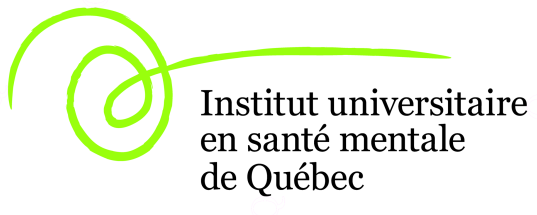
Chapitre 3 : Résolution de systèmes d'équations algébriques

Ibrahima Dione (Université Laval)

21 février 2017



1. André Fortin: *Analyse numérique pour ingénieurs*, Presses Internationales Polytechnique, Montréal, Québec (2016).

Systèmes linéaires

Un système d'équations linéaires est un ensemble d'équations dont les inconnues, notées $x_1, x_2, \cdots, x_n$, sont en relations linéaires à travers chacune de ces équations

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n = b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots + a_{3n}x_n = b_3 \\ \qquad\vdots \qquad\qquad\vdots \qquad\qquad\vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n = b_n \end{cases} \tag{1.1}$$

La notation matricielle du système (1.1) est plus pratique du fait de sa compacité, et s'écrit

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{\vec{x}} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}}_{\vec{b}} \tag{1.2}$$

Remarque(s) : La matrice A et le *membre de droite* $\vec{b}$ sont connus, mais $\vec{x}$ est inconnue.

- Les matrices sont non singulières ou inversibles, pour assurer l'existence d'une solution.
- La solution de l'équation (1.2) s'écrit $\vec{x} = A^{-1}\vec{b}$. Mais calculer A^{-1}, est plus difficile.

Question: Comment résoudre de manière automatique et à des difficultés moindre de tel systèmes ?

Exemple 1

Considérons le système linéaire suivant, sous sa forme matricielle et d'équations linéaires

$$\begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 11 \end{bmatrix} \longleftrightarrow \begin{aligned} 2x_1 + 3x_2 &= 8 \\ 3x_1 + 4x_2 &= 11 \end{aligned}$$

On utilise la méthode classique de résolution consistant à éliminer les équations une à une par *substitution successive*. En effet, en isolant x_1 de la première équation, on a

$$x_1 = \frac{8 - 3x_2}{2}. \tag{1.3}$$

Et si l'on substitue dans la deuxième, on obtient l'équation suivante

$$3\left(\frac{8 - 3x_2}{2}\right) + 4x_2 = 11,$$

dont la solution est $x_2 = 2$; d'où à partir de l'équation (1.3), on tire $x_1 = 1$.

Bien qu'il est théoriquement possible d'étendre la substitution successive à des systèmes de grande taille, il est cependant difficile de la transcrire sous forme d'algorithme afin d'être programmé dans un langage informatique.

Question: Existent-t-ils de types de systèmes linéaires faciles à résoudre ?

Définition 1.1

La matrice $A = (a_{ij})_{1 \leq i,j \leq n}$ est dite diagonale, si ses entrées sont nulles en dehors de sa diagonale (i.e. $a_{ij} = 0,\ i \neq j$). Une telle matrice est de la forme

$$A = \begin{bmatrix} a_{11} & 0 & \dots & \dots & 0 \\ 0 & a_{22} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & a_{nn} \end{bmatrix}$$

Exemple 2

Intéressons nous maintenant au système linéaire suivant, facile à résoudre :

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 9 \end{bmatrix}$$

Il suffit de considérer séparément chaque ligne et on obtient la solution $\vec{x} = [2 \quad 1 \quad 3]^t$.

Exemple 3

Le cas général suivant, où on suppose $a_{ii} \neq 0$, est résolu en considérant pour chaque ligne $i = 1, 2, \cdots, n$, $x_i = \frac{b_i}{a_{ii}}$.

$$\begin{bmatrix} a_{11} & 0 & \cdots & \cdots & 0 \\ 0 & a_{22} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ \vdots \\ \vdots \\ b_n \end{bmatrix}$$

Les systèmes diagonaux sont ainsi faciles à résoudre.

Définition 1.2

La matrice $A = (a_{ij})_{1 \leq i,j \leq n}$ est dite triangulaire inférieure (ou supérieure) si tous les a_{ij} (ou tous les a_{ji}) sont nuls pour $i < j$. La matrice triangulaire inférieure A, a la forme type

$$A = \begin{bmatrix} a_{11} & 0 & \cdots & \cdots & 0 \\ a_{21} & a_{22} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ a_{n1} & \cdots & \cdots & a_{n(n-1)} & a_{nn} \end{bmatrix}$$

Une matrice triangulaire supérieure est tout simplement la transposée d'une matrice triangulaire inférieure.

Exemple 4

Considérons encore l'exemple suivant, où la partie supérieure de la matrice est nulle

$$\begin{bmatrix} 3 & 0 & 0 \\ 1 & 2 & 0 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 7 \\ 14 \end{bmatrix}$$

1. Pour résoudre ce système, il suffit de faire une première descente pour obtenir $x_1 = \frac{9}{3} = 3$.
2. À la deuxième descente, on calcul avec l'aide de x_1 la valeur $x_2 = \frac{7-(1)(3)}{2} = 2$.
3. Et la dernière valeur est obtenue grâce à x_1 et x_2 par $x_3 = \frac{14-(3)(3)-(2)(2)}{1} = 1$.

