Chapitre IV

Systèmes d'Equations Linéaires

"It's **all** linear algebra." (fréquente proclamation de Gene Golub)

Considérons un système d'équations linéaires (a_{ij}, b_j donnés)

$$\begin{array}{c} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \vdots \qquad \vdots \qquad\qquad \vdots \qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{array} \tag{0.1}$$

et cherchons sa solution $x_1, \ldots, x_n$. Très souvent, il est commode d'utiliser la notation matricielle

$$Ax = b. \tag{0.2}$$

La méthode de Wanner. Prenons l'exemple

$$\begin{array}{ll} x_1 + 2x_2 - 3x_3 = 5 & (1) \\ 3x_1 + x_2 - 2x_3 = 1 & (2) \\ 2x_1 - x_2 + x_3 = 2 & (3) \end{array}$$

On soustrait (2) de (1), puis (3) de (2), et (1) de (3). Cela donne

$$\begin{array}{ll} -2x_1 + x_2 - x_3 = 4 & (4) \\ x_1 + 2x_2 - 3x_3 = -1 & (5) \\ x_1 - 3x_2 + 4x_3 = -3 & (6) \end{array}$$

Pour éliminer x_1, on soustrait (5) et (6) pour trouver

$$5x_2 - 7x_3 = 2 \qquad \Rightarrow x_2 = 6, \;\; x_3 = 4 \quad \text{et par (4)} \;\; x_1 = -1.$$

Tous les trois équations (4), (5) et (6) sont satisfaits. Mais il y a une mauvaise surprise si on essaie ces "solutions" dans (1), (2) ou (3). Nous comprenons que des éliminations à l'aveugle ne sont pas sans dangers !...

La règle de Cramer. (Une autre invention genevoise de 1750, cette fois sérieuse...) On apprend en algèbre linéaire que, si $\det(a_{ij}) \neq 0$, les solutions sont données par

$$x_k = \frac{\det(a_{ij}...b_i...a_{ij})}{\det(a_{ij})}$$

(où la k-ème colonne des a_{ij} est remplacée par les b_i) et pour $\det(a_{ij})$ on connait une formule avec une somme sur le groupe de permutation des indices $1, 2, ..., n$, i.e., avec $n!$ de termes (ce qui

donne la même chose que de développer récursivement les déterminants par une ligne). Or si, par exemple, $n = 20$, nous avons $n! = 2.4 \cdot 10^{18}$, et même sur l'ordinateur le plus rapide du monde (disons 10^9 opérations par seconde), ça prendrait $2 \cdot 10^9$ secondes $= 3 \cdot 10^7$ minutes $= 5 \cdot 10^5$ heures $= 3 \cdot 10^3$ jours $= 10$ années de calcul.

Bibliographie sur ce chapitre

Å. Björck (1996): *Numerical Methods for Least Squares Problems.* SIAM. [MA 65/387]

P.G. Ciarlet (1982): *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson.

J.J. Dongarra, C.B. Moler, J.R. Bunch & G.W. Stewart (1979): LINPACK *Users' Guide*. SIAM.

D.K. Faddeev & V.N. Faddeeva (1963): *Computational Methods of Linear Algebra.* Freeman & Co. [MA 65/271]

G.H. Golub & C.F. Van Loan (1989): *Matrix Computations.* Second edition. John Hopkins Univ. Press. [MA 65/214]

N.J. Higham (1996): *Accuracy and Stability of Numerical Algorithms.* SIAM. [MA 65/379]

A.S. Householder (1964): *The Theory of Matrices in Numerical Analysis.* Blaisdell Publ. Comp. [MA 65/262]

G.W. Stewart (1973): *Introduction to Matrix Computations*. Academic Press.

L.N. Trefethen & D. Bau (1997): *Numerical Linear Algebra.* SIAM. [MA 65/388]

J.H. Wilkinson (1969): *Rundungsfehler*. Springer-Verlag.

J.H. Wilkinson & C. Reinsch (1971): *Handbook for Automatic Computation, Volume II, Linear Algebra*. Springer-Verlag.

IV.1 Elimination de Gauss

L'élimination dite "de Gauss" a été pratiquée pendant des siècles sans grand tam-tam, notamment par Newton (voir chapitre II.1) et par Lagrange (en 1781 dans ses calculs astronomiques ; *Oeuvres* V, p. 125-490). Toutefois, Gauss ayant le souci de *prouver* l'existence des solutions pour son *principium nostrum* des moindres carrés (voir notices historiques du cours d'*Algèbre Linéaire*, p. 17) décrit l'algorithme explicitement :

Si $a_{11} \neq 0$, on peut éliminer la variable x_1 dans les équations 2 à n à l'aide de l'équation 1, c.-à-d., on calcule

$$\ell_{i1} = \frac{a_{i1}}{a_{11}} \qquad \text{pour} \qquad i = 2, \ldots, n \tag{1.1}$$

et on remplace la ligne i par

$$\text{ligne } i \;\; - \;\; \ell_{i1} \;*\; \text{ligne } 1.$$

De cette manière, on obtient le système équivalent

$$\begin{aligned} a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + \ldots + a_{1n}^{(1)} x_n &= b_1^{(1)} \\ a_{22}^{(1)} x_2 + \ldots + a_{2n}^{(1)} x_n &= b_2^{(1)} \\ \vdots \qquad\qquad \vdots \quad & \quad \vdots \\ a_{n2}^{(1)} x_2 + \ldots + a_{nn}^{(1)} x_n &= b_n^{(1)} \end{aligned} \tag{1.2}$$

où

$$\begin{aligned} a_{1j}^{(1)} &= a_{1j}, & b_1^{(1)} &= b_1, \\ a_{ij}^{(1)} &= a_{ij} - \ell_{i1} a_{1j} & b_i^{(1)} &= b_i - \ell_{i1} b_1 \end{aligned} \qquad \text{pour} \quad i = 2, \ldots, n \tag{1.3}$$

Le système (1.2) contient un sous-système de dimension $n-1$ sur lequel on peut répéter la procédure pour éliminer x_2 dans les équations 3 à n. On multiplie la ligne 2 de (1.2) par $\ell_{i2} = a_{i2}^{(1)}/a_{22}^{(1)}$ et on la soustrait de la ligne i. Après $n-1$ étapes

$$(A,b) \;\to\; (A^{(1)},b^{(1)}) \;\to\; (A^{(2)},b^{(2)}) \;\to\; \ldots \;\to\; (A^{(n-1)},b^{(n-1)}) =: (R,c)$$

on obtient un système triangulaire

$$\begin{aligned} r_{11}x_1 + r_{12}x_2 + \ldots + r_{1n}x_n &= c_1 \\ r_{22}x_2 + \ldots + r_{2n}x_n &= c_2 \\ \ddots \quad \vdots \quad & \vdots \\ r_{nn}x_n &= c_n \end{aligned} \tag{1.4}$$

qui se résoud facilement par "back substitution"

$$x_n = c_n/r_{nn}, \qquad x_i = (c_i - \sum_{j=i+1}^{n} r_{ij}x_j)/r_{ii} \qquad \text{pour} \quad i = n-1,\ldots,1. \tag{1.5}$$

Astuce pour la programmation : après l'élimination, les places de mémoire pour les $a_{21}, a_{31}, \ldots$ ne seront plus nécessaires (on sait que ces grandeurs sont nulles) ; on peut donc y stocker les $\ell_{21}, \ell_{31}, \ldots$ et tout l'algorithme peut se programmer en quelques lignes :

```
do ir=1,n-1
 do i=ir+1,n
  a(i,ir)=a(i,ir)/a(ir,ir)
  do j=ir+1,n
   a(i,j)=a(i,j)-a(i,ir)*a(ir,j)
  end do
  b(i)=b(i)-a(i,ir)*b(ir)
 end do
end do
```

```
C ---- BACK SUBSTITUTION ---
   x(n)=b(n)/a(n,n)
   do i=n-1,1,-1
    sum=0.
    do j=i+1,n
     sum=sum+a(i,j)*x(j)
    end do
    x(i)=(b(i)-sum)/a(i,i)
   end do
```

Théorème 1.1 *L'élimination de Gauss équivaut à une factorisation*

$$A = LR \tag{1.6}$$

où

$$L = \begin{pmatrix} 1 & & & \\ \ell_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ \ell_{n1} & \ldots & \ell_{n,n-1} & 1 \end{pmatrix}, \qquad R = \begin{pmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ & r_{22} & \ldots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{pmatrix}. \tag{1.7}$$

La formule (1.6) s'appelle "décomposition LR" (left - right) de la matrice A.

Démonstration. En utilisant les matrices

$$L_1 = \begin{pmatrix} 1 & & & & \\ -\ell_{21} & 1 & & & \\ -\ell_{31} & 0 & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ -\ell_{n1} & 0 & \ldots & 0 & 1 \end{pmatrix}, \qquad L_2 = \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & -\ell_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & -\ell_{n2} & \ldots & 0 & 1 \end{pmatrix}, \quad \ldots \tag{1.8}$$

le premier pas de l'élimination de Gauss correspond à une multiplication de A avec L_1, le deuxième avec L_2, etc.,

$$L_1 A = A^{(1)}, \quad L_2 A^{(1)} = A^{(2)}, \quad \ldots \quad , \quad L_{n-1} A^{(n-2)} = A^{(n-1)} = R.$$

Par conséquent,

$$R = (L_{n-1} L_{n-2} \cdot \ldots \cdot L_1) \cdot A \qquad \text{et} \qquad A = (L_{n-1} L_{n-2} \cdot \ldots \cdot L_1)^{-1} \cdot R.$$

Il reste à montrer que la matrice L de (1.7) est égale à $(L_{n-1} L_{n-2} \cdot \ldots \cdot L_1)^{-1}$. Pour ceci, nous appliquons la même procédure à la matrice L. La multiplication de L avec L_1 élimine les éléments de la première colonne en-dessous de la diagonale, puis la multiplication avec L_2 élimine ceux de la deuxième colonne, etc. Finalement, on obtient $(L_{n-1} L_{n-2} \cdot \ldots \cdot L_1) \cdot L = I$ =*identité*, ce qu'il fallait démontrer. □

Calcul du déterminant d'une matrice. La formule (1.6) implique que $\det A = \det L \cdot \det R$. On obtient

$$\det A = r_{11} \cdot \ldots \cdot r_{nn} \tag{1.9}$$

i.e., *le déterminant est le produit des pivots.*

Résolution de systèmes linéaires. En pratique, on rencontre souvent la situation où il faut résoudre une suite de systèmes linéaires $Ax = b$, $Ax' = b'$, $Ax'' = b''$, etc., possédant tous la même matrice. Très souvent, on connaît b' seulement après la résolution du premier système.

C'est la raison pour laquelle on écrit, en général, le programme pour l'élimination de Gauss en deux sous-programmes :

DEC – calculer la décomposition LR (voir (1.6)) de la matrice;

SOL – résoudre le système $Ax = b$. D'abord on calcule le vecteur c (voir (1.4)), défini par $Lc = b$, puis on résoud le système triangulaire $Rx = c$.

Pour le problème ci-dessus, on appelle *une fois* le sous-programme DEC et puis, pour chaque système linéaire, le sous-programme SOL.

