NUMERO DE CONDICION Y DETERMINANTE DE UNA MATRIZ

Jorge Lemagne Pérez, Facultad de Matemática y Computación, Universidad de La Habana

RESUMEN
En la resolución computacional de sistemas de ecuaciones lineales Ax = b mediante métodos directos, juega un papel fundamental el llamado número de condición de A:

$$\mathrm{cond}(A) = ||A||\ ||A^{-1}||$$

ya que este determina la precisión alcanzable en la solución calculada. Con frecuencia se piensa que si la matriz A es mal acondicionada con respecto a la precisión utilizada, su determinante det(A) es cercano a 0 y viceversa. En algunos casos la tradición ha identificado ambas situaciones, sin embargo, aunque existe relación, no son equivalentes. En este trabajo se explica cómo influye el determinante en el número de condición, y recíprocamente, y a continuación se detallan sendos contraejemplos que refutan la falacia expresada anteriormente. Este problema tiene gran importancia dentro de la Matemática aplicada, ya que generalmente los cálculos se realizan dentro de una aritmética de punto flotante, con precisión finita, lo que puede en determinados casos, producir sorpresas desagradables en el resultado final.

ABSTRACT
In computational resolution of linear algebraic systems Ax = b by direct methods, the condition number of A plays an important role:

$$\mathrm{cond}(A) = ||A||\ ||A^{-1}||$$

since the attainable precision of the computer solution is determined by it. We might think that if A is ill conditioned with respect to the used precision, its determinant det(A) is close to 0, and conversely. Sometimes tradition has considered both situations as equivalent; however this is not true, although they are relationed. In this paper, the reciprocal influence between determinant and condition number is explained and examples to disprove this statement are given.

MSC: 65FO5

1. PRELIMINARES

Sea un sistema de ecuaciones lineales Ax = b, donde A = (a_{ij}) es una matriz cuadrada de orden n, de coeficientes reales e inversible y $b \in R^n$. Supongamos que este sistema se va a resolver mediante algún método directo, por ejemplo, el algoritmo de eliminación de Gauss. Es conocido ([2], [5], y [6]) que la precisión obtenible mediante tal método depende en gran medida del número de condición de A:

$$\mathrm{cond}(A) = ||A||\ ||A^{-1}|| \tag{1}$$

donde ||C|| indica alguna norma matricial de C. Además, sabemos que

$$1 \leq \mathrm{cond}(A) < \infty$$

El número de condición de A, según [1], puede interpretarse de la siguiente manera:

$$\frac{1}{\mathrm{cond}(A)} = \min\left\{\frac{||A-B||}{||A||} \,|\, B \text{ es no inversible}\right\} \tag{2}$$

es decir, que el recíproco del número de condición de A es igual a la distancia relativa de A a la matriz B no inversible más cercana.

Aquí se ha hablado de cercanía a una matriz B no inversible, lo cual nos pudiera hacer pensar que existe una relación directa entre det(A) (determinante de A) y cond(A); algunos autores hacen hincapié en esta relación. Recordemos (mediante [3]) que:

$$A^{-1} = \frac{1}{\det(A)} A^{r} \qquad (3)$$

donde A^T es la matriz adjunta de A, o sea A está formada por los cofactores de los elementos de A. Por lo tanto, de (1) se tiene que

$$\operatorname{cond}(A) = \frac{1}{|\det(A)|} \| A \| \| A^T \| \qquad (4)$$

Por supuesto que det(A) influye inversamente en la magnitud de cond(A), pero hay que tomar en cuenta también los otros 2 factores $\|A\|$ y $\|A^T\|$, cuyas magnitudes dependen no solo de los elementos de A en sí, sino también de las posiciones que ocupen; luego, el problema es en realidad más complejo.

2. DESARROLLO

Precisemos esto un poco más: podríamos formularnos las 2 preguntas siguientes:

1) ¿Si |det(A)| es significativamente mayor que 0 entonces cond(A) tiene que ser pequeña?

2) ¿Si det(A) es pequeño entonces cond(A) tiene que ser grande?

Mediante ejemplos demostraremos que *ambas preguntas se responden negativamente*.

1) |det(A)| puede ser significativamente mayor que 0 y cond(A) ser grande.