On peut également généraliser, dans le cas d'un système de taille n, la résolution de ce système à matrice triangulaire inférieure. La solution est déterminée par les relations génériques suivantes

$$x_1 = \frac{b_1}{a_{11}}, \quad x_i = \frac{\left(b_i - \sum_{k=1}^{i-1} a_{ik} x_k\right)}{a_{ii}}, \ i = 2, 3, \cdots, n. \tag{1.4}$$

Dans le cas d'un système à matrice triangulaire supérieure, la solution est donnée

$$x_n = \frac{b_n}{a_{nn}}, \quad x_i = \frac{\left(b_i - \sum_{k=i+1}^{n} a_{ik} x_k\right)}{a_{ii}}, \ i = n-1, n-2, \cdots, 2, 1. \tag{1.5}$$

Note: Une autre réponse à la question précédente est que les systèmes triangulaires sont également faciles à résoudre. Il suffit en effet de commencer par l'équation qui se trouve à la pointe du triangle (la première pour une matrice triangulaire inférieure et la dernière pour une matrice triangulaire supérieure) et de résoudre une à une les équations. On parle de descente triangulaire ou de remontée triangulaire, selon le cas.

Remarque(s) :

1. Les équations (1.4) et (1.5) sont valides si les a_{ii} sont non nuls. Sinon la matrice A n'est pas inversible et, donc, le système $A\vec{x} = \vec{b}$ n'a pas une solution unique.
2. Le déterminant de la matrice triangulaire inférieure A (ou supérieure), est égale au produit des éléments de la diagonale :
$$det(A) = \prod_{i=1}^{n} a_{ii}$$
3. Pour résoudre un système linéaire quelconque, on se ramenera toujours à un système triangulaire plutôt qu'à un système diagonal car ce dernier est rarement rencontré en pratique et exige plus de travail pour s'y ramener.

Question: Comment ramener un système linéaire quelconque à un système triangulaire sans changer la solution du système de départ ?

Opérations élémentaires sur les lignes

La question à répondre ici est, comment transformer le système linéaire

$$A\vec{x} = \vec{b} \tag{1.6}$$

en un système triangulaire sans en modifier la solution. La réponse est de multiplier à gauche de chaque côté du système (1.6) par une matrice inversible W. C'est à dire, on s'intéressera ainsi au système triangulaire suivant

$$WA\vec{x} = W\vec{b} \tag{1.7}$$

Remarque(s) : Les systèmes (1.6) et (1.7) ont la même solution $\vec{x}$. Ce résultat n'est plus vrai si la matrice W n'est pas inversible, i.e. à partir du système (1.7) on ne pourra pas revenir au système (1.6) si la matrice W^{-1} n'existe pas.

Note: Pour transformer un système quelconque en un système triangulaire, il suffit d'utiliser trois opérations élémentaires sur les lignes de la matrice. Ces trois opérations élémentaires correspondent à trois différentes types de matrices W. En notant $\vec{l}_i$ la ligne i de la matrice A du système de départ, on a :

1. L'opération qui consiste à remplacer la ligne i par un multiple d'elle-même notée $(\vec{l}_i \longleftarrow \lambda \vec{l}_i)$;
2. L'opération qui consiste à intervertir la ligne i et la ligne j représentée par $(\vec{l}_i \longleftrightarrow \vec{l}_j)$;
3. L'opération qui est de remplacer la ligne i par la ligne i plus un multiple de la ligne j notée par $(\vec{l}_i \longleftarrow \vec{l}_i + \lambda \vec{l}_j)$.

Chacune de ces opérations élémentaires est équivalente à multiplier le système (1.6) par une matrice inversible.

Multiplication d'une ligne par un scalaire ($\vec{l}_i \longleftarrow \lambda\vec{l}_i$) et matrice équivalente

L'opération ($\vec{l}_i \longleftarrow \lambda\vec{l}_i$), consistant à remplacer la ligne l_i par un multiple d'elle-même, équivaut à multiplier le système linéaire (1.6) par une matrice diagonale inversible W dont tous les éléments diagonaux sont 1, sauf l'élément à la position (i, i), qui vaut λ.

Note: On obtient W à partir de la matrice identité en remplaçant l'élément à la position (i, i) par λ.

Remarque(s) :

1. Dans ce cas, le déterminant de la matrice diagonale W est λ. La matrice est donc inversible si $\lambda \neq 0$.
2. La matrice inverse de W, est celle diagonale dont tous les éléments diagonaux sont 1, sauf l'élément à la position (i, i), qui vaut $\frac{1}{\lambda}$.

Exemple 5

Soit le système suivant dont la solution est donnée par $\vec{x} = [1 \quad 1 \quad 1]^t$

$$\begin{bmatrix} 3 & 1 & 2 \\ 6 & 4 & 1 \\ 5 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 11 \\ 10 \end{bmatrix} \tag{1.8}$$

Si l'on souhaite effectuer l'opération $(\vec{l}_2 \longleftarrow 3\vec{l}_2)$ sur le système (1.8), c'est à dire multiplier la ligne 2 par un facteur 3, alors cela revient à multiplier ce système par la matrice

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Et l'on obtient ainsi le nouveau système suivant

$$\begin{bmatrix} 3 & 1 & 2 \\ 18 & 12 & 3 \\ 5 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 33 \\ 10 \end{bmatrix}$$

dont la solution reste la même que celle du système (1.8) de départ puisque la matrice W est inversible car $det(W) \neq 0$:

$$det(W) = 1 \times 3 \times 1 = 3.$$

Permutation de deux lignes ($\vec{l}_i \longleftrightarrow \vec{l}_j$) et matrice équivalente

L'opération ($\vec{l}_i \longleftrightarrow \vec{l}_j$) qui est à intervertir les lignes i et j, est équivalente à la multiplication du système (1.6) par une matrice inversible W, obtenue en permutant les lignes i et j de la matrice identité.