Calcul de l'inverse d'une matrice. Si on choisit pour les b, b', b'' ci-dessus les vecteurs de base $(1, 0, ..., 0)^T$, $(0, 1, ..., 0)^T$, etc., on obtient pour les x, x', x'', etc., les *colonnes* de la matrice inverse A^{-1}.

Coût de l'élimination de Gauss. Pour le passage de A à $A^{(1)}$, on a besoin de

$n - 1$ divisions (voir (1.1)) et de

$(n-1)^2$ multiplications et additions (voir (1.3)).

Le calcul de $A^{(2)}$ nécessite $n - 2$ divisions et $(n-2)^2$ multiplications et additions, etc. Comme le travail dû aux divisions est ici négligeable, le coût total de la décomposition LR s'élève à environ

$$(n-1)^2 + (n-2)^2 + \ldots + 2^2 + 1^2 \approx \int_0^n x^2 \, dx = \frac{n^3}{3} \qquad \text{opérations}$$

(opération = multiplication + addition).

En revenant à l'exemple ci-dessus d'une matrice de dimension 20×20, l'algorithme de Gauss nécessite ≈ 2600 opérations, d'un facteur 10^{-15} fois plus petit que le travail des déterminants.

Exemple numérique. Prenons une matrice 60×60 avec coefficients choisies aléatoirement entre $-1 \le a_{ij} \le 1$ et calculons A^{-1} par la méthode de Gauss (en simple précision). Puis, on contrôle en double précision l'*erreur* des éléments. Le résultat est présenté en Fig. IV.1 à gauche (noir = 2 décimales justes, blanc = 7 décimales justes). Le résultat semble à désirer!! Mais il donne lieu à une nouvelle découverte: *The Scottish Kilt Phenomenon!!*

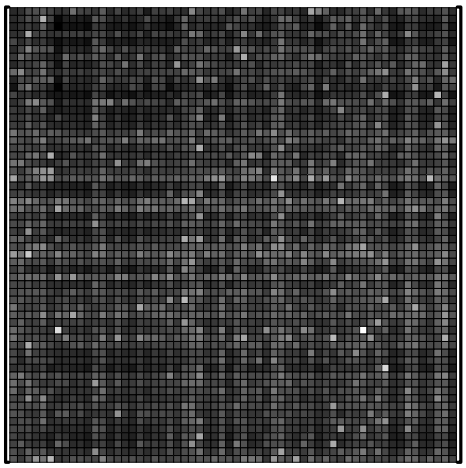
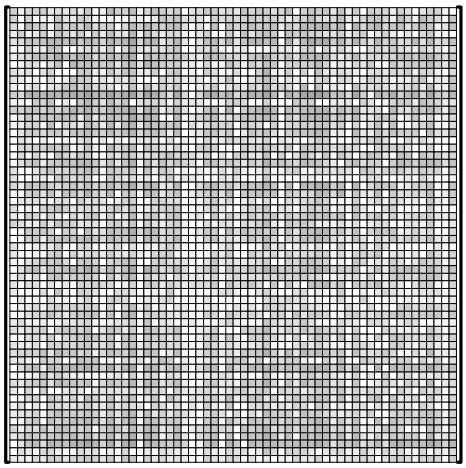
$A^{-1} =$  $A^{-1} =$ 

FIG. IV.1: Erreurs des éléments de A^{-1} d'une matrice aléatoire 60×60; *sans* recherche de pivot (gauche), *avec* recherche de pivot (droite) ; noir $= 2$ décimales justes, blanc $= 7$ décimales justes

Algorithme avec recherche de pivot.

Exemple 1.2 (Forsythe) Considérons le système

$$\begin{aligned} 1 \cdot 10^{-4} \cdot x_1 \ + \ 1 \cdot x_2 \ &= \ 1.0001 \\ 1 \cdot x_1 \ + \ 1 \cdot x_2 \ &= \ 2 \qquad . \end{aligned} \tag{1.10}$$

La solution exacte est, comme on voit, $x_1 = 1$ et $x_2 = 1$. Appliquons l'élimination de Gauss et simulons un calcul en virgule flottante avec 3 chiffres significatifs (en base 10).

Si l'on prend $a_{11} = 1 \cdot 10^{-4}$ comme pivot, on obtient $\ell_{21} = a_{21}/a_{11} = 1.00 \cdot 10^4$, $a_{22}^{(1)} = 1.00 - 1.00 \cdot 10^4 = -1.00 \cdot 10^4$. L'information contenue dans la valeur de $a_{22} = 1$ a tout simplément disparue. Il est clair que le reste du calcul est faux. Regardons: $b_2^{(1)} = 2.00 - 1.00 \cdot 10^4 = -1.00 \cdot 10^4$. Par conséquent, $x_2 = b_2^{(1)}/a_{22}^{(1)} = 1.00$ (exacte!, la première équation n'a pas été endommagé), mais pour x_1 nous obtenons

$$x_1 = (b_1 - a_{12}x_2)/a_{11} = (1.00 - 1.00 * 1.00)/(1.00 \cdot 10^{-4}) = 0.$$

Le résultat numérique, obtenu pour x_1, est faux.

Nous voyons à cet exemple, qu'il faut éviter qu'un des a_{rr} deviendrait trop petit. L'idée est alors de ramener un des $a_{ij} \neq 0$ à la place du a_{rr} par des échanges de lignes (i.e., échange des équations) et/ou des échanges de colonnes (i.e., échange des x_i), pour le rendre le plus grand possible. L'algorithme le plus souvent utilisé dans les codes est le suivant :

Recherche partielle de pivot. On ne se contente pas d'un pivot différent de zéro ($a_{rr} \neq 0$), mais on échange les équations de (0.1) afin que a_{rr} soit le plus grand élément (en valeur absolue) des a_{ir}, $(i = r, r+1, \ldots, n)$. De cette manière on a toujours $|\ell_{ir}| \leq 1$. Pour la programmation, il suffit

d'insérer dans le code ci-dessus, après la première ligne, les commandes :

```
c --- recherche du pivot ----
   pgval=0.
   izero=ir
   do i=ir,n
    valabs=abs(a(i,ir))
    if(valabs.gt.pgval)then
     pgval=valabs
     izero=i
    end if
   end do
```

```
c --- echange -----
   do j=ir,n
    store=a(ir,j)
    a(ir,j)=a(izero,j)
    a(izero,j)=store
   end do
   store=b(ir)
   b(ir)=a(izero)
   b(izero)=store
```

IV.2 Etude des erreurs ; "Backward Error Analysis"

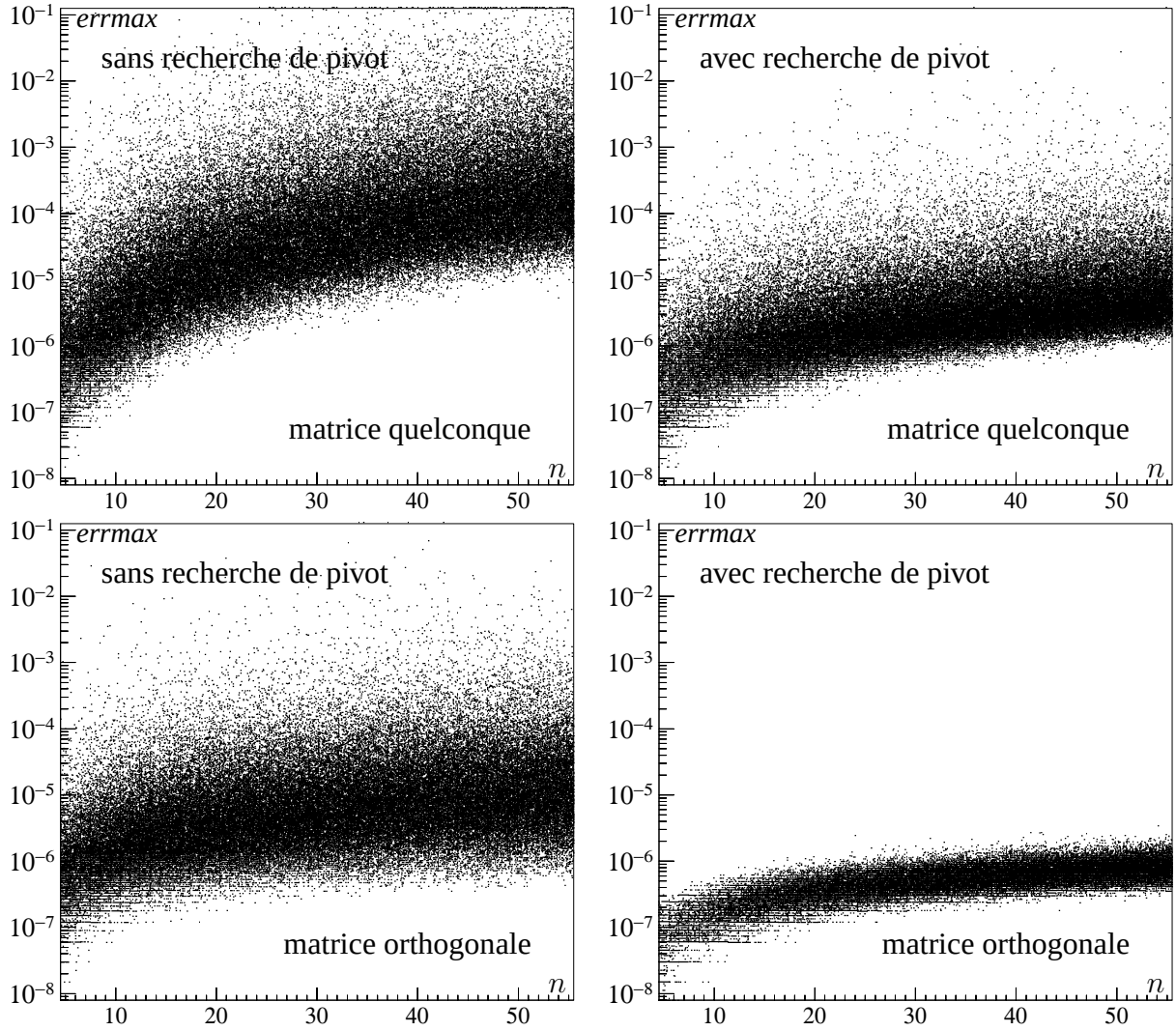


FIG. IV.2: Erreurs pour 1 million de systèmes linéaires de dimensions 5×5 à 55×55

Les pionniers de l'analyse numérique (Hotelling, von Neumann, Goldstine) dans les années '40 ont rencontré d'insurmontables difficultés à analyser les erreurs d'arrondi de la solution des systèmes linéaires. Ils sont arrivés à la conclusion que les dimensions plus grandes que 10 ou 12 seraient impossibles. Malgré ces prédictions pessimistes, les résultats numériques n'ont pas été si mauvais.

