalizada derecha (seudo–inversa de Moore–Penrose) $A^\dagger = (A^T A)^{-1} A^T$.

b) Encuentra $P_{_A} = A\,A^\dagger$ y verifica que esta es una proyección ortogonal en $\mathbb{R}^3$.

c) Sean a_1 y a_2 los vectores columna de la matriz A. Verifica que $P_{_A} a_1 = a_1$ y $P_{_A} a_2 = a_2$.

d) Dado un vector fuera del espacio columna de A, por ejemplo $v = (1,0,0)^T$, verifica que $P_{_A} v$ se encuentra en el espacio columna de A. Es decir, $P_{_A} v = \alpha_1 a_1 + \alpha_2 a_2$ con α_1, $\alpha_2 \in \mathbb{R}$.

8. Demuestra que si Q es una matriz cuadrada con vectores columna ortonormales, entonces $Q^T Q = I$. A una matriz con esta propiedad se le denomima matriz ortogonal. Demuestra que

a) $\det(Q) = \pm 1$. Si A es una matriz cuadrada y $A = QR$, entonces $\det(A) = \pm \det(R)$.

b) $\|Qx\| = \|x\|$, es decir una transformación ortogonal preserva distancias.

c) El producto de matrices ortogonales es ortogonal.

d) La transpuesta de una matriz ortogonal es ortogonal.

9. Sean $\hat{Q} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \\ -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}$ y $\hat{R} = \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$

a) Encontrar la matriz A que tiene factorización reducida $\hat{Q}\hat{R}$.

b) Encontrar la factorización completa QR de A.

c) Utiliza la factorización reducida para encontrar la solución de mínimos cuadrados de $Ax = (1,\, 1,\, 1)^T$. Encuentra la norma del residual.

d) Verifica que $\hat{R}^T\hat{R} = R^TR = A^TA$.

10. Si la matriz A de orden $m \times n$ es de rango completo

a) Demostrar que $A^\dagger A = I$.

b) Demostrar que la solución de mínimos cuadrados del problema sobredeterminado $Ax = b$ es $x = A^\dagger b$.

c) Supóngase que $A = QR$ es una descomposición QR para A. Expresar $A^\dagger$ en términos de Q y R.

d) Si $A = [1\ \ 2;\ 2\ \ 1;\ -1\ \ 2]$, encontrar $A^\dagger$ y usarla para encontrar la solución de mínimos cuadrados de $Ax = (2,\, 2,\, 2)^T$. Calcular el número de condición de A definido por $\kappa(A) = \|A\|\,\|A^\dagger\|$.

11. Encuentra la reflexión de Householder H_1 que transforma el vector $x = (4,\, 0,\, 3)$ en $H_1\,x = (5,\, 0,\, 0)$ y la reflexión de Householder H_2 que lo transforma en $H_2\,x = (-5,\, 0,\, 0)$. Encuentra las ecuaciones de los planos de reflexión H^+ y H^-. Demuestra que la reflexión $H = I - 2v\,v^T$, donde I es la matriz identidad de $n \times n$ y $v \in \mathbb{R}^n$ es un vector unitario, es una transformación ortogonal.

12. Consulta la página del *NIST* http://www.itl.nist.gov/div898/strd/lls/lls.shtml e ingresa a la liga **Wampler4** y después a la liga **Data Data file (ASCII Format)**. Al final de esta liga se encuentra el conjunto de 21 datos. Copia estos datos y calcula el polinomio de ajuste de grado 5 utilizando los métodos de ecuaciones normales y factorización QR. Compara tus resultados con los valores certificados. Encuentra el residual $||A\,c - y||$ en cada caso, así como la diferencia relativa con respecto a los valores certificados. Escribe tus conclusiones.

Capítulo 3

Optimización Cuadrática y Mínimos Cuadrados

La optimización cuadrática significa aplicar lo que se conoce de la optimización de las funciones cuadráticas a la solución de problemas de optimización. Nosotros estamos interesados en el problema de mínimos cuadrados. En este capítulo veremos la conexión entre los problemas de mínimos cuadrados y la minimización de funciones cuadráticas de variable multidimensional, así la relación de ambos con la solución de sistemas de ecuaciones lineales con matrices simétricas y definidas positivas. Asimismo, introduciremos el algoritmo iterativo de gradiente conjugado para resolver estos problemas. Terminaremos el capítulo presentando dos problemas en dimensión infinita cuya discretización resulta en un problema de mínimos cuadrados en dimensión finita. En estos casos resulta conveniente aplicar métodos iterativos en lugar de los métodos de factorización como Choleski y QR, estudiados en el capítulo anterior. Al final quedará clara la estrecha conexión entre solución de sistemas lineales, mínimos cuadrados y optimización cuadrática, así como su aplicación a ecuaciones integrales y ecuaciones diferenciales en la solución de problemas prácticos reales.

3.1. Funciones cuadráticas

3.1.1. Funciones cuadráticas de una variable

Cualquier función cuadrática (polinomio de segundo grado) en una variable, $f : \mathbb{R} \to \mathbb{R}$, es de la forma

$$f(x) = \frac{1}{2}\, a\, x^2 - b\, x + c \tag{3.1}$$

donde a, b y c son constantes reales con $a \neq 0$. Utilizando el procedimiento de completar al cuadrado, esta función puede rescribirse en la siguiente forma

$$f(x) = \frac{a}{2}\left(x - \frac{b}{a}\right)^2 + \left(c - \frac{b^2}{2a}\right). \tag{3.2}$$

De hecho, esta representación corresponde al polinomio de Taylor

$$f(x) = f(x^*) + f'(x^*)(x - x^*) + \frac{f''(x^*)}{2}(x - x^*)^2, \qquad \text{con} \quad x^* = a^{-1}b$$

donde

$$f(x^*) = c - \frac{b^2}{2a}, \qquad f'(x^*) = 0, \qquad f''(x^*) = a\,.$$

Por lo tanto

$$f(x) - f(x^*) = \frac{f''(x^*)}{2}(x - x^*)^2 = \frac{a}{2}\left(x - \frac{b}{a}\right)^2. \tag{3.3}$$

De esta última igualdad encontramos que si $a \neq 0$, entonces para toda $x \neq x^*$

$$\begin{aligned} f(x) - f(x^*) &> 0, \quad \text{si} \quad a > 0, \\ f(x) - f(x^*) &< 0, \quad \text{si} \quad a < 0. \end{aligned}$$

Es decir, para la función cuadrática (3.1)

$$x^* = a^{-1}b \quad \text{es un mínimo global sí} \quad a > 0, \tag{3.4}$$

$$x^* = a^{-1}b \quad \text{es un máximo global sí} \quad a < 0. \tag{3.5}$$

De hecho, por condiciones suficientes de primer orden, sabemos que éste es el único punto crítico. Pero para la función cuadrática podemos decir algo más: Las condiciones (3.4)–(3.5) no solo son necesarias sino que también son suficientes debido a (3.3). Lo anterior puede resumirse en el siguiente resultado:

Teorema 3.1. *Cualquier función cuadrática de una variable se puede escribir en la forma (3.1) o (3.2) con $a \neq 0$. El único punto crítico corresponde al vértice de la parábola, cuya abscisa es $x^* = a^{-1}b$. Este punto crítico*

1. *Es el único mínimo global sí, y solo sí, $a > 0$.*

2. *Es el único máximo global sí, y solo sí, $a < 0$.*

3.1.2. Funciones cuadráticas de varias variables

Cualquier función cuadrática multivariada de valores reales, $f : \mathbb{R}^n \to \mathbb{R}$, se puede expresar en la forma

$$f(\mathbf{x}) = \frac{1}{2}\,\mathbf{x}^T A\,\mathbf{x} - \mathbf{x}^T\,\mathbf{b} + c, \tag{3.6}$$

en donde

- $\mathbf{x} = (x_1, \ldots, x_n)^T$ es el vector de variables independientes.
- $A \in \mathbb{R}^{n\times n}$ es una matriz cuadrada simétrica con coeficientes reales a_{ij}, $1 \leq i, j \leq n$.
- $\mathbf{b} = (b_1, \ldots, b_n)^T$ es un vector constante.
- c es un escalar.