Note: On a W à partir de la matrice identité en permutant les lignes i et j.

Remarque(s) :

1. Le déterminant de la matrice W est -1. En permutant deux lignes, celui du système de départ change de signe.
2. L'inverse de la matrice W est elle même.

Exemple 6

Si l'on souhaite intervertir la ligne 2 et la ligne 3 du système (1.8) de l'exemple précédent,on a

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

On a le nouveau système dont la solution n'a pas changé (car la matrice W est inversible)

$$\begin{bmatrix} 3 & 1 & 2 \\ 5 & 4 & 1 \\ 6 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \\ 11 \end{bmatrix}$$

L'opération $\vec{l}_i \longleftarrow \vec{l}_i + \lambda\vec{l}_j$ et matrice équivalente

L'opération $\vec{l}_i \longleftarrow \vec{l}_i + \lambda\vec{l}_j$ est de remplacer la ligne i par elle même plus un multiple de la ligne j. Elle est équivalente à multiplier le système de départ par W qui vaut 1 sur toute la diagonale et 0 ailleurs, sauf à la position (i, j) qui est λ.

Note: On obtient la matrice W à partir de la matrice identité en ajoutant λ à la position (i, j).

Remarque(s) :

1. On peut montrer facilement que le déterminant de la matrice W est 1.
2. La matrice W est inversible et son inverse est obtenue en remplacant λ par $-\lambda$.

Exemple 7

Au système (1.8), effectuer l'opération $\vec{l}_2 \longleftarrow \vec{l}_2 - 2\vec{l}_1$ équivaut à multiplier ce système par la matrice inversible

$$W = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

On obtient le nouveau système

$$\begin{bmatrix} 3 & 1 & 2 \\ 0 & 2 & -3 \\ 5 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -1 \\ 10 \end{bmatrix}$$

Note: La stratégie de résolution qu'on va adopter, pour transformer un système quelconque en système triangulaire, sera donc basée sur ces trois opérations élémentaires qui seront effectuées à travers deux démarches appelées méthodes directes : la méthode d'élimination de Gauss et la décomposition LU.
En pratique, on ne multiplie jamais explicitement les systèmes considérés par les différentes matrices W, car ce serait trop long. Il faut cependant garder en tête que les opérations effectuées sont équivalentes à ces multiplications.

Élimination de Gauss

Méthode : La méthode de Gauss consiste à éliminer les termes sous la diagonale de la matrice A par le biais des opérations élémentaires, procurant un système triangulaire.

Note: La validité de la méthode de Gauss repose sur le fait que ces opérations équivalent à multiplier le système de départ par une matrice inversible.

Définition 2.1

La matrice augmentée du système linéaire (1.1) est la matrice de dimension n sur $n+1$ que l'on obtient en ajoutant le membre de droite $\vec{b}$ à la matrice A, c'est-à-dire :

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right]$$

Note: Les opérations élémentaires devant être effectuées à la fois sur les lignes de la matrice A et sur celles du vecteur $\vec{b}$, cette notation sera très utile.

Décomposition LU

Méthode : Si on réussit à exprimer la matrice A en un produit de deux matrices triangulaires L et U, alors le système $A\vec{x} = \vec{b}$ peut s'écrire comme suit

$$A\vec{x} = LU\vec{x} = \vec{b}.$$

En posant $U\vec{x} = \vec{y}$, la résolution du système linéaire $A\vec{x} = \vec{b}$ se fait alors en deux étapes

$$\begin{aligned} L\vec{y} &= \vec{b} \\ U\vec{x} &= \vec{y} \end{aligned} \tag{3.1}$$

qui sont deux systèmes triangulaires.

Remarque(s) :

1. On utilise d'abord une descente triangulaire sur la matrice L pour obtenir $\vec{y}$;
2. ensuite, une remontée triangulaire sur la matrice U pour obtenir la solution recherchée $\vec{x}$.

Note: Il est important de souligner que la décomposition LU n'est pas unique.

Décomposition de Crout

Méthode : Dans $A = LU$, si on fait le choix d'imposer des 1 sur la diagonale de U, on a la décomposition de Crout.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} & u_{14} \\ 0 & 1 & u_{23} & u_{24} \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

1. Première colonne de L, faire le produit des lignes de L par la première colonne de U

$$l_{11} = a_{11}, \quad l_{21} = a_{21}, \quad l_{31} = a_{31}, \quad l_{41} = a_{41}. \tag{3.3}$$

2. Première ligne de U, pourvu que $l_{11} \neq 0$,

$$u_{12} = a_{12}/l_{11}, \quad u_{13} = a_{13}/l_{11}, \quad u_{14} = a_{14}/l_{11}, \tag{3.4}$$