Faisons une expérience numérique (voir figure IV.2) : pour chaque $n = 5, 6, \ldots, 55$ nous choisissons 2000 matrices aléatoires avec coefficients a_{ij} uniformément distribués dans $[-1, 1]$ et des solutions x_i uniformément distribuées dans $[-1, 1]$. Alors on calcule en *double précision* les b_j pour cette solution exacte. Ensuite on applique l'algorithme de Gauss, une fois sans recherche de pivot, et une fois avec recherche de pivot, en *simple précision*. L'erreur $\max_i |x_i^{\text{num}} - x_i^{\text{ex}}|$ de chaque résultat est représentée par un petit point dans les dessins supérieurs de la figure IV.2. Bien que nous ne soyons pas surpris par les nombreuses erreurs sans recherche de pivot, *quelques* cas demeurent inacceptables à droite ; bon nombre de résultats restent cependant bons !

Faisons une *deuxième* expérience : une matrice avec a_{ij} uniformément distribués dans $[-1, 1]$ pour $j > i$ est complétée par $a_{ji} = -a_{ij}$, pour assurer que $Q = (I-A)^{-1}(I+A)$ soit orthogonale (Cayley ; voir $\Gamma\epsilon o\mu\epsilon\tau\rho\acute{\iota}\alpha$ II.4). Cette matrice est calculée en double précision, le reste de l'expérience continue comme auparavant (voir les résultats au bas de la figure IV.2). Cette fois-ci il n'y a pas d'exception dans la bonne performance de l'algorithme de Gauss avec recherche de pivot.

La "Backward Error Analysis" de Wilkinson.

L'explication théorique de ces phénomènes a été un des grands *challenges* des années '50. Il paraissait alors difficile d'arriver à un résultat, où même un John von Neumann avait jeté l'éponge !... L'idée miraculeuse a finalement été publiée par Wilkinson (1961, *J. Ass. Comp. Mach.* 8) :

Supposons qu'un système de dimension 2×2 soit à transformer sur forme triangulaire par un pas d'élimination

$$\begin{pmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \end{pmatrix} \quad \text{El. par } \ell_{21} \longrightarrow \quad \begin{pmatrix} a_{11} & a_{12} & b_1 \\ 0 & a_{22} - \ell_{21}a_{12} & b_2 - \ell_{21}b_1 \end{pmatrix}$$

avec $\ell_{21} = a_{21}/a_{11}$.

Première source d'erreur : elle résulte du fait que l'ordinateur calcule avec un *faux*

$$\widehat{\ell}_{21} = \ell_{21} + \epsilon \qquad \text{où } |\epsilon| \leq \textit{eps}$$

(car $|\ell_{21}| \leq 1$ à cause du choix du pivot).

Idée : au lieu de poursuivre les dégâts occasionnés par cette erreur aux calculs ultérieurs et aux solutions, nous *cherchons à modifier les données* pour rendre le calcul (théoriquement) correct : si le a_{21} du début aurait été égal à $a_{21} + \epsilon a_{11}$, ce calcul avait été *sans erreur !*

$$\begin{pmatrix} a_{11} & a_{12} & b_1 \\ a_{21} + \epsilon a_{11} & a_{22} & b_2 \end{pmatrix} \quad \text{El. par } \widehat{\ell}_{21} \longrightarrow \quad \begin{pmatrix} a_{11} & a_{12} & b_1 \\ 0 & a_{22}^{(1)} & b_2^{(1)} \end{pmatrix} .$$

Deuxième source d'erreur. Pour le calcul de $a_{22}^{(1)} = a_{22} - \widehat{\ell}_{21}a_{12}$, il y a une multiplication et une soustraction à faire ; ensuite, le résultat est placé dans une case de mémoire pour $\mathtt{a_{22}}$. Les détails du résultat dépendent de la manière dont le compilateur travaille. Souvent les opérations algébriques en chaîne se font sur un régistre plus long ; seulement la mise en mémoire provoque une erreur d'arrondi notable. Sous cette hypothèse[1], les deux quantités deviennent

$$\widehat{a}_{22}^{(1)} = a_{22}^{(1)} + e_1 \quad \text{et} \quad \widehat{b}_2^{(1)} = b_2^{(1)} + e_2 \,, \qquad \text{où } |e_1| \leq \textit{eps} \cdot |a_{22}^{(1)}| \text{ et } |e_2| \leq \textit{eps} \cdot |b_2^{(1)}|.$$

Nous transportons à nouveau ces erreurs du côté des données, et le calcul

$$\begin{pmatrix} a_{11} & a_{12} & b_1 \\ a_{21} + \epsilon a_{11} & a_{22} + e_1 & b_2 + e_2 \end{pmatrix} \quad \text{El. par } \widehat{\ell}_{21} \longrightarrow \quad \begin{pmatrix} a_{11} & a_{12} & b_1 \\ 0 & \widehat{a}_{22}^{(1)} & \widehat{b}_2^{(1)} \end{pmatrix}$$

[1] qui nous est agréable...

est sans erreur. Nous voyons donc que le résultat *numérique* du système linéaire est le résultat *exact* d'un système dont la deuxième ligne a été modifiée par des quantités $\leq a \cdot eps$ où $a = \max |a_{ij}^{(k)}, b_i^{(k)}|$.

Pour des *systèmes de dimensions supérieures,* on corrige *plusieurs fois* les données a_{ij} ; d'abord pour $i = 2, \ldots, n, \;\; j = 1, \ldots, n$, ensuite pour $i = 3, \ldots, n, \;\; j = 2, \ldots, n$, etc. Nous arrivons au célèbre théorème :

Théorème 2.1 (Wilkinson) *Soit A une matrice inversible et $\widehat{L}$, $\widehat{R}$ le résultat numérique de l'élimination de Gauss (avec recherche de pivot, c.-à-d. $|\widehat{\ell}_{ij}| \leq 1$ pour tout i, j). Alors $\widehat{L}\widehat{R} = \widehat{A}$ avec*

$$|\widehat{a}_{ij} - a_{ij}| \leq a \cdot eps \cdot \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & 0 \\ 1 & 1 & 1 & \ldots & 1 & 1 \\ 1 & 2 & 2 & \ldots & 2 & 2 \\ 1 & 2 & 3 & \ldots & 3 & 3 \\ .. & .. & .. & \ldots & .. & .. \\ 1 & 2 & 3 & \ldots & n-1 & n-1 \end{pmatrix} . \qquad \text{où } \; a = \max_{i,j,k} |a_{ij}^{(k)}|. \qquad (2.1)$$

Définition 2.2 *Un algorithme pour résoudre un problème est numériquement stable (au sens de "backward analysis"), si le résultat numérique peut être interprété comme un résultat exact pour des données légèrement perturbées.*

Par conséquent, si le résultat est faux, ce n'est pas la faute de la méthode, mais bien celle *du problème*. Dans ce cas, on appelle le problème *un problème mal conditionné*. Nous allons étudier ces problèmes plus en détail au paragraphe suivant.

Exemple. Calculons la solution (en simple précision) du système $Ax = b$ avec

$$A = \begin{pmatrix} 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \\ 1/5 & 1/6 & 1/7 & 1/8 \end{pmatrix}, \quad b = \begin{pmatrix} 3511/13860 \\ 277/1540 \\ 40877/291060 \\ 3203/27720 \end{pmatrix}; \qquad \text{sol. exacte} \quad x = \begin{pmatrix} 1/3 \\ 1/11 \\ 1/9 \\ 1/7 \end{pmatrix}.$$

```
.5000000    .3333333    .2500000    .2000000       .2533189
.3333333    .2500000    .2000000    .1666667       .1798701
.2500000    .2000000    .1666667    .1428571       .1404418
.2000000    .1666667    .1428571    .1250000       .1155483
-------------------------------------
.5000000    .3333333    .2500000    .2000000       .2533189
.6666667    .0277778    .0333333    .0333333       .0109909
.5000000    .0333333    .0416667    .0428571       .0137824
.4000000    .0333333    .0428571    .0450000       .0142208
-------------------------------------
.5000000    .3333333    .2500000    .2000000       .2533189
.6666667    .0333333    .0416667    .0428571       .0137824
.5000000    .8333330   -.0013889   -.0023809      -.0004945
.4000000   1.0000000    .0011905    .0021429       .0004384
-------------------------------------
.5000000    .3333333    .2500000    .2000000       .2533189
.6666667    .0333333    .0416667    .0428571       .0137824
.5000000    .8333330   -.0013889   -.0023809      -.0004945
.4000000   1.0000000   -.8571472    .0001020       .0000146
-------------------------------------
x(1)=   .33333951       x(2)=   .09086763
x(3)=   .11118819       x(4)=   .14281443
```

Les résultats montrent que seulement 3 à 4 décimales sont justes. **Mais**, pour l'honneur de notre méthode, nous constatons que les *résidus* de ces solutions res $= Ax - b$ sont correctes :

```
res(1)=  -.00000003      res(2)=  -.00000001
res(3)=   .00000000      res(4)=  -.00000001
```

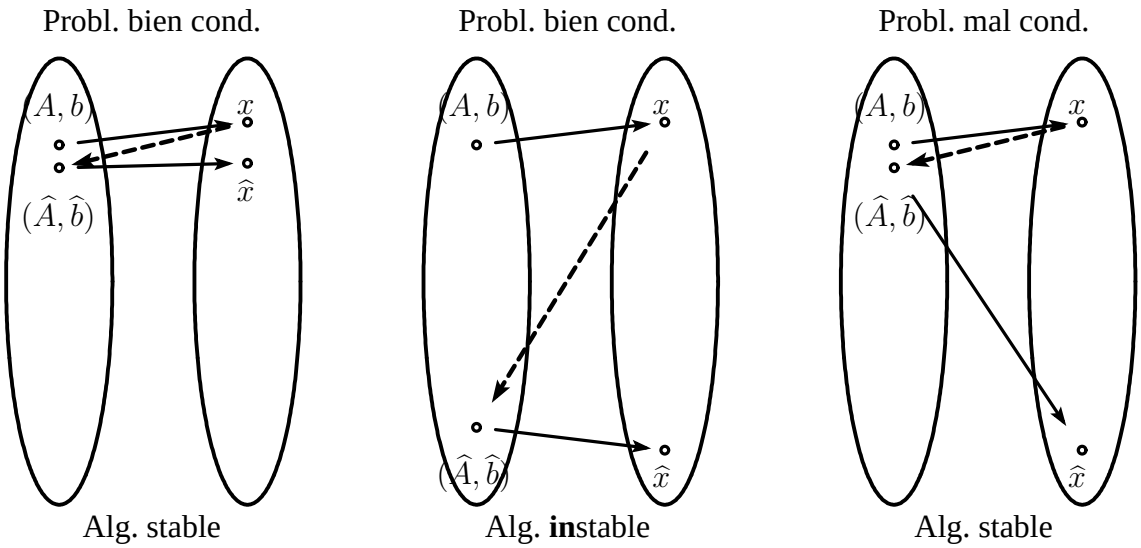


FIG. IV.3: Schéma de la "Backward Error Analysis"

Ce problème est mal conditionné, malgré son apparence débonnaire. Nous sommes donc dans la troisième case du schéma de la Fig. IV.3.

IV.3 La condition d'une matrice

En principe, un problème avec m données et n solutions possède $m \times n$ coefficients décrivant la sensibilité de la n-ème solution par rapport à la m-ème donnée. Devant cette myriade de valeurs, il est parfois préférable d'exprimer la condition *par un seul nombre*. On réussira cela à l'aide de normes de vecteurs et de matrices (recherche initiée par A. Turing 1948).