En las expresiones anteriores utilizamos la notación matricial, de tal forma que el superíndice T se utiliza para indicar la transpuesta de una matriz o vector. Así, para indicar que $\mathbf{x}$ es un vector columna, escribimos el vector de sus coeficientes en forma de reglón y al final lo transponemos. Obsérvese que, por lo tanto,

$$\mathbf{x}^T\,\mathbf{b} = \langle \mathbf{x},\, \mathbf{b}\rangle = \sum_{i=1}^{n} x_i\, b_i,$$

$$\mathbf{x}^T A\,\mathbf{x} = \langle \mathbf{x},\, A\,\mathbf{x}\rangle = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}\, x_i\, x_j$$

es el producto escalar (interior) usual en el espacio n–dimensional $\mathbb{R}^n$. El lector interesado en profundizar en el estudio de las propiedades algebraicas de las funciones cuadráticas, puede consultar textos de álgebra lineal bajo el tema de *formas cuadráticas*.

Las funciones cuadráticas de varias variables son mucho más ricas en posibilidades que las funciones cuadráticas en una dimensión, pero obviamente ofrecen mayores dificultades en su estudio. Por ejemplo, puede que la función cuadrática no tenga un punto crítico, o bien que tenga un punto crítico pero que éste no corresponda a un maximizador o minimizador de la función. Para estudiar estas propiedades primero calculamos el gradiente $\nabla f(\mathbf{x})$, el Jacobiano $J_f(\mathbf{x})$ (matriz de primeras derivadas en $\mathbb{R}^{1\times}n$) y el Hessiano $H_f(\mathbf{x})$ (Matriz de segundas derivadas en $\mathbb{R}^{n\times n}$) de la función cuadrática (3.6):

$$\nabla f(\mathbf{x}) = A\,\mathbf{x} - \mathbf{b}, \tag{3.7}$$

$$J_f(\mathbf{x}) = (\nabla f(\mathbf{x}))^T, \tag{3.8}$$

$$H_f(\mathbf{x}) = A. \tag{3.9}$$

Se dice que $\mathbf{x}^* \in \mathbb{R}^n$ es un punto crítico de $f(\mathbf{x})$ si $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

- Un punto $\mathbf{x}^* \in \mathbb{R}^n$ es un punto crítico de la función cuadrática (3.6) si resuelve el sistema de ecuaciones $A\,\mathbf{x} = \mathbf{b}$. Esta claro que si $\mathbf{b} \neq \mathbf{0}$, entonces esto solo ocurre cuando A es una matriz no singular. Por lo tanto, las funciones cuadráticas de varias variables no siempre tienen puntos críticos.

- Si la matriz A es invertible, entonces la función cuadrática (3.6) tiene un único punto crítico $x^* = A^{-1}\mathbf{b}$. Obsérvese la analogía con el caso unidimensional.

Estamos interesados en explotar las analogías con el caso unidimensional para deducir condiciones necesarias y suficientes para la existencia de un mínimo (máximo) global, suponiendo que este existe. Supongamos que $\mathbf{x}^*$ es un punto crítico de la función cuadrática (3.6), y consideremos su expansión alrededor de este punto crítico:

$$f(\mathbf{x}) = f(\mathbf{x}^*) + J_f(\mathbf{x}^*)\,(\mathbf{x}-\mathbf{x}^*) + \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T\,H_f(\mathbf{x}^*)\,(\mathbf{x}-\mathbf{x}^*), \qquad \text{con} \qquad \mathbf{x}^* = A^{-1}\mathbf{b},$$

donde

$$f(\mathbf{x}^*) = c - \frac{1}{2}\left(A^{-1}\,\mathbf{b}\right)^T\,\mathbf{b}, \qquad J_f(\mathbf{x}^*) = (\nabla f(\mathbf{x}^*))^T = (\mathbf{0})^T, \qquad H_f(\mathbf{x}^*) = A\,. \tag{3.10}$$

De modo que, en forma análoga a (3.3), encontramos que

$$f(\mathbf{x}) - f(\mathbf{x}^*) = \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T\,A\,(\mathbf{x}-\mathbf{x}^*)\,. \tag{3.11}$$

Sin embargo, a diferencia del caso unidimensional, si A es no singular, encontramos tres casos para el punto crítico. Para $\mathbf{x} \neq \mathbf{x}^* = A^{-1}\mathbf{b}$:

$$\begin{aligned}
&f(\mathbf{x}) - f(\mathbf{x}^*) > 0, \quad \text{si } A \text{ es definida positiva,}\\
&f(\mathbf{x}) - f(\mathbf{x}^*) < 0, \quad \text{si } A \text{ es definida negativa,}\\
&f(\mathbf{x}) - f(\mathbf{x}^*) \quad \text{puede tomar valores positivos y negativos si } A \text{ es indefinida.}
\end{aligned}$$

Es decir, para la función cuadrática (3.6)

$$\mathbf{x}^* = A^{-1}\mathbf{b} \quad \text{es un mínimo global sí } A \text{ es definida positiva,} \tag{3.12}$$

$$\mathbf{x}^* = A^{-1}\mathbf{b} \quad \text{es un máximo global sí } A \text{ es definida negativa,} \tag{3.13}$$

$$\mathbf{x}^* = A^{-1}\mathbf{b} \quad \text{es un punto silla si } A \text{ es indefinida.} \tag{3.14}$$

Al igual que en la caso unidimensional, obtenemos que estas condiciones son necesarias y suficientes para la existencia de un mínimo o máximo global cuando A es invertible y definida positiva o negativa, respectivamente. Lo anterior puede resumirse en el siguiente resultado:

Teorema 3.2. *Cualquier función cuadrática se puede escribir en la forma (3.6). Sí además* $\mathbf{x}^*$ *es un punto crítico y* A *es no singular, entonces la función cuadrática se puede escribir en la forma (3.11). Este único punto crítico es igual a* $x^* = A^{-1}\mathbf{b}$*. Además, el punto crítico*

1. *Es el único mínimo global sí, y solo sí,* A *es definida positiva.*
2. *Es el único máximo global sí, y solo sí,* A *es definida negativa.*
3. *Es un punto silla sí, y solo sí,* A *es indefinida.*

El siguiente resultado será muy útil para encontrar soluciones de mínimos cuadrados:

Corolario 3.3. *Sea* A *una matriz simétrica y definida positiva, entonces* $\mathbf{x}^*$ *es un mínimo global estricto de (3.6) sí, y solo sí,* $\mathbf{x}^*$ *resuelve el sistema de ecuaciones* $A\,\mathbf{x} = \mathbf{b}$*.*

Por lo tanto, el punto mínimo se puede encontrar resolviendo el sistema de ecuaciones por medio del método de Choleski introducido en el Capítulo 1. Sin embargo, cuando la matriz A es muy grande conviene utilizar métodos iterativos, principalmente por razones de memoria y de eficiencia computacional.

3.2. Métodos iterativos para minimizar funciones cuadráticas

Sea A una matriz simétrica y definida positiva de orden n, y sea $\mathbf{b}$ un vector en $\mathbb{R}^n$. Consideremos el problema de optimización sin restricciones

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \equiv \frac{1}{2}\,\mathbf{x}^T A\,\mathbf{x} - \mathbf{b}\,\mathbf{x} \tag{3.15}$$

La constante c de la función (3.6) es irrelevante para este problema, dado que el punto mínimo es el mismo cualesquiera que sea esta constante. Por lo tanto sin pérdida de generalidad podemos asumir que $\mathbf{c} = \mathbf{0}$.

Nuestro interés es introducir métodos iterativos o de recurrencia que nos permitan encontrar la solución. Debido a que el problema (3.15) es equivalente a la solución del sistema de ecuaciones

$$A\,\mathbf{x} = \mathbf{b}, \quad \text{con } A \text{ una } \textbf{matriz simética y definida positiva}, \tag{3.16}$$

entonces los métodos iterativos también serán útiles para encontrar soluciones aproximadas de sistemas de ecuaciones lineales con matrices simétricas y definidas positivas. Afortunadamente en un gran número de aplicaciones los sistemas de ecuaciones asociados tienen

matrices de este tipo, y usualmente son de gran tamaño. De ahí la importancia de considerar su estudio por medio de diferentes métodos. En este estudio solo consideraremos dos métodos iterativos del grupo de los llamados ***métodos de descenso***. Para un estudio exhaustivo recomendamos al lector los libros [6] y [7].