3. Produit des lignes de L par la deuxième colonne de U détermine la deuxième colonne de L

$$l_{22} = a_{22} - l_{21}u_{12}, \quad l_{32} = a_{32} - l_{31}u_{12}, \quad l_{42} = a_{42} - l_{41}u_{12}. \tag{3.5}$$

4. Le produit de la deuxième ligne de L par les colonnes de U fournit d'avoir la deuxième ligne de U

$$u_{23} = (a_{23} - l_{21}u_{13})/l_{22}, \quad u_{24} = (a_{24} - l_{21}u_{14})/l_{22}. \tag{3.6}$$

5. Par le produit des lignes de L par la troisième colonne de U, on a la troisième colonne de L

$$l_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23}, \quad l_{43} = a_{43} - l_{41}u_{13} - l_{42}u_{23}. \tag{3.7}$$

6. Le produit de la troisième ligne de L par la quatrième colonne de U, on obtient l'élément

$$u_{34} = (a_{34} - l_{31}u_{14} - l_{32}u_{24})/l_{33}. \tag{3.8}$$

7. Le produit de la quatrième ligne de L par la quatrième colonne de U, fournit dernier coefficient de L

$$l_{44} = a_{44} - l_{41}u_{14} - l_{42}u_{24} - l_{43}u_{34}. \tag{3.9}$$

Remarque(s) : La décomposition (3.2) ne fonctionne que si les l_{ii} sont tous non nuls. Ce n'est pas toujours le cas et il est possible qu'il faille permuter deux lignes pour éviter cette situation, tout comme pour l'élimination de Gauss. Le coefficient l_{ii} est encore appelé pivot.

Définition 3.1

La *notation compacte* de la décomposition LU, d'une matrice de dimension 4 sur 4, est la matrice de coefficients

$$\begin{bmatrix} l_{11} & u_{12} & u_{13} & u_{14} \\ l_{21} & l_{22} & u_{23} & u_{24} \\ l_{31} & l_{32} & l_{33} & u_{34} \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \tag{3.10}$$

Remarque(s) :

1. Les coefficients 1 sur la diagonale de la matrice U ne sont pas indiqués explicitement dans (3.10), mais doivent tout de même être pris en compte.
2. De façon plus rigoureuse, la notation compacte revient à mettre en mémoire la matrice
$$L + U - I$$
et à détruire la matrice A.

Décomposition LU et permutation de lignes

La décomposition LU exige que les pivots l_{ii} soient non nuls, au cas contraire, il faut faire une permutation de lignes. De plus, contrairement à la méthode d'élimination de Gauss, la décomposition LU n'utilise le terme de droite $\vec{b}$ qu'à la toute fin, au moment de la descente triangulaire $L\vec{y} = \vec{b}$. On doit ainsi garder la trace de ces permutations afin de les appliquer sur $\vec{b}$.

Remarque(s) : Dans une décomposition LU, la permutation de lignes s'effectue toujours après le calcul de chaque colonne de L. On place en position de pivot le plus grand terme en valeur absolue de cette colonne (sous le pivot actuel), pour des raisons de précision que nous verrons plus loin.

Théorème 3.1

On peut calculer le déterminant d'une matrice A à l'aide de la méthode de décomposition LU de Crout par

$$det(A) = (-1)^N \prod_{i=1}^{n} l_{ii} \tag{3.11}$$

où N est le nombre de fois où on a interverti deux lignes.

Factorisation de Choleski

Si la matrice A est symétrique, la matrice triangulaire supérieure U dans la résolution du système $A\vec{x} = \vec{b}$ par une décomposition LU, peut être remplacée par la matrice transposée de L pour donner la décomposition $A = LL^t$. C'est ce qu'on appelle la factorisation de Choleski.

Méthode : La décomposition de Choleski de la matrice A consiste à déterminer les coefficients l_{ij} de la matrice L, tel que $A = LL^t$. Si A est de dimension 4 sur 4, on a

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}}_{L^t}$$

Par le produit des matrices L et L^t et se servant des coefficients a_{ij}, on détermine les coefficients l_{ij} comme suit

1. Pour les coefficients de la première colonne de L, on a
$$l_{11} = \sqrt{a_{11}}, \quad l_{21} = \frac{a_{21}}{l_{11}}, \quad l_{31} = \frac{a_{31}}{l_{11}}.$$
2. Pour la deuxième colonne de L, on a
$$l_{22} = \sqrt{a_{22} - l_{21}^2}, \quad l_{32} = \frac{a_{32} - l_{31}l_{21}}{l_{22}}.$$
3. On a le dernier coefficient de L qui est
$$l_{33} = \sqrt{a_{33} - l_{31}^2 - l_{32}^2}.$$

Remarque(s) :

1. La factorisation de Choleski n'est pas unique. On a $A = LL^t = (-L)(-L)^t$ qui sont deux factorisations différentes. On peut s'assurer de l'unicité en imposant $l_{ii} > 0$, ce qui revient au choix naturel de prendre la valeur positive de la racine carrée lors du calcul de l_{ii}.
2. le déterminant de A est donné par

$$det(A) = det(L) \times det(L^t) = (det(L))^2 = \prod_{i=1}^{n} l_{ii}^2$$

3. La factorisation de Choleski n'est menée à terme qu'avec une matrice symétrique définie positive.

Définition 3.2

Une matrice symétrique A est dite définie positive, si on a la condition $A\vec{x} \cdot \vec{x} > 0, \forall\ \vec{x} \neq \vec{0}$.