Rappel sur la norme d'une matrice. Pour une matrice à m lignes et n colonnes, on définit

$$\|A\| = \max_{\|x\|=1} \|Ax\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}, \tag{3.1}$$

c.-à-d., la norme de A est le plus petit nombre $\|A\|$ qui possède la propriété

$$\|Ax\| \leq \|A\| \cdot \|x\| \qquad \text{pour tout} \quad x \in I\!R^n. \tag{3.2}$$

Evidemment, $\|A\|$ dépend des normes choisies dans $I\!R^n$ et $I\!R^m$. Il y a des situations où l'on connaît des formules explicites pour $\|A\|$. Par exemple, si l'on prend la même norme dans les deux espaces alors,

pour $\|x\|_1 = \sum_{i=1}^n |x_i|$, on a

$$\|A\|_1 = \max_{j=1,\ldots,n} \Big(\sum_{i=1}^m |a_{ij}| \Big); \tag{3.3}$$

pour $\|x\|_2 = (\sum_{i=1}^n |x_i|^2)^{1/2}$, on a

$$\|A\|_2 = \sqrt{\text{plus grande valeur propre de } \ A^T A}; \tag{3.4}$$

pour $\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$, on a

$$\|A\|_\infty = \max_{i=1,\dots,m} \Big(\sum_{j=1}^{n} |a_{ij}| \Big). \tag{3.5}$$

La norme $\|A\|$ d'une matrice satisfait toutes les propriétés d'une norme. En plus, elle vérifie $\|I\| = 1$ pour la matrice d'identité et $\|A \cdot B\| \le \|A\| \cdot \|B\|$.

Après ce rappel sur la norme d'une matrice, essayons d'estimer la condition du problème $Ax = b$. Pour ceci, considérons un deuxième système linéaire $\widehat{A}\widehat{x} = \widehat{b}$ avec des données perturbées

$$\begin{aligned} \widehat{a}_{ij} &= a_{ij}(1+\epsilon_{ij}), \qquad |\epsilon_{ij}| \le \epsilon_A, \\ \widehat{b}_i &= b_i(1+\epsilon_i), \qquad\quad |\epsilon_i| \le \epsilon_b, \end{aligned} \tag{3.6}$$

où ϵ_A et ϵ_b spécifient la précision des données (par exemple $\epsilon_A \le$ *eps*, $\epsilon_b \le$ *eps* où *eps* est la précision de l'ordinateur). Les hypothèses (3.6) impliquent (au moins pour les normes $\|\cdot\|_1$ et $\|\cdot\|_\infty$) que

$$\|\widehat{A} - A\| \le \epsilon_A \cdot \|A\|, \qquad \|\widehat{b} - b\| \le \epsilon_b \cdot \|b\|. \tag{3.7}$$

Notre premier résultat donne une estimation de $\|\widehat{x} - x\|$, en supposant que (3.7) soit vrai. Un peu plus loin, on donnera une estimation améliorée valable si (3.6) est satisfait.

Théorème 3.1 *Considérons les deux systèmes linéaires $Ax = b$ et $\widehat{A}\widehat{x} = \widehat{b}$ où A est une matrice inversible. Si (3.7) est vérifié et si $\epsilon_A \cdot \kappa(A) < 1$, alors on a*

$$\frac{\|\widehat{x} - x\|}{\|x\|} \le \frac{\kappa(A)}{1 - \epsilon_A \cdot \kappa(A)} \cdot (\epsilon_A + \epsilon_b) \tag{3.8}$$

où $\kappa(A) := \|A\| \cdot \|A^{-1}\|$. Le nombre $\kappa(A)$ s'appelle condition de la matrice A.

Démonstration. De $\widehat{b} - b = \widehat{A}\widehat{x} - Ax = (\widehat{A} - A)\widehat{x} + A(\widehat{x} - x)$, nous déduisons que

$$\widehat{x} - x = A^{-1}\Big(-(\widehat{A} - A)\widehat{x} + (\widehat{b} - b) \Big). \tag{3.9}$$

Maintenant, prenons la norme de (3.9), utilisons l'inégalité du triangle, les estimations (3.7), $\|\widehat{x}\| \le \|x\| + \|\widehat{x} - x\|$ et $\|b\| = \|Ax\| \le \|A\| \cdot \|x\|$. Nous obtenons ainsi

$$\|\widehat{x} - x\| \le \|A^{-1}\| \Big(\epsilon_A \cdot \|A\| \cdot (\|x\| + \|\widehat{x} - x\|) + \epsilon_b \cdot \|A\| \cdot \|x\| \Big).$$

Ceci donne l'estimation (3.8). □

La formule (3.8) montre que pour $\epsilon_A \cdot \kappa(A) \ll 1$, l'amplification maximale de l'erreur des données sur le résultat est de $\kappa(A)$.

Propriétés de $\kappa(A)$. *Soit A une matrice inversible. Alors,*

a) $\kappa(A) \ge 1$ *pour toute A,*

b) $\kappa(\alpha A) = \kappa(A)$ *pour $\alpha \ne 0$,*

c) $\kappa(A) = \max_{\|y\|=1} \|Ay\| \Big/ \min_{\|z\|=1} \|Az\|$.

La propriété (c) permet d'étendre la définition de $\kappa(A)$ aux matrices de dimension $m \times n$ avec $m \ne n$.

Démonstration. La propriété (a) est une conséquence de $1 = \|I\| = \|AA^{-1}\| \le \|A\| \cdot \|A^{-1}\|$. La propriété (b) est évidente. Pour montrer (c), nous utilisons

$$\|A^{-1}\| = \max_{x \ne 0} \frac{\|A^{-1}x\|}{\|x\|} = \max_{z \ne 0} \frac{\|z\|}{\|Az\|} = \left(\min_{z \ne 0} \frac{\|Az\|}{\|z\|} \right)^{-1}.$$

□

TAB. IV.1: Conditions de matrices de Hilbert et Vandermonde

n	2	4	6	8	10	12
$\kappa(H_n)$	27	$2.8 \cdot 10^4$	$2.9 \cdot 10^7$	$3.4 \cdot 10^{10}$	$3.5 \cdot 10^{13}$	$3.8 \cdot 10^{16}$
$\kappa(V_n)$	8	$5.6 \cdot 10^2$	$3.7 \cdot 10^4$	$2.4 \cdot 10^6$	$1.6 \cdot 10^8$	$1.0 \cdot 10^{10}$

Exemples de matrices ayant une grande condition. Considérons les matrices H_n (matrice de Hilbert) et V_n (matrice de Vandermonde) définies par ($c_j = j/n$)

$$H_n = \Big(\frac{1}{i+j-1}\Big)_{i,j=1}^{n}, \qquad V_n = \Big(c_j^{i-1} \Big)_{i,j=1}^{n}.$$

Leur condition pour la norme $\|\cdot\|_\infty$ est donnée dans le tableau IV.1. La matrice V_n est précisément la matrice du problème d'interpolation polynomiale pour noeuds équidistants. La mauvaise condition de cette matrice est liée au mauvais comportement de cette interpolation que nous avons remcontré au chapitre II.4.

Exemples de matrices ayant une petite condition. Une matrice U est orthogonale si $U^T U = I$. Pour la norme euclidienne, sa condition vaut 1 car $\|U\|_2 = 1$ et $\|U^{-1}\|_2 = 1$ (l'inverse $U^{-1} = U^T$ est aussi orthogonale).

Concernant l'interpolation avec des fonctions splines, nous avons rencontré la matrice (voir le paragraphe II.8, cas équidistant)

$$A = \frac{1}{h} \left(\begin{array}{cccc} 4 & 1 & & \\ 1 & 4 & 1 & \\ & 1 & \ddots & \ddots \\ & & \ddots & \ddots \end{array} \right) \quad \Big\} \, n \tag{3.10}$$

Le facteur $1/h$ n'influence pas $\kappa(A)$. Posons alors $h = 1$. Avec la formule (3.5), on vérifie facilement que $\|A\|_\infty = 6$. Pour estimer $\|A^{-1}\|_\infty$, écrivons A sous la forme $A = 4(I + N)$ où I est l'identité et N contient le reste. On voit que $\|N\|_\infty = 1/2$. En exprimant A^{-1} par une série géométrique, on obtient

$$\|A^{-1}\|_\infty \leq \frac{1}{4}\big(1 + \|N\|_\infty + \|N\|_\infty^2 + \|N\|_\infty^3 + \ldots\big) \leq \frac{1}{2}.$$

Par conséquent, $\kappa_\infty(A) \leq 3$ indépendamment de la dimension du système.

IV.4 L'algorithme de Cholesky

Soit B une matrice quelconque et posons

$$A = B^T B \, , \tag{4.1}$$

i.e., l'élément a_{ij} de A est le produit scalaire des colonnes i et j de B [2]. Alors

$$A \quad \text{est } \textit{symétrique} \quad (\text{car le prod. scal. est symétrique; ou car } \; A^T = B^T (B^T)^T = B^T B = A) \tag{4.2}$$

[2] On appelle cela aussi une *Matrice de Gram.*

et

$$A \text{ est } \textit{définie positive} \quad (\text{i.e., } x^T Ax > 0 \text{ pour } x \neq 0), \tag{4.3}$$

car $x^T Ax = x^T B^T Bx = (Bx)^T (Bx) = y^T y > 0$). Il est nécessaire pour l'inégalité stricte que les colonnes de B sont linéairement indépendantes [3]. Si $Ax = \lambda x$, alors $x^T Ax = \lambda \cdot x^T x > 0$, on voit que chaque valeur propre d'une matrice symétrique et positive définie doit être réelle (voir cours d'Algèbre) et > 0.

Question: Existe-t-il une "décomposition LR" **symétrique**

$$A = L\,L^T \quad \text{ou} \quad \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} \ell_{11} & & & \\ \ell_{21} & \ell_{22} & & \\ \ell_{31} & \ell_{32} & \ell_{33} & \\ \ell_{41} & \ell_{42} & \ell_{43} & \ell_{44} \end{pmatrix} \begin{pmatrix} \ell_{11} & \ell_{21} & \ell_{31} & \ell_{41} \\ & \ell_{22} & \ell_{32} & \ell_{42} \\ & & \ell_{33} & \ell_{43} \\ & & & \ell_{44} \end{pmatrix} \tag{4.4}$$

Il est clair, par (4.1), que A doit être symétrique et définie positive.

Théorème. *Pour chaque matrice symétrique et définie-positive existe une décomposition dite "de Cholesky"* [4] *(4.4). L'algorithme de Cholesky ci-dessous est toujours numériquement stable. Il n'est pas nécessaire de faire une recherche de pivot.*

Calcul des ℓ_{ij}.

Pas 1a. Calculons dans (4.4) la valeur de a_{11}. Elle est

$$a_{11} = (\ell_{11})^2 \qquad \text{donc} \qquad \ell_{11} = \sqrt{a_{11}}\,. \tag{4.5}$$

Question. Est-on sûr que $a_{11} > 0$? Oui, il suffit de poser dans la condition (4.3) le vecteur $x = (1, 0, 0, ...)^T$.