En un método iterativo se propone un valor inicial (generalmente arbitrario) $\mathbf{x}^0$, y después se genera un conjunto de valores $\mathbf{x}^1, \ldots, \mathbf{x}^m$ por medio de una iteración de la forma

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \, \mathbf{d}^k. \tag{3.17}$$

donde

- $\mathbf{d}^k$ es la dirección de búsqueda.
- α_k es el mínimo en la dirección de búsqueda, dado por

$$\alpha_k = \arg \min_{\alpha} \; f(\mathbf{x}^k + \alpha \, \mathbf{d}^k). \tag{3.18}$$

Esta última condición asegura que el método sea de descenso, es decir que

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k)$$

3.2.1. Método de descenso máximo

En 1847 Cauchy publicó un artículo en una revista francesa con el título *Métodos generales para la resolución de los sistemas de ecuaciones simultaneas* [8], en donde propone el uso del negativo del gradiente como dirección de descenso. El propone

$$\mathbf{d}^k = -\mathbf{g}^k, \quad \text{donde} \quad \mathbf{g}^k = \nabla f(\mathbf{x}^k)). \tag{3.19}$$

El algoritmo completo, tomando un criterio de paro, es:

Algoritmo de descenso máximo

1. **Inicialización**: Dado $\mathbf{x}^0$, calcular $\mathbf{g}^0 = \nabla f(\mathbf{x}^0)$.
2. **Descenso**: Suponiendo conocidos $\mathbf{x}^k$, $\mathbf{g}^k$, calcular $\mathbf{x}^{k+1}$, $\mathbf{g}^{k+1}$ por medio de

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \, \mathbf{g}^k, \quad \text{donde} \quad \alpha_k = \arg \min_{\alpha} f(\mathbf{x}^k - \alpha \, \mathbf{g}^k),$$
$$\mathbf{g}^{k+1} = \nabla f(\mathbf{x}^{k+1}).$$

3. **Prueba de convergencia**:

   Si $\langle \mathbf{g}^{k+1}, \mathbf{g}^{k+1} \rangle \leq \epsilon \langle \mathbf{g}^0, \mathbf{g}^0 \rangle$, tomar $\mathbf{x}^* = \mathbf{x}^{k+1}$ y parar.

   En caso contrario, hacer $k = k + 1$ y volver a 2.

En este algoritmo el valor de α_k se calcula minimizando la función de una variable $\phi(\alpha) = f(\mathbf{x}^k - \alpha\, \mathbf{g}^k)$, ya sea en forma exacta (cuando esto es posible) o bien en forma aproximada. Obsérvese que el método es general, es decir calcula en forma aproximada mínimos de funciones arbitrarias siempre que esto sea posible. Sin embargo, para el caso de funciones cuadráticas el método se puede simplificar debido a que es posible calcular de manera exacta el paso α_k, obteniendo

$$\alpha_k = \langle \mathbf{g}^k, \mathbf{g}^k \rangle / \langle \mathbf{g}^k, A\, \mathbf{g}^k \rangle. \tag{3.20}$$

Algoritmo de descenso máximo para funciones cuadráticas

1. **Inicialización**: Dado $\mathbf{x}^0$: $\mathbf{g}^0 = A\, \mathbf{x}^0 - \mathbf{b}$.

2. **Descenso**: Suponiendo conocidos $\mathbf{x}^k$, $\mathbf{g}^k$, calcular $\mathbf{x}^{k+1}$, $\mathbf{g}^{k+1}$

$$\begin{aligned} \mathbf{x}^{k+1} &= \mathbf{x}^k - \alpha_k\, \mathbf{g}^k, \quad \text{donde} \quad \alpha_k = \langle \mathbf{g}^k, \mathbf{g}^k \rangle / \langle \mathbf{g}^k, A\, \mathbf{g}^k \rangle. \\ \mathbf{g}^{k+1} &= \mathbf{g}^k - \alpha_k\, A\, \mathbf{g}^k. \end{aligned}$$

3. **Prueba de convergencia**:

   Si $\langle \mathbf{g}^{k+1}, \mathbf{g}^{k+1} \rangle \leq \epsilon \langle \mathbf{g}^0, \mathbf{g}^0 \rangle$, tomar $\mathbf{x}^* = \mathbf{x}^{k+1}$ y parar.

   E caso contrario, hacer $k = k + 1$ y volver a 2.

Ejemplo 3.4. *Consideremos la función cuadrática* $f : \mathbb{R}^2 \to \mathbb{R}$ *definida por* $f(x_1, x_2) = 2\, x_1^2 + 2\, x_1\, x - 2 + 3\, x_2^2 - 4\, x_1 + 8\, x_2 + 12$. *Esta función tiene un mínimo global en* $\mathbf{x}^* = (x_1^*, x_2^*) = (2, -2)$. *Su matriz Hessiana* A *y vector* $\mathbf{b}$ *son*

$$A = \begin{bmatrix} 4 & 2 \\ 2 & 6 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 4 \\ 8 \end{bmatrix}.$$

Obsérvese que A es simétrica. Aplicando el método de descenso máximo encontramos que para reducir el valor de $||\mathbf{g}^k||$ a 2.0386×10^{-16} se necesitaron 43 iteraciones. En la Figura 3.2.1 se muestran tres pasos de la iteración de descenso máximo sobre la superficie $z = f(x, y)$ con $x = x_1$ y $y = x_2$, y sobre las líneas de contorno de f en el plano. Desgraciadamente este método converge muy lentamente a la solución exacta. La estimación teórica del error se proporciona en términos de la una norma denominada A–norma y que es definida por

Figura 3.1: Visualización de tres pasos del método de descenso máximo.

$||\mathbf{x}||_A^2 = \mathbf{x}^T A\,\mathbf{x}$. En esta norma la estimación del de la disminución del error de una iteración a la siguiente esta dada por

$$||\mathbf{x}^{k+1} - \mathbf{x}^*||_A \leq \left(\frac{r-1}{r+1}\right) ||\mathbf{x}^k - \mathbf{x}^*||_A, \qquad r = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}. \tag{3.21}$$

A los métodos iterativos en donde la reducción del error es proporcional al error en la iteración anterior, se les denomina métodos con convergencia lineal. Se observa que para matrices mal condicionadas, en donde $r >> 1$, la disminución del error de un paso al siguiente puede ser realmente muy pequeño. Existen métodos que convergen más rápido como, por ejemplo, aquellos en donde la reducción del error en una iteración es proporcional al cuadrado del error en el paso anterior. Cuando el error se comporta de esta manera, se dice que el método tiene convergencia cuadrática. Es bien conocido por los investigadores que los métodos con convergencia cuadrática, como el método de Newton, son caros computacionalmente, especialmente para problemas grandes (n grande). Por eso es que se prefieren utilizar variantes más económicas que, aunque pierdan la convergencia cuadrática, sean superiores a los métodos con convergencia lineal. Por ejemplo, el método de gradiente conjugado es un método intermedio entre el método de descenso máximo y el método de Newton.

3.2.2. Método de gradiente conjugado

El método de gradiente conjugado fue introducido por Magnus R. Hestenes y Eduard Stiefel en 1952 para resolver sistemas lineales con matrices simétricas y definidas positivas, [9]. Actualmente este algoritmo y sus variantes son de los métodos más exitosos para el cómputo científico de gran escala, principalmente sistemas de ecuaciones lineales del orden de cientos de miles y millones de ecuaciones, así como la optimización de funciones y funcionales cuadráticos (dimensión finita e infinita, respectivamente). La ascendencia y reputación del algoritmo de gradiente conjugado es tal que, en el año 2000 fue considerado uno de los algoritmos más destacados del siglo XX, por el Instituto Americano de Física y la Sociedad de Computación del Instituto de Investigaciones Eléctricas y Electrónica de los Estados Unidos, [10].

Una de las propiedades cruciales del método de gradiente conjugado es su habilidad para generar de manera muy económica un conjunto de vectores de descenso, denominados

vectores conjugados. Dada una matriz cuadrada, un conjunto de vectores $\{\mathbf{d}^0, \mathbf{d}^1, \ldots, \mathbf{d}^m\}$ se denomina A–***conjugado*** si

$$\langle \mathbf{d}^k, \mathbf{d}^l \rangle_A \doteq \langle \mathbf{d}^k, A\,\mathbf{d}^l \rangle = (\mathbf{d}^k)^T A\,\mathbf{d}^l = 0, \qquad \text{si} \quad k \neq l.$$

La importancia de este tipo de vectores radica en que se puede minimizar la función cuadrática $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\,\mathbf{x} - \mathbf{x}^T\mathbf{b}$, definida de $\mathbb{R}^n$ a $\mathbb{R}$ en a lo más n iteraciones. La minimización procede a lo largo de estas direcciones conjugadas, de tal manera que después de n iteraciones de descenso del tipo (3.17) se obtiene

$$\mathbf{x}^* = \mathbf{x}^0 + \alpha_0\mathbf{d}^0 + \cdots + \alpha_{n-1}\mathbf{d}^{n-1}.$$

el cual debe ser el punto mínimo de la función cuadrática debido a que las direcciones conjugadas siempre son linealmente independientes.