Note: A est définie positive si le produit scalaire de $\vec{x}$ et $A\vec{x}$ est strictement positif.

Question: Peut-t-on caractériser les matrices symétriques et définies positives ?

Théorème 3.2

Les énoncés suivants sont équivalents :

1. A est une matrice symétrique et définie positive ;
2. Toutes les valeurs propres de A sont réelles et strictement positives ;
3. Le déterminant des sous-matrices principales de A est strictement positif ;
4. Il existe une factorisation de Choleski $A = LL^t$.

Remarque(s) :

1. Les critères du théorème sont difficiles à vérifier au préalable et il semble bien que la meilleure façon de s'assurer qu'une matrice est définie positive est d'y appliquer l'algorithme de factorisation de Choleski.
2. Dans le cas d'une matrice symétrique non définie positive, on doit recourir à la factorisation LU classique.

Proposition 3.1

Soit A une matrice vérifiant

1. A est symétrique et $a_{ii} > 0$,
2. A est à diagonale strictement dominante, c'est-à-dire que,

$$|a_{ii}| > \sum_{j=1}^{n} |a_{ij}|, \ i = 1, \cdots, n,$$

alors A admet une décomposition de Cholesky.

Note: La proposition ne dit rien dans le cas où la matrice ne vérifie pas les conditions ; elle est donc une condition suffisante mais pas nécessaire.

Exemple 8

Voir exemple 12 des notes de cours !

Effets de l'arithmétique flottante

Jusqu'ici, nous n'avons effectué les opérations qu'en utilisant l'arithmétique exacte. Il est donc important de voir si l'arithmétique flottante, utilisée par les ordinateurs, a une influence sur ces calculs. Nous allons voir que certaines matrices sont très sensibles aux effets de l'arithmétique flottante (appelées matrices mal conditionnées).

Exemple 9

Soit le système linéaire suivant, dont la solution est $\vec{x} = [4 \quad 3]^t$:

$$\begin{bmatrix} 1 & 2 \\ 1,1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 10 \\ 10,4 \end{bmatrix}$$

En remplaçant le terme $1,1$ de la matrice par $1,05$, la solution devient $\vec{x} = [8 \quad 1]^t$.

Remarque(s) : Cet exemple démontre qu'une petite modification sur un terme de la matrice peut entraîner une grande modification de la solution exacte. En pratique, l'arithmétique flottante provoque inévitablement de petites modifications de chaque terme de la matrice et de sa décomposition LU. Il est alors tout à fait possible que ces petites erreurs aient d'importantes répercussions sur la solution et, donc, que les résultats numériques soient très éloignés de la solution exacte

Remarque(s) : En arithmétique flottante à m chiffres dans la mantisse, on doit effectuer chaque opération arithmétique en représentant les opérandes en notation flottante et en arrondissant le résultat de l'opération au m^e chiffre de la mantisse (voir chapitre 1).

Exemple 10

Voir exemple 14 des notes de cours !

Exemple 11

Lire les exemples 3.19 et 3.20 du livre[a] qui illustrent comment la stratégie du pivot et la mise à l'échelle améliorent la précision de la solution du système, respectivement !

a. André Fortin : *Analyse numérique pour ingénieurs*, Presses Internationales Polytechnique, Montréal (2016).

Définition 4.1

La mise à l'échelle consiste à diviser chaque ligne du système linéaire par le plus grand terme (en valeur absolue) de la ligne correspondante de la matrice A. *On ne tient pas compte du terme de droite* $\vec{b}$ *pour déterminer le plus grand terme de chaque ligne.*

Conditionnement d'une matrice

La section précédente illustre clairement la sensibilité qu'ont certains systèmes linéaires aux erreurs dues à l'arithmétique flottante. Ici, nous allons essayer de déterminer cette sensibilité en mesurant l'écart entre la solution numérique et celle exacte.

Définition 5.1

Une norme vectorielle est une application de $\mathbb{R}^n$ dans $\mathbb{R}$ ($\mathbb{R}$ désigne l'ensemble des réels) qui associe à un vecteur $\vec{x}$, un scalaire noté $\|\vec{x}\|$ qui vérifie les trois propriétés suivantes :

1. La norme d'un vecteur est toujours strictement positive (sauf si $\vec{x} = \vec{0}$)
$$\|\vec{x}\| > 0.$$
2. Si α est un scalaire (où $|\alpha|$ est la valeur absolue de α), alors on a
$$\|\alpha\vec{x}\| = |\alpha|\|\vec{x}\|.$$
3. L'inégalité triangulaire est toujours vérifiée entre deux vecteurs $\vec{x}$ et $\vec{y}$ quelconques
$$\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$$

Note: Toute application vérifiant ces trois propriétés est une norme vectorielle dont les plus connues sont la norme euclidienne, la norme l_1 et la norme l_∞.