Pas 1b. Calculons dans (4.4) les valeurs de a_{i1} pour $i = 2, 3, 4, ...$ On obtient

$$a_{i1} = \ell_{i1} \cdot \ell_{11} \qquad \text{donc } (\ell_{21} \text{ est connu}) \qquad \ell_{i1} = a_{i1}/\ell_{11}\,. \tag{4.6}$$

La division par ℓ_{11} ne pose pas de problème, car $\ell_{11} > 0$.

Pas 2a. Calculons dans (4.4) la valeur de a_{22}. Elle est

$$a_{22} = (\ell_{21})^2 + (\ell_{22})^2 \qquad \text{donc} \qquad \ell_{22} = \sqrt{a_{22} - (\ell_{21})^2}\,. \tag{4.7}$$

Question. Est-on sûr que $a_{22} - (\ell_{21})^2 > 0$? C'est déjà plus difficile. On pose dans la condition (4.3) le vecteur $x = (u, 1, 0, ...)^T$. Ainsi, $x^T Ax$, que nous savons positif, devient

$$a_{11}u^2 + 2a_{21}u + a_{22} > 0 \tag{4.8}$$

pour *chaque* u. Nous obtenons l'information la meilleure, si nous posons pour u la valeur pour laquelle $a_{11}u^2 + 2a_{21}u$ it est minimale, i.e., où la dérivée $2a_{11}u + 2a_{21} = 0$, i.e., $u = -a_{21}/a_{11}$. Ainsi (4.8) devient, par (4.6) et (4.5),

$$a_{22} - \frac{a_{21}^2}{a_{11}} = a_{22} - \frac{\ell_{21}^2 \ell_{11}^2}{\ell_{11}^2} = a_{22} - (\ell_{21})^2 > 0\,.$$

[3] Sinon, la matrice est *définie semi-positive*.

[4] Le "Commandant Cholesky" (1875–1918) entra à l'École Polytechnique à l'âge de vingt ans et en sortit dans l'arme de l'Artillerie. Affecté à la Section de Géodésie du Service géographique, en juin 1905, il s'y fit remarquer de suite par une intelligence hors ligne, une grande facilité pour les travaux mathématiques, un esprit chercheur, des idées originales, parfois même paradoxales, mais toujours empreintes d'une grande élévation de sentiments et qu'il soutenait avec une extrême chaleur. (...) Cholesky aborda ce problème en apportant dans ses solutions, ... une originalité marquée. Il imagina pour la résolution des équations de condition par la méthode des moindres carrés un procédé de calcul très ingénieux ... (copié du *Bulletin géodésique* No. 1, 1922).

Pas 2b. Calculons dans (4.4) les valeurs de a_{i2} pour $i = 3, 4, ...$ On obtient

$$a_{i2} = \ell_{i1} \cdot \ell_{21} + \ell_{i2} \cdot \ell_{22} \qquad \text{donc} \qquad \ell_{i2} = (a_{i2} - \ell_{i1} \cdot \ell_{21})/\ell_{22} \,. \tag{4.9}$$

La division par ℓ_{22} ne pose pas de problème, car $\ell_{22} > 0$.

Pas 3a. Pour la valeur de a_{33} dans (4.4) on obtient

$$a_{33} = (\ell_{31})^2 + (\ell_{32})^2 + (\ell_{33})^2 \qquad \text{donc} \qquad \ell_{33} = \sqrt{a_{33} - (\ell_{31})^2 - (\ell_{32})^2} \,. \tag{4.10}$$

QUESTION. Est-on sûr que $a_{33} - (\ell_{31})^2 - (\ell_{32})^2 > 0$? Cette fois-ci on va poser dans (4.3) le vecteur $x = (u, v, 1, 0, ...)^T$, i.e.,

$$(\, u \quad v \quad 1 \,) \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = a_{11}u^2 + 2a_{21}uv + \ldots + a_{33} > 0 \tag{4.11}$$

pour tout u et v. On va de nouveau chercher la valeur minimale de cette expression quadratique. Pour ne pas nous perdre dans les calculs, observons que

$$\begin{pmatrix} \ell_{11} & & \\ \ell_{21} & \ell_{22} & \\ \ell_{31} & \ell_{32} & 0 \end{pmatrix} \begin{pmatrix} \ell_{11} & \ell_{21} & \ell_{31} \\ & \ell_{22} & \ell_{32} \\ & & 0 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & \ell_{31}^2 + \ell_{32}^2 \end{pmatrix} \,. \tag{4.12}$$

Ainsi, l'expression (4.11) est égale à

$$y^T y + a_{33} - \ell_{31}^2 - \ell_{32}^2 > 0 \qquad \text{avec} \quad y = \begin{pmatrix} \ell_{11} & \ell_{21} & \ell_{31} \\ & \ell_{22} & \ell_{32} \\ & & 0 \end{pmatrix} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \,. \tag{4.13}$$

Pour $\ell_{11}u + \ell_{21}v = -\ell_{31}$ et $\ell_{22}v = -\ell_{32}$ nous avons $y = 0$ et (4.13) devient l'estimation recherchée.

Algorithme de Cholesky. On continue anisi avec les pas 3b, 4a, 4b, etc., et on obtient l'algorithme suivant :

for $k := 1$ **to** n **do**
  $\ell_{kk} := \sqrt{a_{kk} - \sum_{j=1}^{k-1} \ell_{kj}^2}$;
  for $i := k + 1$ **to** n **do**
    $\ell_{ik} := (a_{ik} - \sum_{j=1}^{k-1} \ell_{ij}\ell_{kj})/\ell_{kk}$.

Coût de cet algorithme. En négligeant les n racines, le nombre d'opérations nécessaires est d'environ

$$\sum_{k=1}^{n} (n-k) \cdot k \approx \int_0^n (n-x)x \, dx = \frac{n^3}{6}.$$

L'algorithme est deux fois plus rapide que la décomposition LR de Gauss.

Solution du système linéraire. Pour résoudre le système $Ax = b$, on calcule d'abord la décomposition de Cholesky $A = LL^T$. Puis

$$L \underbrace{L^T x}_{c} = b \quad \Rightarrow \text{ résoudre successivement les systèmes } \; Lc = b \; \text{ et } \; L^T x = c$$

dont les matrices sont triangulaires.

IV.5 Systèmes surdéterminés – méthode des moindres carrés

Considérons un système d'équations linéaires

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \vdots \qquad \vdots \qquad\qquad \vdots \quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m \end{array} \tag{5.1}$$

où $m \geq n$ (matriciellement: $Ax = b$ avec $x \in \mathbb{R}^n$ et $b \in \mathbb{R}^m$; A est une matrice $m \times n$). Evidemment, le système (5.1) ne possède, en général, pas de solution. L'idée est de chercher un vecteur x tel que

$$\|Ax - b\|_2 \to \min \tag{5.2}$$

pour la norme euclidienne. Une justification probabiliste de cette condition sera donnée dans le paragraphe IV.7. Le nom "méthode des moindres carrés" indique le choix de la norme dans (5.2) (la somme des carrés des erreurs doit être minimale).

Théorème 5.1 *Soit A une matrice $m \times n$ (avec $m \geq n$) et soit $b \in \mathbb{R}^m$. Le vecteur x est solution de (5.2) si et seulement si*

$$A^T Ax = A^T b. \tag{5.3}$$

Les équations du système (5.3) s'appellent "équations normales".

Démonstration. Les minima de la fonction quadratique

$$f(x) := \|Ax - b\|^2 = (Ax - b)^T(Ax - b) = x^T A^T Ax - 2x^T A^T b + b^T b$$

sont donnés par $0 = f'(x) = 2(x^T A^T A - b^T A)$. □

Interprétation géométrique. L'ensemble $E = \{Ax \mid x \in \mathbb{R}^n\}$ est un sous-espace linéaire de $\mathbb{R}^m$. Pour un $b \in \mathbb{R}^m$ arbitraire, x est une solution de (5.2) si et seulement si Ax est la projection orthogonale de b sur E. Ceci signifie que $Ax - b \perp Az$ pour tout $z \in \mathbb{R}^n$. On en déduit que $A^T(Ax - b) = 0$ et on a ainsi établi une deuxième démonstration de (5.3).

Exemple 5.2 Pour étudier le phénomène de la thermo-électricité, on fait l'expérience suivante. On soude un fil de cuivre avec un fil de constantan de manière à obtenir une boucle fermée. Un point de soudure est maintenu à température fixe ($T_0 \approx 24°$C), alors que l'on fait varier la température T de l'autre. Ceci génère une tension U, laquelle est mesurée en fonction de T (voir le tableau IV.2 et la fig. IV.4). Les données du tableau IV.2 sont prises du livre de P.R. Bevington[5].

On suppose que cette dépendance obéit à la loi

$$U = a + bT + cT^2 \tag{5.4}$$

et on cherche à déterminer les paramètres a, b et c. Les données du tableau IV.2 nous conduisent au système surdéterminé ($n = 3$, $m = 21$)

$$U_i = a + bT_i + cT_i^2, \qquad i = 1, \ldots, 21. \tag{5.5}$$

En résolvant les équations normales (5.3) pour ce problème, on obtient $a = -0.886$, $b = 0.0352$ et $c = 0.598 \cdot 10^{-4}$. Avec ces paramètres, la fonction (5.4) est dessinée dans la fig. IV.4. On observe une très bonne concordance avec les données.

TAB. IV.2: Tensions mesurées en fonction de la température T

i	T_i°C	U_i	i	T_i°C	U_i	i	T_i°C	U_i
1	0	−0.89	8	35	0.42	15	70	1.88
2	5	−0.69	9	40	0.61	16	75	2.10
3	10	−0.53	10	45	0.82	17	80	2.31
4	15	−0.34	11	50	1.03	18	85	2.54
5	20	−0.15	12	55	1.22	19	90	2.78
6	25	0.02	13	60	1.45	20	95	3.00
7	30	0.20	14	65	1.68	21	100	3.22

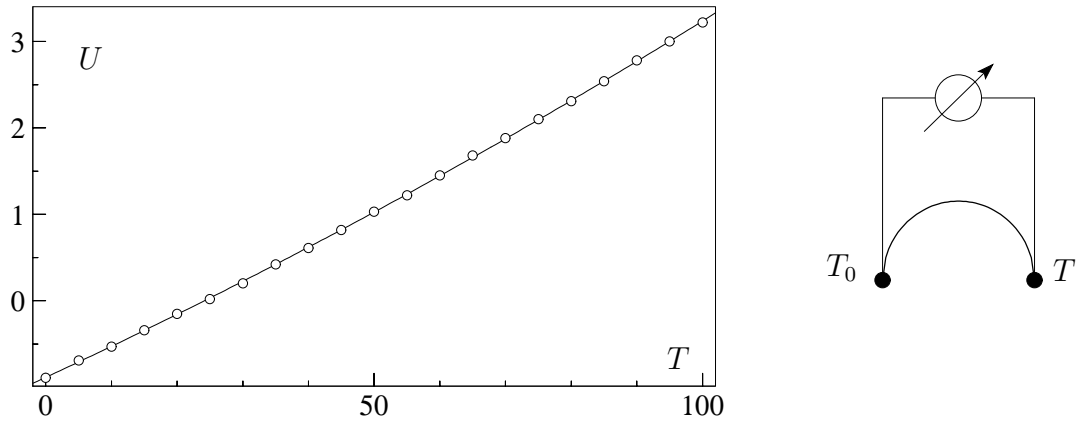


FIG. IV.4: Tension en fonction de la température et schéma de l'expérience

Remarque. Les équations normales (5.3) possèdent toujours au moins une solution (la projection sur E existe toujours). La matrice A^TA est symétrique et non-négative ($x^TA^TAx = \|Ax\|^2 \geq 0$). Elle est définie positive si les colonnes de A sont linéairement indépendantes ($Ax \neq 0$ pour $x \neq 0$). Dans cette situation, on peut appliquer l'algorithme de Cholesky pour résoudre le système (5.3). Mais, souvent, il est préférable de calculer la solution directement de (5.2) sans passer par les équations normales (5.3).