Generación de direcciones conjugadas

Todo lo que necesitamos para generar el algoritmo de gradiente conjugado es saber como generar las direcciones conjugadas. Hay algunas formas, como las siguientes

- Si A es una matriz simétrica, sus **vectores propios** $\mathbf{v}_1, \ldots, \mathbf{v}_n$ son A–conjugados y también mutuamente ortogonales. Sin embargo, computacionalmente es más caro calcular vectores propios que resolver el sistema de ecuaciones directamente, por lo que este método no es adecuado desde el punto de vista numérico.

- **Algoritmo de Gram–Schmidt**. Se puede adaptar este método de ortogonalización para generar las direcciones conjugadas a partir de los vectores canónicos. El problema es que este método también es costoso para problemas de gran escala, aparte de ser inestable.

- **Idea de Hestenes y Stiefel**. La idea de estos investigadores fue calcular la dirección conjugada $\mathbf{d}^k$ usando solamente una combinación lineal del residual (gradiente) y la dirección conjugada anterior, en la forma siguiente

$$\mathbf{d}^k = -\mathbf{g}^k + \beta_k\mathbf{d}^{k-1}$$
$$(\mathbf{d}^{k-1})^T A\,\mathbf{d}^k = 0 \quad \Longleftrightarrow \beta_k = \frac{\langle \mathbf{g}^k, A\,\mathbf{d}^k \rangle}{\langle \mathbf{d}^k, A\,\mathbf{d}^k \rangle}$$

Este último método es realmente el que se utiliza debido a su gran simplicidad y efectividad, además de que genera progresivamente las direcciones conjugadas conforme se avanza en

las iteraciones. El método completo de gradiente conjugado para funciones cuadráticas se muestra a continuación.

Algoritmo de gradiente conjugado

1. **Inicialización**: Dado $\mathbf{x}^0$, calcular $\mathbf{g}^0 = A\,\mathbf{x}^0 - \mathbf{b}$ y $\mathbf{d}^0 = -\mathbf{g}^0$.

2. **Descenso**: Suponiendo conocidos $\mathbf{x}^k$, $\mathbf{g}^k$, $\mathbf{d}^k$, calcular $\mathbf{x}^{k+1}$, $\mathbf{g}^{k+1}$, $\mathbf{d}^{k+1}$ por medio de

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k\,\mathbf{d}^k, \qquad \text{con} \qquad \alpha_k = \frac{-\langle \mathbf{g}^k,\, \mathbf{d}^k\rangle}{\langle \mathbf{d}^k,\, A\,\mathbf{d}^k\rangle}$$
$$\mathbf{g}^{k+1} = A\,\mathbf{x}^{k+1} - \mathbf{b}$$

3. **Prueba de convergencia y nueva dirección conjugada**:

$$\text{Si} \quad \langle \mathbf{g}^{k+1},\, \mathbf{g}^{k+1}\rangle \leq \epsilon \langle \mathbf{g}^0,\, \mathbf{g}^0\rangle \quad \text{tomar} \quad \mathbf{x}^* = \mathbf{x}^{k+1} \quad \text{y parar.}$$

   En caso contrario, tomar

$$\mathbf{d}^{k+1} = -\mathbf{g}^{k+1} + \beta_k\,\mathbf{d}^k, \qquad \text{con} \qquad \beta_k = \frac{\langle \mathbf{g}^{k+1},\, A\,\mathbf{d}^k\rangle}{\langle \mathbf{d}^k,\, A\,\mathbf{d}^k\rangle}.$$

   Hacer $k = k + 1$ y volver a 2.

Este método básico se puede mejorar para calcular de manera equivalente pero más eficiente los escalares α_k y β_k, así como calcular el vector residual $\mathbf{g}^{k+1}$ en términos del anterior. El lector interesado en los detalles puede consultar las referencias dadas anteriormente. Aquí solo presentaremos la versión práctica del método.

Algoritmo práctico de gradiente conjugado

1. **Inicialización**: Dado $\mathbf{x}^0$, calcular $\mathbf{g}^0 = A\,\mathbf{x}^0 - \mathbf{b}$ y $\mathbf{d}^0 = -\mathbf{g}^0$.

2. **Descenso**: Suponiendo conocidos $\mathbf{x}^k$, $\mathbf{g}^k$, $\mathbf{d}^k$, calcular $\mathbf{x}^{k+1}$, $\mathbf{g}^{k+1}$, $\mathbf{d}^{k+1}$ por medio de

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k\,\mathbf{d}^k, \qquad \text{con} \qquad \alpha_k = \frac{\langle \mathbf{g}^k,\, \mathbf{g}^k\rangle}{\langle \mathbf{d}^k,\, A\,\mathbf{d}^k\rangle}$$
$$\mathbf{g}^{k+1} = \mathbf{g}^k + \alpha_k A\,\mathbf{d}^k$$

3. **Prueba de convergencia y nueva dirección conjugada**:

$$\text{Si} \quad \langle \mathbf{g}^{k+1},\, \mathbf{g}^{k+1}\rangle \leq \epsilon \langle \mathbf{g}^0,\, \mathbf{g}^0\rangle \quad \text{tomar} \quad \mathbf{x}^* = \mathbf{x}^{k+1} \quad \text{y parar.}$$

   En caso contrario, tomar

$$\mathbf{d}^{k+1} = -\mathbf{g}^{k+1} + \beta_k\,\mathbf{d}^k, \qquad \text{con} \qquad \beta_k = \frac{\langle \mathbf{g}^{k+1},\, \mathbf{g}^{k+1}\rangle}{\langle \mathbf{g}^k,\, \mathbf{g}^k\rangle}.$$

   Hacer $k = k + 1$ y volver a 2.

Ejemplo 3.5. *Consideremos de nuevo la función cuadrática de dos variables del ejemplo anterior,* $f(x_1, x_2) = 2\,x_1^2 + 2\,x_1\,x - 2 + 3\,x_2^2 - 4\,x_1 + 8\,x_2 + 12$*, con punto mínimo* $\mathbf{x}^* = (2,\, -2)$*. Aplicamos el método de gradiente conjugado para encontrar este mínimo, tomando como valor de comienzo* $\mathbf{x}^0 = (0,\, 0)^T$*.*

La Figura 3.2 muestra las líneas de nivel de la función con las iteraciones de gradiente conjugado. Se observa que el algoritmo converge en dos iteraciones a la solución exacta. Recordemos que el algoritmo de descenso máximo debe realizar 43 iteraciones para reducir el gradiente al orden de 2×10^{-16}. Obviamente el algoritmo de gradiente conjugado es superior al de descenso máximo.

Figura 3.2: Iteraciones del algorimo de gradiente conjugado.

En suma, el método de gradiente conjugado, es un método iterativo de descenso el cual genera las direcciones conjugadas de manera muy económica. En cada paso,

$$\mathbf{x}^k \text{ minimiza } f(\mathbf{x}) \text{ sobre } \mathbf{x}^0 + Span\left\{\mathbf{d}^0,\, \mathbf{d}^1,\, \ldots\, \mathbf{d}^{k-1}\right\}.$$

Si la matriz A tiene solo p diferentes valores propios, entonces el método de gradiente conjugado converge en a lo más p iteraciones, y la estimación del error viene dada por

$$||\mathbf{x}^{k+1} - \mathbf{x}^*||_A \leq 2\left(\frac{\sqrt{r}-1}{\sqrt{r}+1}\right)||\mathbf{x}^k - \mathbf{x}^*||_A.$$

con $r = \lambda_{max}(A)/\lambda_{min}(A)$ el condicionamiento de la matriz.