Définition 5.2

La norme euclidienne d'un vecteur $\vec{x}$ est notée $\|\vec{x}\|_2$ et est définie par

$$\|\vec{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

Tandis que les normes l_1 et l_∞ d'un vecteur $\vec{x}$, notées respectivement $\|\vec{x}\|_1$ et $\|\vec{x}\|_\infty$, sont définies par

$$\|\vec{x}\|_1 = \sum_{i=1}^{n} |x_i|, \quad \|\vec{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

Exemple 12

Soit le vecteur $\vec{x} = [1 \quad -3 \quad -8]^t$. Sa norme euclidienne, l_1 et l_∞ sont définies par

$$\|\vec{x}\|_1 = 1 + 3 + 8 = 12$$

$$\|\vec{x}\|_\infty = \max(1, 3, 8) = 8$$

$$\|\vec{x}\|_2 = \sqrt{1 + 9 + 64} = \sqrt{74}$$

Définition 5.3

Une norme matricielle est une application qui associe à une matrice A un scalaire noté $\|A\|$ vérifiant les propriétés :

1. La norme d'une matrice est toujours strictement positive, sauf si la matrice a toutes ses composantes nulles

$$\|A\| > 0, \text{ sauf si } A = 0 \tag{5.1}$$

2. Si α est un scalaire, alors

$$\|\alpha A\| = |\alpha|\|A\| \tag{5.2}$$

3. L'inégalité triangulaire est toujours vérifiée entre deux matrices A et B quelconques, c'est-à-dire

$$\|A + B\| \leq \|A\| + \|B\| \tag{5.3}$$

4. Une quatrième propriété est nécessaire pour les matrices

$$\|AB\| \leq \|A\|\|B\| \tag{5.4}$$

Théorème 5.1

Pour toute norme vectorielle $\|\cdot\|_*$, l'application $\|A\|$ définie pour toute matrice A par

$$\|A\| = \sup_{\|\vec{x}\|_*=1} \|A\vec{x}\|_*$$

est une norme matricielle (car vérifiant les quatre propriétés (5.1)-(5.4)).

Note: Les normes matricielles induites par les normes vectorielles $\|\cdot\|_1$, $\|\cdot\|_\infty$ et $\|\cdot\|_2$ sont définies par

$$\|A\|_1 = \sup_{\|\vec{x}\|_1=1} \|A\vec{x}\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|$$

$$\|A\|_\infty = \sup_{\|\vec{x}\|_\infty=1} \|A\vec{x}\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$$

$$\|A\|_2 = \sup_{\|\vec{x}\|_2=1} \|A\vec{x}\|_2 = \left(\rho(A^tA)\right)^{\frac{1}{2}}$$

où ρ désigne le rayon spectral, c'est-à-dire la plus grande valeur propre.

Remarque(s) : La norme suivante, dite de Frobenius, n'est induite par aucune norme vectorielle

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}$$

Exemple 13

Soit la matrice A suivante

$$A = \begin{bmatrix} 1 & -2 & 5 \\ -3 & 1 & -5 \\ 1 & -9 & 0 \end{bmatrix}$$

Les différentes normes matricielles précédentes de la matrice A prennent alors les valeurs suivantes

$$\|A\|_1 = \max(5, 12, 10) = 12$$

$$\|A\|_\infty = \max(8, 9, 10) = 10$$

$$\|A\|_F = \sqrt{1 + 4 + 25 + 9 + 1 + 25 + 1 + 81} = \sqrt{147}$$

Définition 5.4

Une norme vectorielle et une norme matricielle sont dites *compatibles* si la condition

$$\|A\vec{x}\| \leq \|A\|\|\vec{x}\| \tag{5.5}$$

est valide quels que soient la matrice A et le vecteur $\vec{x}$.

Remarque(s) :

- Toutes les normes matricielles induites sont compatibles avec leurs normes vectorielles

$$\|A\vec{x}\|_1 \leq \|A\|_1\|\vec{x}\|_1, \quad \|A\vec{x}\|_\infty \leq \|A\|_\infty\|\vec{x}\|_\infty, \quad \|A\vec{x}\|_2 \leq \|A\|_2\|\vec{x}\|_2$$

- La norme de Frobenius est compatible avec la norme euclidienne

$$\|A\vec{x}\|_2 \leq \|A\|_F\|\vec{x}\|_2$$

Note: La norme matricielle induite par la norme euclidienne est rarement utilisée sauf dans un cadre théorique. On préférera la norme l_1 ou la norme l_∞, plus faciles et demandant moins d'effort de calcul.

Exemple 14

Considérons de nouveau le vecteur $\vec{x} = [1 \quad -3 \quad -8]^t$ et la matrice A suivante

$$A = \begin{bmatrix} 1 & -2 & 5 \\ -3 & 1 & -5 \\ 1 & -9 & 0 \end{bmatrix}$$

Ainsi le produit $A\vec{x}$ donne le vecteur $[-33 \quad 34 \quad 28]$ et donc

$$\|A\vec{x}\|_1 = 95 \quad \|A\vec{x}\|_\infty = 34 \quad \|A\vec{x}\|_2 = \sqrt{3029}$$

L'inégalité (5.5) devient, respectivement aux calculs précédents, ce-ci

$$95 \leq (12)(12) \qquad \text{en norme } l_1$$

$$34 \leq (10)(8) \qquad \text{en norme } l_\infty$$

$$\sqrt{3029} \leq (\sqrt{147})(\sqrt{74}) \qquad \text{en norme euclidienne}$$

Définition 5.5

Le conditionnement d'une matrice (noté *cond* A) est défini par le produit des normes de A et de son inverse

$$cond\ A = \|A\|\|A^{-1}\|$$

Note: Le conditionnement dépend de la norme matricielle utilisée et la norme $\|\cdot\|_\infty$ sera souvent utilisée. Quelle que soit la norme matricielle utilisée, le conditionnement d'une matrice A est un nombre supérieur ou égal à 1 :

$$1 \leq cond\ A < \infty$$

Si l'on revient au système linéaire $A\vec{x} = \vec{b}$ dont la solution exacte est $\vec{x}$, on sait que l'effet de l'arithmétique flottante est qu'une solution approximative $\vec{x}^*$ est obtenue. Il est donc souhaitable (même si ce n'est pas toujours le cas) que ces deux vecteurs soient près l'un de l'autre, c'est-à-dire que la norme de l'erreur $\vec{e} = \vec{x} - \vec{x}^*$ soit petite.