IV.6 Décomposition QR d'une matrice

Dans l'élimination de Gauss, on a multiplié l'équation $Ax = b$ par la matrice triangulaire $L_{n-1} \cdot \ldots \cdot L_2 \cdot L_1$. De cette manière, on a réduit le problème original à $Rx = c$ où R est une matrice triangulaire supérieure. Malheureusement, la multiplication de $Ax - b$ avec L_i ne conserve pas la norme du vecteur.

Pour résoudre (5.2), nous cherchons une matrice orthogonale Q telle que

$$Q^T(Ax - b) = Rx - c = \begin{pmatrix} R' \\ 0 \end{pmatrix} x - \begin{pmatrix} c' \\ c'' \end{pmatrix} \tag{6.1}$$

où R' (une matrice carrée de dimension n) est triangulaire supérieure et $(c', c'')^T$ est la partition de $c = Q^Tb$ telle que $c' \in \mathbb{R}^n$ et $c'' \in \mathbb{R}^{m-n}$. Comme le produit par une matrice orthogonale ne

[5]P.R. Bevington (1969): *Data reduction and error analysis for the physical sciences*. McGraw-Hill Book Company).

change pas la norme du vecteur, on a

$$\|Ax-b\|_2^2 = \|Q^T(Ax-b)\|_2^2 = \|Rx-c\|_2^2 = \|R'x-c'\|_2^2 + \|c''\|_2^2. \tag{6.2}$$

On obtient alors la solution de (5.2) en résolvant le système

$$R'x = c'. \tag{6.3}$$

Le problème consiste à calculer une matrice orthogonale Q (c.-à-d., $Q^TQ = I$) et une matrice triangulaire supérieure R telles que $Q^TA = R$ ou de façon équivalente

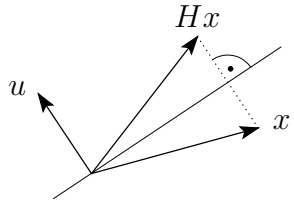
$$A = QR. \tag{6.4}$$

Cette factorisation s'appelle la "décomposition QR" de la matrice A. Pour arriver à ce but, on peut se servir des rotations de Givens (voir exercice **??** du chapitre V) ou des réflexions de Householder.

Réflexions de Householder (1958). Une matrice de la forme

$$H = I - 2uu^T \qquad \text{où} \qquad u^Tu = 1 \tag{6.5}$$

a les propriétés suivantes :



- H est une réflexion à l'hyper-plan $\{x \mid u^Tx = 0\}$ car $Hx = x - u\cdot(2u^Tx)$ et $Hx + x \perp u$.
- H est symétrique.
- H est orthogonale, car

$$H^TH = (I-2uu^T)^T(I-2uu^T) = I - 4uu^T + 4uu^Tuu^T = I.$$

En multipliant A avec des matrices de Householder, nous allons essayer de transformer A en une matrice de forme triangulaire.

L'algorithme de Householder - Businger - Golub. Dans une *première étape*, on cherche une matrice $H_1 = I - 2u_1u_1^T$ ($u_1 \in \mathbb{R}^m$ et $u_1^Tu_1 = 1$) telle que

$$H_1A = \begin{pmatrix} \alpha_1 & \times & \cdots & \times \\ 0 & \times & \cdots & \times \\ \vdots & \vdots & & \vdots \\ 0 & \times & \cdots & \times \end{pmatrix}. \tag{6.6}$$

Si l'on dénote par A_1 la première colonne de A, il faut que $H_1A_1 = \alpha_1e_1 = (\alpha_1, 0, \ldots, 0)^T$ et on obtient $|\alpha_1| = \|H_1A_1\|_2 = \|A_1\|_2$. La forme particulière de H_1 implique que

$$H_1A_1 = A_1 - 2u_1\cdot u_1^TA_1 = \alpha_1e_1.$$

L'expression $u_1^TA_1$ est un scalaire. Par conséquent,

$$u_1 = C\cdot v_1 \qquad \text{où} \qquad v_1 = A_1 - \alpha_1e_1 \tag{6.7}$$

et la constante C est déterminée par $\|u_1\|_2 = 1$. Comme on a encore la liberté de choisir le signe de α_1, posons

$$\alpha_1 = -\text{sign}(a_{11})\cdot\|A_1\|_2 \tag{6.8}$$

pour éviter une soustraction mal conditionnée dans le calcul de $v_1 = A_1 - \alpha_1 e_1$.

Calcul de $H_1 A$. Notons par A_j et $(H_1 A)_j$ les $j^{\text{èmes}}$ colonnes de A et $H_1 A$ respectivement. Alors, on a

$$(H_1 A)_j = A_j - 2u_1 u_1^T A_j = A_j - \beta \cdot v_1^T A_j \cdot v_1 \qquad \text{où} \qquad \beta = \frac{2}{v_1^T v_1}. \tag{6.9}$$

Le facteur β peut être calculé à l'aide de

$$\beta^{-1} = \frac{v_1^T v_1}{2} = \frac{1}{2}\left(A_1^T A_1 - 2\alpha_1 a_{11} + \alpha_1^2\right) = -\alpha_1(a_{11} - \alpha_1). \tag{6.10}$$

Dans une *deuxième étape*, on applique la procédure précédente à la sous-matrice de dimension $(m-1) \times (n-1)$ de (6.6). Ceci donne un vecteur $\bar{u}_2 \in I\!R^{m-1}$ et une matrice de Householder $\bar{H}_2 = I - 2\bar{u}_2 \bar{u}_2^T$. En posant $u_2 = (0, \bar{u}_2)^T$, une multiplication de (6.6) par la matrice $H_2 = I - 2u_2 u_2^T$ donne

$$H_2 H_1 A = H_2 \begin{pmatrix} \alpha_1 & \times & \cdots & \times \\ 0 & & & \\ \vdots & & C & \\ 0 & & & \end{pmatrix} = \begin{pmatrix} \alpha_1 & \times & \cdots & \times \\ 0 & & & \\ \vdots & & \bar{H}_2 C & \\ 0 & & & \end{pmatrix} = \begin{pmatrix} \alpha_1 & \times & \times & \cdots & \times \\ 0 & \alpha_2 & \times & \cdots & \times \\ 0 & 0 & \times & \cdots & \times \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \times & \cdots & \times \end{pmatrix}.$$

En continuant cette procédure, on obtient après n étapes (après $n-1$ étapes si $m = n$) une matrice triangulaire

$$\underbrace{H_n \cdot \ldots \cdot H_2 H_1}_{Q^T} A = R = \begin{pmatrix} R' \\ 0 \end{pmatrix}.$$

Ceci donne la décomposition (6.4) avec $Q^T = H_n \cdot \ldots \cdot H_2 H_1$.

Coût de la décomposition QR. La première étape exige le calcul de α_1 par la formule (6.8) ($\approx m$ opérations), le calcul de $2/v_1^T v_1$ par la formule (6.10) (travail négligeable) et le calcul de $(H_1 A)_j$ pour $j = 2, \ldots, n$ par la formule (6.9) ($\approx (n-1) \cdot 2 \cdot m$ opérations). En tout, cette étape nécessite environ $2mn$ opérations. Pour la décomposition QR, on a alors besoin de

$2(n^2 + (n-1)^2 + \ldots + 1) \approx 2n^3/3$ opérations si $m = n$ (matrice carrée);

$2m(n + (n-1) + \ldots + 1) \approx mn^2$ opérations si $m \gg n$.

En comparant encore ce travail avec celui de la résolution des équations normales ($\approx mn^2/2$ opérations pour le calcul de $A^T A$ et $\approx n^3/6$ opérations pour la décomposition de Cholesky de $A^T A$), on voit que la décomposition QR coûte au pire le double.

Remarque. Si les colonnes de la matrice A sont linéairement indépendantes, tous les α_i sont non nuls et l'algorithme de Householder–Businger–Golub est applicable. Une petite modification (échange des colonnes de A) permet de traiter aussi le cas général.

Concernant la programmation, il est important de ne calculer ni les matrices H_i, ni la matrice Q. On retient simplement les valeurs α_i et les vecteurs v_i (pour $i = 1, \ldots, n$) qui contiennent déjà toutes les informations nécessaires pour la décomposition. Comme pour l'élimination de Gauss, on écrit deux sous-programmes. DECQR fournit la décomposition QR de la matrice A (c.-à-d. les α_i, v_i et la matrice R). Le sous-programme SOLQR calcule $Q^T b$ et la solution du système triangulaire $R'x = c'$ (voir (6.3)). Le calcul de $Q^T b = H_n \cdot \ldots \cdot H_2 H_1 b$ se fait avec une formule analogue à (6.9).

Exemple 6.1 Si les colonnes de A sont "presque" linéairement dépendantes, la résolution du problème (5.2) à l'aide de la décomposition QR est préférable à celle des équations normales. Considérons, par exemple,

$$A = \begin{pmatrix} 1 & 1 \\ \epsilon & 0 \\ 0 & \epsilon \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

où ϵ est une petite constante, disons $\epsilon^2 < \mathit{eps}$. Avec un calcul exact, on obtient

$$A^T A = \begin{pmatrix} 1+\epsilon^2 & 1 \\ 1 & 1+\epsilon^2 \end{pmatrix}, \qquad A^T b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

et la solution est donnée par

$$x_1 = x_2 = \frac{1}{2+\epsilon^2} = \frac{1}{2} + \mathcal{O}(\epsilon^2).$$

Un calcul en virgule flottante fait disparaître le ϵ^2 dans $A^T A$ et cette matrice devient singulière. On n'obtient pas de solution.

Par contre, l'algorithme de Householder–Businger–Golub donne (en négligeant ϵ^2) $\alpha_1 = -1$, $v_1 = (2, \epsilon, 0)^T$,... et à la fin

$$R = \begin{pmatrix} -1 & -1 \\ 0 & \sqrt{2}\cdot\epsilon \\ 0 & 0 \end{pmatrix}, \qquad Q^T b = \begin{pmatrix} -1 \\ \epsilon/\sqrt{2} \\ -\epsilon/\sqrt{2} \end{pmatrix}.$$

La résolution de (6.3) donne une bonne approximation de la solution exacte.