Finalmente, conviene mencionar que es posible acelerar sustancialmente la convergencia del método de gradiente conjugado, especialmente con sistemas muy grandes. Es se realiza por medio de una técnica denominada ***precondicionamiento***, la cual consiste en transformar el sistema de ecuaciones en uno equivalente cuya nueva matriz tenga una distribución más compacta de sus valores propios. Es decir, que su número de condición r sea mucho más pequeño que la matriz del sistema original. En este curso no abordaremos esta técnica, el lector interesado puede estudiarla en las referencias [4], [11]. El método de gradiente conjugado también adaptarse para resolver problemas de optimización no lineales sin necesidad de almacenar Hessianos y extenderse sin problema a problemas elípticos en dimensión infinita. Al final de este capítulo veremos un ejemplo.

3.3. Mínimos cuadrados y funciones cuadráticas. Problemas en dimensión infinita

3.3.1. La relación entre ambos problemas

Retomando el problema de mínimos cuadrados

$$\min_{c\in\mathbb{R}^n} \frac{1}{2}\|A\,\mathbf{x}-\mathbf{b}\|^2\,, \tag{3.22}$$

obsérvese que

$$\frac{1}{2}\|A\,\mathbf{x}-\mathbf{b}\|^2 = \frac{1}{2}\langle A\,\mathbf{x},\, A\,\mathbf{x}\rangle - \langle A\,\mathbf{x},\, \mathbf{b}\rangle + \frac{1}{2}\langle \mathbf{b},\, \mathbf{b}\rangle = \frac{1}{2}\langle \mathbf{x},\, A^T A\,\mathbf{x}\rangle - \langle \mathbf{x},\, A^T\mathbf{b}\rangle + \frac{1}{2}\langle \mathbf{b},\, \mathbf{b}\rangle.$$

Es decir,

$$f(\mathbf{x}) = \frac{1}{2}\|A\,\mathbf{x}-\mathbf{b}\|^2 = \frac{1}{2}\mathbf{x}^T(A^T A)\,\mathbf{x} - \mathbf{x}^T\,A^T\mathbf{b} + \mathbf{c}. \tag{3.23}$$

Esta última igualdad indica que el cuadrado de la norma del residual es una función cuadrática. Sabemos que si la matriz A es de rango completo, entonces $A^T A$ es simétrica y definida positiva. Por lo tanto, se obtiene inmediatamente el siguiente resultado:

Teorema 3.6. *Si $A \in \mathbb{R}^{m\times n}$ con $m > n$ es una matriz de rango completo, el problema de mínimos cuadrados (3.22) equivale a minimizar la función cuadrática (3.23). El punto mínimo $\mathbf{x}^*$ resuelve el sistema de ecuaciones normales*

$$A^T A\,\mathbf{x} = A^T\mathbf{b}, \tag{3.24}$$

y es posible encontrarlo por un método directo, como el algoritmo de Choleski, o por un método iterativo, como el algoritmo de gradiente conjugado.

Algoritmo de gradiente conjugado para resolver el problema de mínimos cuadrados.

Dados A de $m \times n$ de rango completo y $b \in \mathbb{R}^m$

1 Calcular y el vector $A^T b$ y la matriz de ecuaciones normales $A^T A$.

1. **2]** Escoger un valor inicial $\mathbf{x}^0$, un valor ϵ para el criterio de paro de las iteraciones, y un valor $kmax$ para el número máximo de iteraciones permitidas.

2. **1]** Utilizar el algoritmo iterativo de gradiente conjugado de la sección anterior, con la matriz $A^T A$ en lugar de A.

Nota importante. Queremos hacer énfasis que este método no es más eficiente que el método de Choleski o el método QR, para resolver problemas de mínimos cuadrados, cuando el número de variables nos es grande. Por ejemplo, para los problemas de mínimos cuadrados del Capítulo 2, dado que los sistemas correspondientes son de tamaño pequeño, el método de gradiente conjugado no es la mejor opción. Sin embargo, el método de gradiente conjugado proporciona algoritmos más eficientes en problemas donde intervienen cientos de miles o millones de variables o ecuaciones. Estos problemas de gran escala generalmente aparecen en la modelación de fenómenos por medio de ecuaciones diferenciales o ecuaciones integrales, como son la dinámica molecular, los sistemas complejos, aplicaciones industriales, nanotecnología, problemas inversos, sistemas atmosféricos y biológicos, entre muchos otros. Las simulaciones correspondientes se realizan en supercomputadoras o sistemas en paralelo que permiten la realización de un gran volumen de cálculos en forma muy veloz. La discusión de problemas de este tipo no forma parte de los objetivos del presente escrito, pero cerraremos el presente trabajo presentando dos problemas de mínimos cuadrados, definidos en espacios de dimensión infinita, con aplicaciones a problemas reales, y en donde se utiliza el método de gradiente conjugado como parte del método de solución.

3.3.2. Problema inverso de la ecuación de calor

Los problemas inversos usualmente están relacionados con la determinación de las causas que provocan un efecto observado ó deseado. Los problemas inversos surgen en disciplinas tales como geofísica, imagen médica (como por ejemplo en la tomografía axial computerizada), sensores remotos, tomografía acústica oceánica, pruebas no invasivas o destructivas, astronomía. Sin embargo, éstos normalmente son problemas ***mal planteados en el sentido de Hadamard***. Es decir, que en estos problemas hay pérdida de existencia de soluciones, pérdida de unicidad y/o pérdida de dependencia continúa de las soluciones respecto de los datos. En el estudio de problemas inversos, la pérdida de existencia y unicidad no constituyen problemas graves. Es la pérdida de dependencia continua, respecto de los datos, la que origina los mayores inconvenientes debido a que, en tal caso, pequeños errores o ruidos en los datos pueden producir una amplificación del error muy grande en las soluciones del problema inverso. Para abordar el mal planteamiento de los problemas inversos se utilizan los métodos de regularización, los cuales buscan restaurar la estabilidad de las soluciones con relación a los datos. En términos generales la regularización es una aproximación de un problema mal planteado por medio de una familia cercana de problemas bien planteados.

Algunos avances en la ciencia y la tecnología han sido posibles gracias a la solución de problemas inversos, y por tanto, el campo de estos problemas es uno de los que han tenido un

mayor crecimiento en matemáticas aplicadas e industriales. No obstante, el crecimiento en la investigación de los problemas inversos también se debe al desarrollo de computadoras más poderosas y a métodos numéricos más efectivos para la solución de los problemas asociados.

Abordaremos un ejemplo de un problema inverso extremadamente mal planteado: ***el problema del calor con retroceso en el tiempo***. Para ello consideremos la región $D = (0, X) \times (0, Y)$, y sea ∂D su frontera. La dinámica de difusión o propagación del calor en esta región, dada una distribución inicial de temperaturas se modela mediante la ecuación de calor

$$\begin{aligned} u_t &= a^2(u_{xx} + u_{yy}), \text{ en } D,\, t > t_0 > 0, \\ u(x, y, t_0) &= \varphi(x, y), \text{ en } D, \text{ Condición inicial}, \\ u(x, y, t) &= 0 \text{ sobre } \partial D,\, t \leq 0, \end{aligned} \tag{3.25}$$

en donde $u = u(x, y, t)$ representa la temperatura en la posición $(x,\, y) \in D$ en el instante $t > t_0$.

El problema directo es: Dada la distribución inicial de temperaturas $\varphi(x, y)$ (estado inicial), encontrar la distribución de temperaturas en la región D en cualquier otro instante $t > t_0 \geq 0$ (estado en un tiempo posterior). Este problema se puede resolver por medio de separación de variables y series de Fourier. La solución analítica del problema es:

$$u(x, y, t) = \int_0^Y \int_0^X \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \frac{4\varphi(\xi, \zeta)}{XY}\, u_{mn}(\xi, \zeta)\, e^{-\left[\left(\frac{m\pi a}{X}\right)^2 + \left(\frac{n\pi a}{Y}\right)^2\right](t - t_0)}\, u_{mn}(x, y)\, d\xi\, d\zeta, \tag{3.26}$$

donde las funciones propias (modos de Fourier) u_{mn} se definen para todo $(x, y) \in D$ de la siguiente manera

$$u_{mn}(x,\, y) = \sin\left(\frac{m\pi x}{X}\right) \sin\left(\frac{n\pi y}{Y}\right).$$

Por lo tanto, resolver el problema directo no ocasiona ningún problema. El problema tiene solución, esta es única, y cualquier perturbación en el dato $\varphi(x, y)$ desaparece rápidamente, por el amortiguamiento debido a la exponencial negativa en (3.26).