Note: Le vecteur $\vec{r} = \vec{b} - A\vec{x}^* = A\vec{x} - A\vec{x}^* = A(\vec{x} - \vec{x}^*) = A\vec{e}$ est appelé résidu associé à la solution approximative $\vec{x}^*$.

Théorème 5.2

Soit A une matrice de dimension n sur n et $\vec{b} \in \mathbb{R}^n$. Si le vecteur $\vec{x} \in \mathbb{R}^n$ est la solution du système linéaire $A\vec{x} = \vec{b}$ et $\vec{x}^* \in \mathbb{R}^n$ une solution approximative, alors on a

$$\frac{1}{cond\ A} \frac{\|\vec{r}\|}{\|\vec{b}\|} \leq \frac{\|\vec{e}\|}{\|\vec{x}\|} \leq cond\ A \frac{\|\vec{r}\|}{\|\vec{b}\|} \tag{5.6}$$

Remarque(s) : Plusieurs remarques s'imposent pour bien comprendre l'inégalité (5.6)

1. Le terme $\frac{\|\vec{e}\|}{\|\vec{x}\|}$ est l'erreur relative entre les solutions exacte $\vec{x}$ et approximative $\vec{x}^*$.
2. Si *cond A* est près de 1, $\frac{\|\vec{e}\|}{\|\vec{x}\|}$ est coincé entre deux valeurs très près l'une de l'autre. Si la norme du résidu est petite, l'erreur relative est également petite et la précision de la solution approximative a toutes les chances d'être satisfaisante.
3. Par contre, si le conditionnement de la matrice A est grand, la valeur de l'erreur relative est quelque part entre 0 et un nombre possiblement très grand. *Il est donc à craindre que l'erreur relative soit alors grande, donc que la solution approximative soit de faible précision et même, dans certains cas, complètement fausse*.
4. *Même si la norme du résidu est petite, il est possible que l'erreur relative liée à la solution approximative soit quand même très grande*.
5. Plus *cond A* est grand, plus on doit être attentif à l'algorithme de résolution utilisé.
6. Un mauvais algorithme peut mener à un résultat erroné $\hat{m}$ si A est bien conditionnée.

Note: Le calcul de l'inverse d'une matrice étant numériquement coûteux, il n'est pas utile de calculer le conditionnement. On peut cependent déterminer une borne inférieure facilement computable de celui-ci à partir de (5.6)

$$\begin{aligned} cond\ A &\geq \max\left\{ \frac{\|\vec{e}\|\|\vec{b}\|}{\|\vec{x}\|\|\vec{r}\|}, \quad \frac{\|\vec{x}\|\|\vec{r}\|}{\|\vec{e}\|\|\vec{b}\|} \right\} \\ &= \max\left\{ \frac{\|\vec{x}-\vec{x}^*\|}{\|\vec{x}\|}\frac{\|\vec{b}\|}{\|\vec{b}-A\vec{x}^*\|}, \quad \frac{\|\vec{x}\|}{\|\vec{x}-\vec{x}^*\|}\frac{\|\vec{b}-A\vec{x}^*\|}{\|\vec{b}\|} \right\} \end{aligned}$$

Théorème 5.3

Soit A une matrice de dimension n sur n et $\vec{b} \in \mathbb{R}^n$. Si le vecteur $\vec{x} \in \mathbb{R}^n$ est la solution du système linéaire $A\vec{x} = \vec{b}$ et $\vec{x}^* \in \mathbb{R}^n$ une solution approximative, alors on a

$$\frac{\|\vec{x} - \vec{x}^*\|}{\|\vec{x}^*\|} \leq cond\ A \frac{\|E\|}{\|A\|} \tag{5.7}$$

où E est la perturbation du système $A\vec{x} = \vec{b}$ qui devient $(A + E)\vec{x}^* = \vec{b}$.

Remarque(s) : Les remarques suivantes permettent de bien mesurer la portée de (5.7).

1. Le terme de gauche est une approximation de l'erreur relative entre la solution exacte et la solution du système perturbé. (On devrait avoir $\|\vec{x}\|$ au dénominateur pour représenter vraiment l'erreur relative).
2. Le terme de droite est en quelque sorte l'erreur relative liée aux coefficients de la matrice A multipliée par le conditionnement de A.
3. Si *cond* A est petit, une petite perturbation sur la matrice A entraîne une petite perturbation sur la solution $\vec{x}$.
4. Par contre, si *cond* A est grand, une petite perturbation sur la matrice A pourrait résulter en une très grande perturbation sur la solution du système. Il est par conséquent possible que les résultats numériques soient peu précis et même, dans certains cas, complètement faux.