Calcul pour l'exemple 5.2 (dont les données sont multipliées par 100 ; voir équation (7.11) ci-dessous) :

$$\begin{pmatrix} 100 & 0 & 0 & -88 \\ 100 & 500 & 2500 & -68 \\ 100 & 1000 & 10000 & -52 \\ 100 & 1500 & 22500 & -33 \\ 100 & 2000 & 40000 & -14 \\ 100 & 2500 & 62500 & 2 \\ 100 & 3000 & 90000 & 20 \\ 100 & 3500 & 122500 & 42 \\ 100 & 4000 & 160000 & 61 \\ 100 & 4500 & 202500 & 82 \\ 100 & 5000 & 250000 & 103 \\ 100 & 5500 & 302500 & 122 \\ 100 & 6000 & 360000 & 145 \\ 100 & 6500 & 422500 & 168 \\ 100 & 7000 & 490000 & 188 \\ 100 & 7500 & 562500 & 210 \\ 100 & 8000 & 640000 & 231 \\ 100 & 8500 & 722500 & 254 \\ 100 & 9000 & 810000 & 278 \\ 100 & 9500 & 902500 & 300 \\ 100 & 10000 & 1000000 & 322 \end{pmatrix} \qquad \begin{pmatrix} -457 & -22912 & -1565712 & -494 \\ 0 & -3603 & -277963 & -141 \\ 0 & -3103 & -270463 & -125 \\ 0 & -2603 & -257963 & -106 \\ 0 & -2103 & -240463 & -87 \\ 0 & -1603 & -217963 & -70 \\ 0 & -1103 & -190463 & -52 \\ 0 & -603 & -157963 & -30 \\ 0 & -103 & -120463 & -11 \\ 0 & 396 & -77963 & 9 \\ 0 & 896 & -30463 & 30 \\ 0 & 1396 & 22036 & 49 \\ 0 & 1896 & 79536 & 72 \\ 0 & 2396 & 142036 & 95 \\ 0 & 2896 & 209536 & 115 \\ 0 & 3396 & 282036 & 137 \\ 0 & 3896 & 359536 & 158 \\ 0 & 4396 & 442036 & 181 \\ 0 & 4896 & 529536 & 205 \\ 0 & 5396 & 622036 & 227 \\ 0 & 5896 & 719536 & 249 \end{pmatrix}$$

$$\begin{pmatrix} -457 & -22912 & -1565712 & -494 \\ 0 & 13874 & 1387444 & 572 \\ 0 & 0 & 25324 & 1 \\ 0 & 0 & -9816 & 0 \\ 0 & 0 & -39957 & -1 \\ 0 & 0 & -65098 & -4 \\ 0 & 0 & -85238 & -7 \\ 0 & 0 & -100379 & -5 \\ 0 & 0 & -110520 & -6 \\ 0 & 0 & -115661 & -6 \\ 0 & 0 & -115802 & -5 \\ 0 & 0 & -110943 & -7 \\ 0 & 0 & -101083 & -4 \\ 0 & 0 & -86224 & -2 \\ 0 & 0 & -66365 & -2 \\ 0 & 0 & -41506 & 0 \\ 0 & 0 & -11647 & 0 \\ 0 & 0 & 23212 & 2 \\ 0 & 0 & 63071 & 5 \\ 0 & 0 & 107930 & 7 \\ 0 & 0 & 157789 & 9 \end{pmatrix} \quad \begin{pmatrix} -457 & -22912 & -1565712 & -494 \\ 0 & 13874 & 1387444 & 572 \\ 0 & 0 & -374437 & -21 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

IV.7 Etude de l'erreur de la méthode des moindres carrés

Supposons d'avoir un système surdéterminé

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \qquad i = 1, \ldots, m \,. \tag{7.1}$$

En pratique, les b_i sont des mesures légèrement erronées et il est naturel de les considérer comme des valeurs plus ou moins aléatoires. L'étude de l'erreur de la solution x, obtenue par la méthode des moindres carrés, se fait alors dans le cadre de la théorie des probabilités.

Rappel sur la théorie des probabilités. Considérons des *variables aléatoires* X (dites "continues") qui sont spécifiées par une fonction de densité $f : I\!R \to I\!R$, c.-à-d., la probabilité de l'événement que la valeur de X se trouve dans l'intervalle $[a, b)$ est donnée par

$$P(a \leq X < b) = \int_a^b f(x)\, dx \tag{7.2}$$

avec $f(x) \geq 0$ pour $x \in I\!R$ et $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

On appelle *espérance* (mathématique) de la variable aléatoire X le nombre réel

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x f(x)\, dx, \tag{7.3}$$

et *variance* la valeur

$$\sigma_X^2 = \mathit{Var}\,(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)\, dx = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu_X^2. \tag{7.4}$$

Exemple 7.1 Si une variable aléatoire satisfait (7.2) avec (voir la fig. IV.5)

$$f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\exp\Big(-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big) \tag{7.5}$$

alors on dit que la variable aléatoire satisfait la *loi normale* ou la *loi de Gauss – Laplace* que l'on symbolise par $N(\mu,\sigma^2)$. On vérifie facilement que μ est l'espérance et σ^2 la variance de cette variable aléatoire.

La loi normale est parmi les plus importantes en probabilités. Une raison est due au "théorème de la limite centrale" qui implique que les observations pour la plupart des expériences physiques obéissent à cette loi.

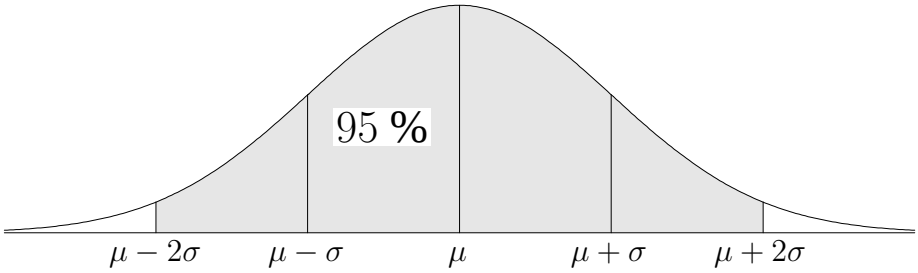


FIG. IV.5: Fonction de densité pour la loi normale

Rappelons aussi que n variables aléatoires $X_1,\dots,X_n$ sont indépendantes si, pour tout a_i, b_i, on a

$$P(a_i \le X_i < b_i,\ i=1,\dots,n) = \prod_{i=1}^{n} P(a_i \le X_i < b_i). \tag{7.6}$$

Lemme 7.2 *Soient X et Y deux variables aléatoires indépendantes avec comme fonctions de densité $f(x)$ et $g(y)$ respectivement et soient $\alpha,\beta\in\mathbb{R}$ avec $\alpha\neq 0$. Alors, les variables aléatoires $\alpha X+\beta$ et $X+Y$ possèdent les fonctions de densité*

$$\frac{1}{|\alpha|}\,f\Big(\frac{x-\beta}{\alpha}\Big) \qquad \textit{et} \qquad (f*g)(z) = \int_{-\infty}^{\infty} f(z-y)g(y)\,dy. \tag{7.7}$$

Leur espérance mathématique est

$$E(\alpha X+\beta) = \alpha E(X)+\beta, \qquad E(X+Y) = E(X)+E(Y) \tag{7.8}$$

et leur variance satisfait

$$\textit{Var}\,(\alpha X+\beta) = \alpha^2 \textit{Var}\,(X), \qquad \textit{Var}\,(X+Y) = \textit{Var}\,(X)+\textit{Var}\,(Y). \tag{7.9}$$

Démonstration. La fonction de densité pour la variable aléatoire $\alpha X+\beta$ découle de (pour $\alpha>0$)

$$P(a\le \alpha X+\beta < b) = P\Big(\frac{a-\beta}{\alpha}\le X < \frac{b-\beta}{\alpha}\Big) = \int_{(a-\beta)/\alpha}^{(b-\beta)/\alpha} f(x)\,dx = \int_a^b \alpha^{-1} f\Big(\frac{t-\beta}{\alpha}\Big)\,dt.$$

Les propriétés (7.8) et (7.9) pour $\alpha X+\beta$ en sont une conséquence directe.

Comme X et Y sont supposées indépendantes, on obtient (en posant $z=x+y$)

$$P(a\le X+Y<b) = \iint_{a\le x+y<b} f(x)g(y)\,dx\,dy = \int_a^b \int_{-\infty}^{\infty} f(z-y)g(y)\,dy\,dz$$

et on trouve la fonction de densité pour $X+Y$. Un calcul direct donne

$$E(X+Y) = \int_{-\infty}^{\infty} z \int_{-\infty}^{\infty} f(z-y)g(y)\,dy\,dz = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)f(x)g(y)\,dy\,dx = E(X)+E(Y)$$

et, de façon similaire, on obtient

$$\begin{aligned} \mathit{Var}\,(X+Y) &= \int_{-\infty}^{\infty} z^2 \int_{-\infty}^{\infty} f(z-y)g(y)\,dy\,dz - \mu_{X+Y}^2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)^2 f(x)g(y)\,dy\,dx - (\mu_X+\mu_Y)^2 = \mathit{Var}\,(X) + \mathit{Var}\,(Y). \end{aligned}$$ □

Remarque. Si X et Y sont deux variables aléatoires indépendantes qui obéissent à la loi normale, les variables aléatoires $\alpha X+\beta$ et $X+Y$ obéissent aussi à cette loi (exercice 13).

Retour au problème (7.1). Pour pouvoir estimer l'erreur du résultat numérique x, faisons les hypothèses suivantes :

H1: La valeur b_i est la réalisation d'une épreuve pour une variable aléatoire B_i. On suppose que les B_i soient indépendantes et qu'elles obéissent à la loi de Gauss–Laplace avec β_i comme espérance et σ_i^2 comme variance (les β_i sont inconnus, mais les σ_i^2 sont supposés connus).

H2: Le système surdéterminé (7.1) possède une solution unique si l'on remplace les b_i par les nombres β_i, c.-à-d. qu'il existe un vecteur $\xi \in I\!R^n$ tel que $A\xi = \beta$ où $\beta = (\beta_1, \ldots, \beta_m)^T$.

Une illustration de cette situation est donnée en Fig. IV.6.

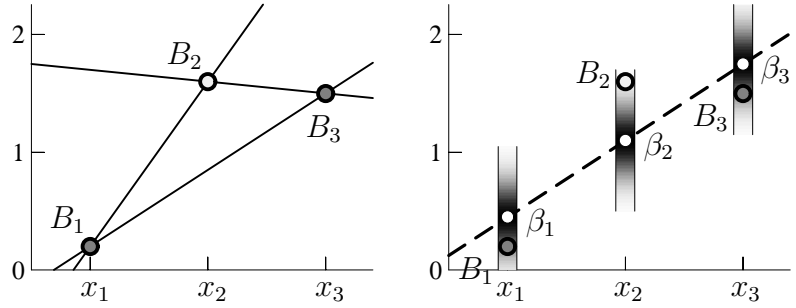


FIG. IV.6: Illustration pour les hypothèses H1 et H2 (les probabilités sont repésentées par un dégradé de gris).