Tomemos

$$A = \begin{bmatrix} 10^7 & 1 \\ 10^{14} & 10^{21} \end{bmatrix}$$

En nuestro caso,

$\det(A) = 10^{28} - 10^{14} \approx 10^{28}$

Algunos autores calculan el "determinante normalizado" de A:

$$\det_{norm}(A) = \det(A') \qquad (5)$$

donde A' es el resultado de normalizar cada vector fila de A; por lo tanto, si A' = $A' = \left(a'_{ij}\right)_{n\times n}$, entonces

$$a'_{ij} = \frac{a_{ij}}{f_i} : f_i = \left\| \left(a_{ij}\right)_{j=1,\cdots,n} \right\|_{\infty} = \max_j | a_{ij} |$$

Aquí hemos tomado la $\| \ \|_{\infty}$ por facilidad, y es la que utilizaremos preferiblemente en este artículo.

Aplicando esta definición en (5) obtenemos que

$$\det_{norm}(A) = \frac{\det(A)}{f_1 f_2 \cdots f_n} \qquad (6)$$

Supongamos ahora que n = 2, y que

$$sg(a_{11}a_{22}) = sg(a_{21}a_{12}) \qquad (7)$$

o sea, que los productos de los elementos de ambas diagnonales tengan el mismo signo. En particular, se cumple (7) si todos los elementos de A tienen el mismo signo. Entonces:

$$|\det(A)| = |a_{11}a_{22} - a_{21}a_{12}| \leq \max\{|a_{11}a_{22}|, |a_{21}a_{12}|\} = \max\{|a_{11}|\,|a_{22}|, |a_{21}|\,|a_{12}|\} \leq f_1 f_2$$

Por lo que, si se cumple (7):

$$|\det(A)| \leq f_1 f_2 \qquad (8)$$

y en este caso a partir de (6) y de (8), se obtiene que:

$$|\det_{norm}(A)| = \frac{|\det(A)|}{f_1 f_2} \leq \frac{f_1 f_2}{f_1 f_2} = 1 \qquad (9)$$

Cuando $\det_{norm}(A)$ es mucho menor que 1, algunos autores afirman por definición que A es mal acondicionada.

Veamos qué sucede en nuestro ejemplo. Aplicando (6) se obtiene que

$$\det_{norm}(A) \approx \frac{10^{28}}{10^7 \cdot 10^{21}} = 1,$$

que de acuerdo con (9) es el mayor valor posible para n = 2. Por lo tanto, aún en la variante normalizada, el determinante es grande; entonces para cualquier variante A no debería presentar ningún problema de mal acondicionamiento.

Sin embargo, calculemos ahora cond(A). Primeramente, utilizaremos la $\|\ \|_\infty$, teniendo en cuenta que para matrices $C = (c_{ij})_{n\times n}$: (según [4]).

$$\|C\|_\infty = \max_i \sum_{j=1}^{n} |c_{ij}| \qquad (10)$$

entonces de (3):

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{bmatrix}^T = \frac{1}{\det(A)} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \qquad (11)$$

De (4), (10) y (11):

$$\text{cond}(A) = \max\{|a_{11}| + |a_{12}|, |a_{21}| + |a_{22}|\} \cdot \frac{1}{|\det(A)|} \max\{|a_{11}| + |a_{21}|, |a_{12}| + |a_{22}|\} \quad (12)$$

por lo que en nuestro ejemplo

$$\text{cond}(A) \approx 10^{21} \cdot \frac{1}{10^{28}} \cdot 10^{21} = 10^{14}$$

Más abajo analizaremos la significación de este número, pero primero recordemos lo siguiente:

Sea u la unidad de redondeo en la aritmética de punto flotante con que se realizan los cálculos. Si en dicha aritmética la base es β y la precisión es t, entonces ([1])

$u = ½\ \beta^{1-t}$ si el redondeo es simétrico

$u = \beta^{1-t}$ si el redondeo es truncado.

Por ejemplo, si en una computadora trabajamos con simple precisión, u es del orden de 10^{-7}, y si es en doble precisión u es del orden de 10^{-14}.

Decir que el número de condición de A es pequeño o grande es por supuesto relativo; de aquí que cobre mucha mayor importancia la definición de **matriz mal acondicionada con respecto a la precisión utilizada**, es decir, cuando cond(A) es del orden de $u^{-1.}$ De acuerdo con esto, la matriz A del ejemplo es mal acondicionada (trabajando con la $\|\ \|_\infty$), inclusive con doble precisión.