Un problema inverso es: Dada la distribución de temperaturas en un tiempo futuro $T \geq t_0$, es decir dada $u(x, y, T) = g(x, y)$ sobre la placa D, encontrar la distribución inicial de temperaturas $\varphi(x, y) = u(x, y, t_0)$. A este problema se le denomina el **problema de la ecuación de calor con retroceso en el tiempo**. Podemos evaluar (3.26) en $t = T$ y sustituir el dato $g(x, y)$, obteniendo la siguiente ecuación, denominada ecuación de Fredholm

de primer tipo

$$(\mathbb{K}\varphi)(x,y) \equiv \int_0^Y \int_0^X \mathcal{K}(x,y,\xi,\zeta)\ \varphi(\xi,\zeta)\ d\xi\, d\zeta = g(x,y), \qquad (\mathbb{K}\ \varphi = g) \tag{3.27}$$

con kernel (núcleo)

$$\mathcal{K}(x,y,\xi,\zeta) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} \frac{4}{XY}\ u_{mn}(x,\, y)\ u_{mn}(\xi,\, \zeta)\ e^{-\left[\left(\frac{m\pi a}{X}\right)^2+\left(\frac{n\pi a}{Y}\right)^2\right](T-t_0)}. \tag{3.28}$$

Enfatizamos que, en este problema, el dato es la función $g(x,y)$ y la incógnita es la función $\varphi(\xi,\zeta)$. Así que para encontrar la incógnita debemos resolver la ecuación integral (3.27), mientras que el problema directo para encontrar la incógnita tuvimos que resolver la ecuación diferencial (3.25). Esta no es la única diferencia entre el problema directo y el problema inverso. Por ejemplo, si pudieramos despejar de alguna manera φ de (3.27), entonces podemos anticipar que el dato $g(x,y)$ se verá afectado por la misma exponencial pero con signo positivo. Esto indica que cualquier perturbación en el dato $g(x,y)$ se amplificará en forma exponencial al tratar de encontrar φ. Es decir, el problema inverso es extremadamente sensible a perturbaciones en los datos, por lo que se trata de un problema extremadamente mal condicionado.

Es necesario regularizar para resolver el problema inverso en forma estable, es decir para controlar la amplificación de posibles perturbaciones. Uno de los métodos de regularización más común es el de Tijonov, el cual consiste en minimizar

$$\left\{||\mathbb{K}\varphi - g||^2 + \alpha||\varphi||\right\} \quad \text{en lugar de} \quad ||\mathbb{K}\varphi - g||^2,$$

donde $J(\varphi) = ||\mathbb{K}\varphi - g||^2 + \alpha||\varphi||$ es un funcional cuadrático, α es el parámetro de regularización y $||\cdot||$ es una norma en el espacio de funciones adecuada. Por el momento no discutiremos específicamente de que norma se trata, y dejamos este tema para cuando abordemos el segundo problema (recuperación de campos de viento). En realidad, én la práctica, para resolver el problema inverso se debe discretizar la ecuación (3.27).

Una forma de discretizar el problema (3.27) es aproximar la integral por medio de la ***cuadratura del punto medio***. Para ello se dividen los intervalos $(0,X)$ y $(0,Y)$ en M y N subintervalos, respectivamente, y se toma $\Delta\xi = 1/M$, $\Delta\zeta = 1/N$. Se construye una malla en la región D con ***puntos medios***

$$(\xi_l,\, \zeta_p) = \left((l-\frac{1}{2})\Delta\xi,\ (p-\frac{1}{2})\Delta\zeta\right), \quad \text{para } 1 \le l \le M \text{ y } 1 \le p \le N.$$

Entonces, la ecuación integral (3.27) se transforma en el sistema lineal de $M \times N$ ecuaciones con $M \times N$ incógnitas

$$\sum_{p=1}^{\infty}\sum_{l=1}^{\infty}\mathcal{K}(x_i, y_j, \xi_l, \zeta_p)\ \varphi(\xi_l, \zeta_p)\ \Delta\xi\ \Delta\zeta = g(x_i,\ y_j), \quad \forall \text{ punto medio } (x_i,\ y_j). \tag{3.29}$$

Sí denotamos por K a la matriz del sistema, sus coeficientes son

$$[K]_{i,j,l,p} = \sum_{p=1}^{\infty}\sum_{l=1}^{\infty}\mathcal{K}(x_i, y_j, \xi_l, \zeta_p)\Delta\xi\ \Delta\zeta, \tag{3.30}$$

y esta matriz es simétrica y definida positiva, pero excesivamente mal condicionada. Cualquier perturbación en los datos se amplificará exponencialmente debido a los modos de alta frecuencia, por lo que es necesario regularizar. Regularizando el sistema mediante la estrategia de Tijonov, se encuentra que es necesario resolver el siguiente sistema de ecuaciones

$$(K^T K + \alpha I)\vec{\varphi} = K^T\,\vec{g},$$

en donde α es el parámetro de regularización, $\vec{g} = \{g(x_i,\ y_j\}_{i\,j}$ es el vector de datos sobre los puntos medios de la malla, y $\vec{\varphi} = \{\xi_l,\ \zeta_p\}_{l\,p}$ es el vector de incógnitas. Por el tamaño del sistema resultante es conveniente utilizar el método de gradiente conjugado con la matriz $A = K^T K + \alpha I$ (la cual es simétrica y definida positiva) y lado derecho $\mathbf{b} = \vec{g}$.

Ejemplo 3.7. *Consideremos la siguiente condición inicial para el problema de calor (3.25) en el cuadrado unitario ($X = 1$, $Y = 1$) con $a = 1$.*

$$\varphi(x,y)\ = \begin{cases} 2y, & si\ x \geq y,\ 1-x \geq y; \\ 2(1-x), & si\ x \geq y,\ 1-x \leq y; \\ 2(1-y), & si\ x \leq y,\ 1-x \leq y; \\ 2x, & si\ x \leq y,\ 1-x \geq y. \end{cases} \tag{3.31}$$

Supongamos que conocemos la solución $u(x, y, T) = g(x, y)$ en el tiempo T =0.01. La Figura 3.3 muestra ambas funciones. Nuestro propósito es resolver el problema inverso partiendo de $g)x,\ y) = u(x,\ y,\ 0{,}01)$, utilizando regularización de Tijonov y gradiente conjugado. Se consideran solo 3 modos de la serie (3.29). Esta serie representa kernel del problema (3.27).

En la Figura 3.4 se muestra la mejor solución recuperada, comparada con la exacta. La solución recuperada difiere de la "exacta" en dos aspectos: su amplitud es menor, y no se pueden recuperar las discontinuidades de la derivada a lo largo de las líneas diagonales del dominio. Obsérvese que, la diferencia entre la condición inicial exacta y la recuperada por regularización es pequeña, es decir, la condición inicial recuperada tiene un error del 4 % comparada con la exacta.

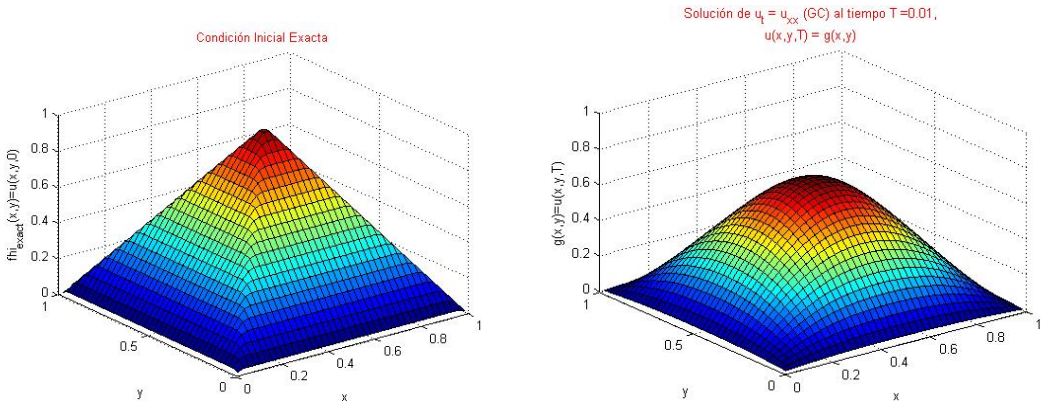

Figura 3.3: Condición inicial (izquierda) y solución en T = 0.01 (lderecha).