Note: La matrice E étant la perturbation de A, par définition de la précision machine ε_m et de la norme l_∞, on a

$$\|E\|_\infty \leq \varepsilon_m \|A\|_\infty$$

On peut ainsi réécrire la relation (5.7) sous la forme suivante

$$\frac{\|\vec{x} - \vec{x}^*\|_\infty}{\|\vec{x}^*\|_\infty} \leq \varepsilon_m cond\ A = \varepsilon_m \|A\|_\infty \|A^{-1}\|_\infty$$

Plus le conditionnement est élevé, plus la précision machine ε_m doit être petite (si la simple précision est insuffisante, on recourt à la double précision).

Systèmes non linéaires

Les méthodes de résolution des systèmes non linéaires sont nombreuses (voir chapitre 2). Nous ne présentons ici que la méthode la plus utilisée en pratique, soit *la méthode de Newton*.
L'application de cette méthode à un système de deux équations non linéaires est suffisante pour illustrer le cas général.
Le problème est de trouver le vecteur $\vec{x} = [x_1 \;\; x_2]^t$, vérifiant les 2 équations non linéaires

$$\begin{cases} f_1(x_1, x_2) = 0 \\ f_2(x_1, x_2) = 0 \end{cases} \tag{6.1}$$

où f_1 et f_2 sont deux fonctions de deux variables que nous supposons différentiables.
En révisant le développement de la méthode de Newton pour une équation non linéaire (voir le chapitre 2), on considère une approximation initiale (x_1^0, x_2^0) de la solution du système (6.1) qu'on cherchera à corriger.

Note: Cette approximation initiale est cruciale et doit toujours être choisie avec soin.

Méthode :

- Par un développement de Taylor en deux variables pour chacune des équations du système (6.1) (où les termes d'ordre supérieur sont négligés), on détermine une correction $\delta\vec{x} = (\delta x_1, \delta x_2)$ à (x_1^0, x_2^0) en résolvant le système

$$\begin{aligned} \frac{\partial f_1}{\partial x_1}(x_1^0, x_2^0)\delta x_1 + \frac{\partial f_1}{\partial x_2}(x_1^0, x_2^0)\delta x_2 &= -f_1(x_1^0, x_2^0) \\ \frac{\partial f_2}{\partial x_1}(x_1^0, x_2^0)\delta x_1 + \frac{\partial f_2}{\partial x_2}(x_1^0, x_2^0)\delta x_2 &= -f_2(x_1^0, x_2^0) \end{aligned} \tag{6.2}$$

  où sous la forme matricielle, on écrira le système comme suit

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x_1^0, x_2^0) & \frac{\partial f_1}{\partial x_2}(x_1^0, x_2^0) \\ \frac{\partial f_2}{\partial x_1}(x_1^0, x_2^0) & \frac{\partial f_2}{\partial x_2}(x_1^0, x_2^0) \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} f_1(x_1^0, x_2^0) \\ f_2(x_1^0, x_2^0) \end{bmatrix} \tag{6.3}$$

- On construit ainsi une nouvelle approximation de la solution du système non linéaire (6.1) en posant

$$\begin{aligned} x_1^1 &= x_1^0 + \delta x_1 \\ x_2^1 &= x_2^0 + \delta x_2 \end{aligned}$$

- On cherchera par la suite à corriger (x_1^1, x_2^1) d'une nouvelle quantité $\delta\vec{x}$ en réitérant le procédé ci-dessus

Remarque(s) : Le système linéaire (6.3) s'écrit également sous une forme plus compacte

$$J(x_1^0, x_2^0)\vec{\delta x} = -\vec{R}(x_1^0, x_2^0) \tag{6.4}$$

où $J(x_1^0, x_2^0)$ désigne la matrice des dérivées partielles ou *matrice jacobienne* évaluée au point (x_1^0, x_2^0), $\vec{\delta x}$ est le vecteur des corrections relatives à chaque variable et où $-\vec{R}(x_1^0, x_2^0)$ est *le vecteur résidu* évalué en (x_1^0, x_2^0). Le déterminant de la matrice jacobienne est appelé *le jacobien*. Le jacobien doit bien entendu être différent de 0 pour que la matrice jacobienne soit inversible.

Exemple 15

Voir exemple 20 des notes de cours !

Exercices suggérés du manuel !

- Exercices[a] suggérés : 1-8, 10a), 11-14, 16, 21 26-28, 30, 31, 34, 36, 37.
- Exercices fortement suggérés : 2, 8, 10a), 16.

a. André Fortin : *Analyse numérique pour ingénieurs*, Presses Internationales Polytechnique

À Retenir !

Je dois pouvoir répondre aux questions[a] suivantes :

1. Je sais faire une méthode d'élimination de Gauss d'un système linéaire et définir les matrices équivalentes aux opérations élémentaires.
2. Je sais faire une décomposition et une résolution par LU et Cholesky.
3. Je distingue Cholesky et Crout.
4. Je sais reconnaitre des matrices permettant d'appliquer ou d'exclure Cholesky.
5. Je comprends la notion de conditionnement d'une matrice.
6. Je connais le théorème sur le conditionnement.
7. Je sais appliquer le théorème de conditionnement pour caractériser une solution et son erreur relative.
8. Je sais construire une borne inférieure du conditionnement et je comprends ses limitations.
9. Je sais calculer une matrice Jacobienne.
10. Je sais utiliser la méthode de Newton.

a. André Fortin : *Analyse numérique pour ingénieurs*, Presses Internationales Polytechnique 2016.