Vamos ahora a generalizar el resultado anterior para cualquier norma matricial; para ello utilizaremos el **radio espectral** de la matriz:

$$\rho(C) = \max_i \{|\lambda_i(C)|\} \quad (13)$$

es decir el máximo entre los módulos de los valores propios de C. Es sabido además que

$$\rho(C) \le \|C\|, \text{ para cualquier } \|\ \| \quad (14)$$

De (1) y (14) se obtiene que:

$$\rho(A)\ \rho(A^{-1}) \le \|A\|\ \|A^{-1}\| = \text{cond}(A) \quad (15)$$

Calculemos $\rho(A)\ \rho(A^{-1})$ para la A del ejemplo:

$$\det(A - \lambda I) = \det\begin{bmatrix} 10^7 - \lambda & 1 \\ 10^{14} & 10^{21} - \lambda \end{bmatrix} = \lambda^2 - (10^{21} + 10^7)\lambda + 10^{28} - 10^{14} = 0$$

El mayor $|\lambda_i|$ es aproximadamente igual a 10^{21}, por lo que

$\rho(A) \approx 10^{21}$ (16)

Por otra parte, con respecto a A^{-1}, sustituyendo en (11):

$$\det(A^{-1} - \lambda I) = \det\begin{bmatrix} 10^{-7} - \lambda & -10^{-28} \\ -10^{-14} & 10^{-21} - \lambda \end{bmatrix} = \lambda^2 - (10^{-7} + 10^{-21})\lambda + 10^{-28} - 10^{-42} = 0$$

El mayor $|\lambda_i|$ es aproximadamente igual a 10^{-7}, por lo que

$\rho(A^{-1}) \approx 10^{-7}$ (17)

Sustituyendo (16) y (17) en (15) se obtiene que

$\text{cond}(A) \geq 10^{21} \cdot 10^{-7} = 10^{14}$,

independientemente de la norma matricial utilizada.

Conclusión 1: *Para cualquier norma la matriz del ejemplo tiene un número de condición grande, inclusive si trabajamos con doble precisión, a pesar de que el determinante es grande.*

Pasemos ahora a responder la pregunta 2:

2) det(A) puede ser pequeño y cond(A) no ser grande.

En este caso trabajaremos con una aritmérica flotante de 14 dígitos decimales (doble precisión), y la matriz

$$A = \begin{bmatrix} 10^{-7} & 10^{-7} & 1 \\ 10^{-7} & 2\cdot 10^{-7} & 1 \\ 2\cdot 10^{-7} & 10^{-7} & 1 \end{bmatrix}$$

$\det(A) = \det_{norm}(A) \approx -10^{-14}$, (18)

cuyo módulo es un número pequeño, tanto desde el punto de vista absoluto como relativo a los datos, ya que es 10^7 veces inferior al menor de los coeficientes. Inclusive en nuestro ejemplo det(A) es el orden de la unidad de redondeo u de la aritmética flotante. De (10):

$||A||_\infty = 1.0000003$ (19)

Por otra parte

$$A^T = \begin{bmatrix} 10^{-7} & 0 & -10^{-7} \\ 10^{-7} & -10^{-7} & 0 \\ -3\cdot 10^{-14} & 10^{-14} & 10^{-14} \end{bmatrix}$$

Entonces, por (3), aplicando además (10) y (18):

$$\| A^{-1} \|_\infty = 10^{14} \cdot 2{\cdot}10^{-7} = 2{\cdot}10^{7}$$

Sustituyendo el resultado anterior y (19) en (1) se tiene finalmente que

$$\text{cond}(A) \approx 2{\cdot}10^{7}$$

De acuerdo con la definición, esta matriz *no* es mal acondicionada con respecto a la precisión utilizada, pues cond(A) es mucho menor que 10^{14}. Si nuestra precisión fuera de 7 cifras, entonces sí lo sería.

Conclusión 2: *El determinante de una matriz puede ser relativamente pequeño, pero esta no ser mal acondicionada con respecto a la precisión utilizada*.

Resumiendo en general, podemos decir que de acuerdo con (4), existe una relación entre número de condición y determinante de una matriz, pero que el hecho de que este último sea relativamente pequeño no implica que la matriz sea mal acondicionada con respecto a la precisión utilizada, y viceversa.

Faltaría analizar ahora qué problemas computacionales se presentan al resolver sistemas lineales mal acondicionados y sistemas con matrices cuyos determinantes sean pequeños. Esto podría ser objetivo de un trabajo futuro.

REFERENCIAS

[1] CONTE, S.D. and C. de BOOR: **Elementary Numerical Analysis, an algorithmic approach**, Third edition.

[2] FORSYTHE, G.; M. MALCOLM and C. MOLER (1972): **Computer Methods for Mathematical Computations**, Computer Science Department, Stanford University.

[3] HADLEY, G. (1968): **Linear algebra**, Ciencia y técnica.

[4] SUAREZ, M. (1980): **Matemática numérica**, Editorial de libros para la educación.

[5] VOLKOV, E.A. (1990): **Métodos numéricos**, Editorial MIR, Moscú.

[6] WILKINSON, J.H. and C. REINSCH (1971): **Linear Algebra**, Springer-Verlag.