3.3.3. Recuperación de campos de viento en meteorología

En meteorología existen muchos problemas muy complicados en donde es necesaria la modelación y simulación computacional como por ejemplo: la dispersión de contaminantes en la atmósfera, la predicción del clima, la obtención de escenarios de contingencia, energía eólica, entre otros muchos. Los expertos en el campo utilizan modelos y programas muy sofisticados, producto del esfuerzo de muchos investigadores a nivel mundial durante muchos años. Tales programas simulan la dinámica atmosférica y para echarlos a andar es necesario alimentarlos con un estado inicial. Por ejemplo, necesitan de variables atmosféricas como la temperatura, densidad, presión, velocidad del viento, en un momento inicial. Los datos se toman de los que registran las estaciones meteorológicas y se generan mapas de las mismas en una región dada. La Figura 3.5 muestra el mapa de velocidad de viento en un instante dado en una región determinada. Las flechas indican su dirección y el tamaño su magnitud. Esta claro que estos mapas contienen información que no es exacta debido a los errores de medición y a procesos de interpolación para obtener datos en regiones y posiciones regulares dentro de una malla. Por si fuera poco, las estaciones meteorológicas generalmente no miden la componente vertical del viento, a menos de que se cuente con equipo muy sofisticado. Así que, generalmente solo se obtiene información parcial e imprecisa del campo de velocidades del viento.

Un problema de de mucho interés en meteorología es: ***¿Cómo a partir de la información horizontal del viento podemos reconstruir la componente vertical?***. Siendo un poco más ambiciosos, podríamos pedir que el campo reconstruido satisfaga propiedades de los campos de viento reales. Por ejemplo, una propiedad física muy importante que debe

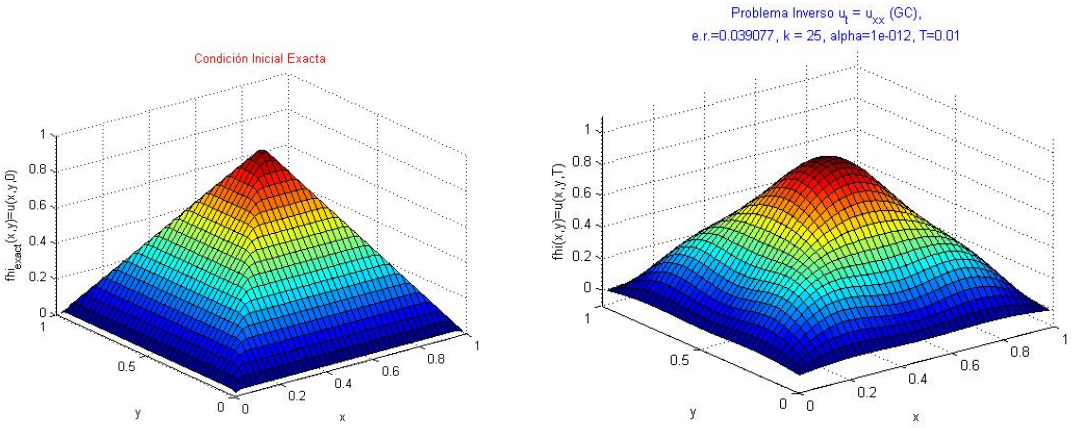

Figura 3.4: Condición incial exacta (izquierda) y recuperada(derecha).

Figura 3.5: Mapa de velocidades en una región dada.

satisfacer un campo de velocidades de viento es la de **conservación de masa**.

Conservación de masa Si $\mathbf{u}(\mathbf{x}) = (u(\mathbf{x}), v(\mathbf{x}), w(\mathbf{x}))$, denota la velocidad del viento en la posición $\mathbf{x} = (x_1, x_2, x_3)$, con funciones componentes horizontales u, v, y la función componente vertical w. Suponiendo al aire como un fluido incompresible (que no se puede comprimir), de las leyes de la dinámica de fluidos, se deduce que este campo de viento conserva la masa si satisface la ecuación

$$\boldsymbol{\nabla}\cdot\mathbf{u} \equiv \frac{\partial u}{\partial x_1} + \frac{\partial v}{\partial x_2} + \frac{\partial w}{\partial x_3} = 0. \tag{3.32}$$

en todo punto $\mathbf{x} \in \mathbb{R}^3$ de su dominio. Esta ecuación, denominada ***ecuación de continuidad***, debe ser satisfecha por el campo de velocidades reconstruido. Si denotamos por $\mathbf{u}^0$ al campo medido por las estaciones meteorológicas, éste no satisface la ecuación de continuidad, debido a los errores de medición en interpolación y debido a que no se proporciona la componente vertical.

Formulación del problema: Dado un ***campo inicial de velocidad*** $\mathbf{u}^0$ en una región Ω (obtenido de un conjunto discreto de datos proporcionados por estaciones meteorológicas),

calcular un campo vectorial $\mathbf{u}$ tal que

$$\boldsymbol{\nabla} \cdot \mathbf{u} = 0 \quad \text{en } \Omega, \quad \text{(Restricción física)} \tag{3.33}$$

$$\mathbf{u} \cdot \mathbf{n} = 0 \quad \text{sobre } \Gamma_N\,, \quad \text{(Condición de frontera)} \tag{3.34}$$

en donde $\mathbf{n}$ denota la componente normal exterior a la frontera Γ_N, y que indica la topografía del terreno. La primera condición es la ecuación de continuidad y la segunda es una condición de frontera sobre Γ_N (la topografía), e indica que el viento no tiene componente perpendicular al terreno, es decir el viento al hacer contacto con el terreno, se resbala, ver Figura 3.6. Este problema se puede reformular pensando que el campo que queremos recu-

Figura 3.6: Región del problema de recuperación del campo de viento

perar $\mathbf{u}$ se encuentra en el conjunto de funciones de divergencia libre (condición (3.33)) y que resbalan en la superficie del terreno (condición (3.34)). Para encontrar una formulación (modelo) matemática del problema debemos introducir algunos espacios de funciones.

El espacio de funciones básico es

$$L_2(\Omega) = \left\{ f : \Omega \mapsto \mathbb{R} \;\middle|\; \int_\Omega f(\mathbf{x})^2 d\mathbf{x} < \infty \right\} \tag{3.35}$$

A los elementos de este conjunto se les denominan funciones cuadrado integrables. Obsérvese que éste es un conjunto más amplio que el conjunto de funciones continuas sobre $\Omega \in \mathbb{R}^d$, donde d indica la dimensión, es decir $\mathbf{x} = (x_1, \ldots, x_d)$. Para el problema de meteorología d puede ser 2 o 3, dependiendo si se trata del caso bidimensional o tridimensional. No discutiremos aspectos técnicos o teóricos de este conjunto, como por ejemplo que la función debe ser medible y que la condición debe cumplirse salvo un conjunto de medida cero. Estos aspectos forman parte de los temas del análisis matemático y del análisis funcional. Lo importante es que este conjunto es un espacio vectorial con un producto interior definido por

$$\langle f,\, g \rangle = \int_\Omega f(\mathbf{x})\, g(\mathbf{x})\, d\mathbf{x},$$

donde la integral es una integral multidimensional: doble si $d = 2$ y triple si $d = 3$. Este producto interior produce una norma en $f \in \mathbf{L}_2(\Omega)$ que se define como $|| f || = \langle f,\, f \rangle^{1/2}$.

El siguiente paso es generalizar esta idea al caso de funciones (campos) vectoriales, que son en las que estamos interesados. Estas se definen como productos tensoriales en la siguiente forma:

$$\mathbf{L}_2(\Omega) = L_2(\Omega)^d = L_2(\Omega) \times L_2(\Omega) \cdots \times L_2(\Omega) \quad (d \text{ veces}), \tag{3.36}$$

donde $d = 2$ o $d = 3$ produce campos vectoriales bidimensionales y tridimensionales, respectivamente. El producto interno en esta ocasión es obviamente:

$$\langle \mathbf{u},\, \mathbf{v} \rangle = \int_\Omega \mathbf{u} \cdot \mathbf{v}\, d\mathbf{x} \tag{3.37}$$

donde $\mathbf{u}(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x}) = \sum_{i=1}^{d} u_i(\mathbf{x})\, v_i(\mathbf{x})$. Por último definimos el espacio de funciones en donde buscaremos el campo vectorial de velocidad del viento $\mathbf{u}$ que satisface las condiciones (3.33) y (3.34):

$$\mathbf{V} = \{\mathbf{v} \in \mathbf{L}_2(\Omega) \mid \boldsymbol{\nabla}\cdot\mathbf{v} \in L_2(\Omega),\ \boldsymbol{\nabla}\cdot\mathbf{v} = 0,\ \ \mathbf{v}\cdot\mathbf{n} = 0 \text{ on } \Gamma_N\}. \tag{3.38}$$

La razón por la que buscamos la solución en este conjunto de funciones, es que éste es un subespacio vectorial cerrado de $\mathbf{L}_2(\Omega)$ y que $\mathbf{u}^0 \in \mathbf{L}_2(\Omega)$. Esto nos permite buscar la solución como la proyección del campo inicial $\mathbf{u}^0$ en el subespacio cerrado $\mathbf{V}$, es decir nos permite formular el problema como uno de mínimos cuadrados. Como el dato $\mathbf{u}^0$ proporciona más información en la dirección horizontal y no proporciona información en la dirección vertical, los meteorólogos asignan mayor peso a la dirección horizontal que a la dirección vertical. Esto se logra introduciendo un producto interior ponderado por medio de una matriz diagonal $S \in \mathbb{R}^{d\times d}$ con coeficientes $\alpha_i > 0$, de la siguiente forma:

$$\langle \mathbf{u},\, \mathbf{v} \rangle_S = \int_\Omega S\mathbf{u} \cdot \mathbf{v}\, d\mathbf{x} = \int_\Omega \alpha_1\, u_1(\mathbf{x})\, v_1(\mathbf{x}) + \cdots + \alpha_d\, u_d(\mathbf{x})\, v_d(\mathbf{x})\, d\mathbf{x}, \tag{3.39}$$

En este caso la distancia entre $\mathbf{u}^0$ y cualquier elemento $\mathbf{v} \in \mathbf{V}$ es:

$$||\mathbf{v} - \mathbf{u}^0||_S^2 = \int_\Omega S(\mathbf{v} - \mathbf{u}^0) \cdot (\mathbf{v} - \mathbf{u}^0)\, d\mathbf{x} \tag{3.40}$$

Como nos interesa aquel campo $\mathbf{v} \in \mathbf{V}$ que minimiza el cuadrado de la distancia, entonces nuestro problema consiste en resolver el:

Problema de mínimos cuadrados. Dado $\mathbf{u}^0$ en $\mathbf{L}_2(\Omega)$, encontrar el campo vectorial $\mathbf{u}$ en $\mathbf{V}$ que minimice el funcional

$$J(\mathbf{v}) = \frac{1}{2}\, ||\mathbf{v} - \mathbf{u}^0||_S^2 = \frac{1}{2} \int_\Omega S(\mathbf{v} - \mathbf{u}^0) \cdot (\mathbf{v} - \mathbf{u}^0)\, d\mathbf{x}. \tag{3.41}$$

Como los espacios de funciones vectoriales $\mathbf{L}_2(\Omega)$ y $\mathbf{V}$ son de dimensión infinita, entonces éste es un problema de mínimos cuadrados en dimensión infinita. Obsérvese la similitud del funcional con

Para resolver este problema computacionalmente es necesario primero reducir este problema a uno de dimensión finita, por medio de la aproximación de los espacios de funciones en donde se busca la solución. El resultado, al final, consiste en la minimización con restricciones de un funcional cuadrático sobre $\mathbb{R}^n$ en donde n puede ser del orden de miles (en el caso bidimensional) o cientos de miles (en el caso tridimensional). Para resolver el problema aproximado, es posible aplicar las técnicas de solución ya vistas para funciones cuadráticas. Como los sistemas resultantes son muy grandes, el método de gradiente conjugado es un método muy conveniente en este caso. Para ver los detalles del método de aproximación y los algoritmos de solución se pueden consultar la referencias [13], [14]. A continuación se muestran algunos resultados.

Ejemplo 3.8. *Consideremos el campo vectorial bidimensional* $\mathbf{u}^0(x,\, z) = (x,\, 0)$ *definido sobre* $\Omega = (1,\, 2)\times(0,\, 1)$*. En este caso es muy fácil ver directamente que el complemento para obtener el campo ajustado es* $\mathbf{w} = (0,\, -z)$*. Es decir, la solución es* $\mathbf{u} = \mathbf{u}^0 + \mathbf{w} = (x,\, -z)$*. Aplicando el algoritmo de aproximación utilizado en [13] y [14] al dato inicial* $\mathbf{u}^0$ *se obtiene el campo mostrado en la Figura 3.7. Se escogieron los parámetros* $\alpha_1 = 1$*,* $\alpha_3 = 0.001$ *para la matriz* S*. La aproximación se realizó sobre una malla de* 80×80*, lo cual produce sistemas del orden de 6,400 ecuaciones con 6,400 incógnitas. La diferencia relativa entre la solución exacta y la aproximada es de* 5.9×10^{-4}*, es decir el error es del orden de 0.05 por ciento.*

Figura 3.7: Campo ajustado. El error respecto del exacto es del 0.05 %.

En la Figura 3.8 se muestran los resultados de otros dos ejemplos, en donde el campo vectorial bidimensional ya no es tan trivial, debido a que la topografía es más complicada: un campo vectorial con una elevación real del terreno, y otro con una elevación de tipo cosinusoidal. En cada caso se recuperó el campo vectorial de la componente vertical de los mismos. Los parámetros utilizados son los mismos y la discretización de los problemas es análoga a la del ejemplo anterior.

Ejemplo 3.9. *Por último, presentamos un ejemplo donde se obtiene la recuperación de un campo tridimensional. El campo inicial tomado es* $\mathbf{u}^0(x,\, y,\, z) = (x,\, y,\, 0)$*. Los coeficientes*

Figura 3.8: Ejemplo con una elevación real del terreno (izq.) y de tipo cosinusoidal (der.).

de la matriz S *se tomaron como* $\alpha_1 = \alpha_2 = 1$ *y* $\alpha_3 = 0.001$. *El sistema discreto resultante fue del orden de 15,000 con una malla muy gruesa. El error del campo ajustado obtenido numéricamente respecto del exacto* $\mathbf{u}(x, y, z) = (x, y, -2z)$ *fue del orden de 1 %. Una discretización con el mismo nivel de precisión que en el caso bidimensionales daría sistemas del orden de* $80^3 = 512,000$ *ecuaciones.*

Figura 3.9: Campo tridimensional recuperado. Porcentaje de error 0.95 %.

Bibliografía

[1] W.Gautschi, *Numerical Analysis: An Introduction*, Birkhauser, 1998.

[2] L. N. Trefethen, D. Bau III, *Numerical Linear Algebra*, SIAM, 1997.

[3] Cleve Moler, *Numerical Computing Whit Matlab*, Disponible en linea: http://www.mathworks.com/moler/index.html

[4] G. Golub, C. Van Loan, *Matrix Computation*, John Hopkins University Press, 2nd. Ed. 1989, 3rd. Ed. 1996.

[5] Alston S. Householder, *Unitary Triangularization of a Nonsymmetric Matrix*, Journal ACM, 5 (4), pp. 339–342, 1958.

[6] Jorge Nocedal and Stephen J. Wright, *Numerical Optimization*, second edition, Springer, 2006.

[7] R. Fletcher, *Practical Methods for Optimization*, 2nd. edition. Wiley, 1987.

[8] A. Cauchy, *Methodes generales pour la resolution des systémes d'equationes simultanees*, C.R. Acad. Sci. Par., 25:536–538, 1847.

[9] Magnus R. Hestenes y Eduard Stiefel, *Methods of Conjugate Gradients for Solving Linear Systems*, Journal of Research of the National Bureau of Standards, Vol. 49, No. 6, December **1952**, Research Paper 2379.

[10] *Top ten algorithms in the 20th century*: January/February 2000 issue of Computing in Science & Engineering (a joint publication of the American Institute of Physics and the IEEE Computer Society)

[11] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, 1996.

[12] Diana Assaely León Velasco, *Reguralización de problemas mal planteados*, tesis de maestría en matemáticas, Universidad Autónoma Metropolitana, Iztapalapa, 2010.

[13] Ciro F. Flores Rivera, *Modelación computacional de problemas en mecánica de fluidos*, tesis de doctorado en matemáticas, Universidad Autónoma Metropolitana, Iztapalapa, 2009.

[14] Ciro Flores, Héctor Juárez, Marco Nuñez, María Luisa Sandoval, *Algorithms for Vector Field Generation in Mass Consistent Models*, Numerical Methods for Partial Differential Equations, Vol. 26, No. 4, pp. 826–842, 2010.