Motivation de la méthode des moindres carrés par "maximum likelihood". Par l'hypothèse H1, la probabilité que B_i soit dans l'intervalle $[b_i, b_i+db_i)$ avec db_i (infinimement) petit est

$$P(b_i \le B_i < b_i + db_i) \approx \frac{1}{\sqrt{2\pi}\cdot\sigma_i} \cdot \exp\Big(-\frac{1}{2}\Big(\frac{b_i-\beta_i}{\sigma_i}\Big)^2\Big)\cdot db_i.$$

Comme les B_i sont indépendants, la formule (7.6) implique que

$$\begin{aligned} P(b_i \le B_i < b_i + db_i,\ i=1,\ldots,m) &\approx \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\cdot\sigma_i} \cdot \exp\Big(-\frac{1}{2}\Big(\frac{b_i-\beta_i}{\sigma_i}\Big)^2\Big)\cdot db_i \qquad (7.10) \\ &= C\cdot\exp\Big(-\frac{1}{2}\sum_{i=1}^{m}\Big(\frac{b_i-\beta_i}{\sigma_i}\Big)^2\Big) = C\cdot\exp\Big(-\frac{1}{2}\sum_{i=1}^{m}\Big(\frac{b_i-\sum_{j=1}^{n}a_{ij}\xi_j}{\sigma_i}\Big)^2\Big). \end{aligned}$$

Selon une idée de Gauss (1812), la "meilleure" réponse x_i pour les ξ_i (inconnus) est celle pour laquelle la probabilité (7.10) est maximale ("maximum likelihood"). Alors, on calcule $x_1, \ldots, x_n$ de façon à ce que

$$\sum_{i=1}^{m}\Big(\frac{b_i}{\sigma_i} - \sum_{j=1}^{n}\frac{a_{ij}}{\sigma_i}\cdot x_j\Big)^2 \to \min. \qquad (7.11)$$

Si l'on remplace b_i/σ_i par b_i et a_{ij}/σ_i par a_{ij}, la condition (7.11) est équivalente à (5.2). Par la suite, nous supposerons que cette normalisation soit déjà effectuée (donc, $\sigma_i = 1$ pour $i = 1, \ldots, n$).

Estimation de l'erreur. La solution de (7.11) est donnée par $x = (A^TA)^{-1}A^Tb$. La solution théorique satisfait $\xi = (A^TA)^{-1}A^T\beta$. Alors,

$$x - \xi = (A^TA)^{-1}A^T(b - \beta) \qquad \text{ou} \qquad x_i - \xi_i = \sum_{j=1}^{m} \alpha_{ij}(b_j - \beta_j)$$

où α_{ij} est l'élément (i, j) de la matrice $(A^TA)^{-1}A^T$. L'idée est de considérer la valeur x_i comme la réalisation d'une variable aléatoire X_i définie par

$$X_i = \sum_{j=1}^{m} \alpha_{ij}B_j \qquad \text{ou} \qquad X_i - \xi_i = \sum_{j=1}^{m} \alpha_{ij}(B_j - \beta_j). \tag{7.12}$$

Théorème 7.3 *Soient $B_1, \ldots, B_m$ des variables aléatoires indépendantes avec β_i comme espérance et $\sigma_i = 1$ comme variance. Alors, la variable aléatoire X_i, définie par (7.12), satisfait*

$$E(X_i) = \xi_i \qquad \textit{et} \qquad \textit{Var}\,(X_i) = \epsilon_{ii} \tag{7.13}$$

où ϵ_{ii} est le $i^{\text{ème}}$ élément de la diagonale de $(A^TA)^{-1}$.

Remarque. Les autres éléments de $(A^TA)^{-1}$ sont les covariances de X_i et X_j.

Démonstration. La formule (7.8) donne $E(X_i) = \xi_i$. Pour calculer la variance de X_i, nous utilisons le fait que $\textit{Var}\,(B_i) = 1$ et la formule (7.9). Ceci donne avec $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ que

$$\sigma^2_{X_i} = \sum_{j=1}^{m} \alpha_{ij}^2 = \|e_i^T(A^TA)^{-1}A^T\|_2^2 = e_i^T(A^TA)^{-1}A^TA(A^TA)^{-1}e_i = e_i^T(A^TA)^{-1}e_i = \epsilon_{ii}. \qquad \square$$

Exemple 7.4 Pour l'expérience sur la thermo-électricité (voir le paragraphe IV.5), on a supposé que les mesures b_i ont été faites avec une précision correspondant à $\sigma_i = 0.01$. Pour le système surdéterminé (on écrit x_1, x_2, x_3 pour a, b, c et b_i pour U_i)

$$\frac{1}{\sigma_i} \cdot x_1 + \frac{T_i}{\sigma_i} \cdot x_2 + \frac{T_i^2}{\sigma_i} \cdot x_3 = \frac{b_i}{\sigma_i}, \qquad i = 1, \ldots, 21$$

la matrice $(A^TA)^{-1}$ devient

$$(A^TA)^{-1} = \begin{pmatrix} 0.356 \cdot 10^{-4} & -0.139 \cdot 10^{-5} & 0.113 \cdot 10^{-7} \\ -0.139 \cdot 10^{-5} & 0.765 \cdot 10^{-7} & -0.713 \cdot 10^{-9} \\ 0.113 \cdot 10^{-7} & -0.713 \cdot 10^{-9} & 0.713 \cdot 10^{-11} \end{pmatrix} \tag{7.14}$$

et on obtient

$$\sigma_{X_1} = 0.60 \cdot 10^{-2}, \qquad \sigma_{X_2} = 0.28 \cdot 10^{-3}, \qquad \sigma_{X_3} = 0.27 \cdot 10^{-5}.$$

Ceci implique qu'avec une probabilité de 95%, la solution exacte (si elle existe) satisfait

$$a = -0.886 \pm 0.012, \qquad b = 0.0352 \pm 0.0006, \qquad c = 0.598 \cdot 10^{-4} \pm 0.054 \cdot 10^{-4}.$$

Test de confiance du modèle. Etudions encore si les données sont compatibles avec l'hypothèse H2.

En utilisant la décomposition QR de la matrice A, le problème surdéterminé $Ax = b$ se transforme en (voir (6.1))

$$\begin{pmatrix} R' \\ 0 \end{pmatrix} x = \begin{pmatrix} c' \\ c'' \end{pmatrix} \qquad \text{où} \qquad \begin{pmatrix} c' \\ c'' \end{pmatrix} = Q^T b. \tag{7.15}$$

La grandeur de $\|c''\|_2^2$ est une mesure de la qualité du résultat numérique. Théoriquement, si l'on a β à la place de b et ξ à la place de x, cette valeur est nulle.

Notons les éléments de la matrice Q par q_{ij}. Alors, les éléments du vecteur $c = Q^T b$ sont donnés par $c_i = \sum_{j=1}^m q_{ji} b_j$ et ceux du vecteur c'' satisfont aussi $c_i = \sum_{j=1}^m q_{ji}(b_j - \beta_j)$. Il est alors naturel de considérer les variables aléatoires

$$C_i = \sum_{j=1}^m q_{ji}(B_j - \beta_j), \qquad i = n+1, \ldots, m. \tag{7.16}$$

Le but est d'étudier la fonction de densité de $\sum_{i=n+1}^m C_i^2$.

Lemme 7.5 *Soient $B_1, \ldots, B_m$ des variables aléatoires indépendantes satisfaisant la loi normale $N(\beta_i, 1)$. Alors, les variables aléatoires $C_{n+1}, \ldots, C_m$, définies par (7.16), sont indépendantes et satisfont aussi la loi normale avec*

$$E(C_i) = 0, \qquad \mathit{Var}\,(C_i) = 1. \tag{7.17}$$

Démonstration. Pour voir que les C_i sont indépendants, calculons la probabilité $P(a_i \le C_i < b_i, i = n+1, \ldots, m)$. Notons par S l'ensemble $S = \{y \in I\!R^m \mid a_i \le y_i < b_i,\ i = n+1, \ldots, m\}$ et par C et B les vecteurs $(C_1, \ldots, C_m)^T$ et $(B_1, \ldots, B_m)^T$. Alors, on a

$$\begin{aligned} P(a_i \le C_i < b_i, i = n+1, \ldots, m) &= P(C \in S) = P(Q^T(B - \beta) \in S) \\ &= P(B - \beta \in Q(S)) \overset{\text{(a)}}{=} \iint_{Q(S)} \frac{1}{(\sqrt{2\pi})^m} \exp\Big(-\frac{1}{2}\sum_{i=1}^m y_i^2\Big)\, dy_1 \ldots dy_m \\ &\overset{\text{(b)}}{=} \iint_S \frac{1}{(\sqrt{2\pi})^m} \exp\Big(-\frac{1}{2}\sum_{i=1}^m z_i^2\Big)\, dz_1 \ldots dz_m = \prod_{i=n+1}^m \int_{a_i}^{b_i} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{z_i^2}{2}\Big)\, dz_i. \end{aligned} \tag{7.18}$$

L'identité (a) est une conséquence de l'indépendance des B_i et (b) découle de la transformation $y = Qz$, car $\det Q = 1$ et $\sum_i y_i^2 = \sum_i z_i^2$ (la matrice Q est orthogonale). En utilisant $S_i = \{y \in I\!R^m \mid a_i \le y_i < b_i\}$, on déduit de la même manière que

$$P(a_i \le C_i < b_i) = P(C \in S_i) = \ldots = \int_{a_i}^{b_i} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{z_i^2}{2}\Big)\, dz_i. \tag{7.19}$$

Une comparaison de (7.18) avec (7.19) démontre l'indépendance de $C_{n+1}, \ldots, C_m$ (voir la définition (7.6)).

Le fait que les C_i satisfont la loi normale $N(0,1)$ est une conséquence de (7.19). □

Théorème 7.6 (Pearson) *Soient $Y_1, \ldots, Y_n$ des variables aléatoires indépendantes qui obéissent à la loi normale $N(0,1)$. Alors, la fonction de densité de la variable aléatoire*

$$Y_1^2 + Y_2^2 + \ldots + Y_n^2 \tag{7.20}$$

est donnée par (voir fig. IV.7)

$$f_n(x) = \frac{1}{2^{n/2} \cdot \Gamma(n/2)} \cdot x^{n/2-1} \cdot e^{-x/2} \tag{7.21}$$

pour $x > 0$ et par $f_n(x) = 0$ pour $x \le 0$ ("loi de χ^2 à n degrés de liberté"). L'espérance de cette variable aléatoire vaut n et sa variance $2n$.

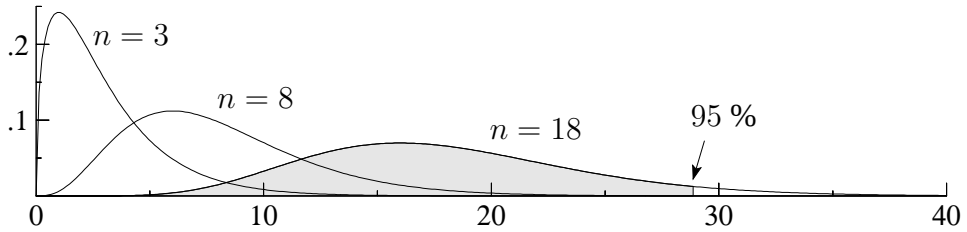


FIG. IV.7: Fonction de densité (7.21)

Démonstration. Considérons d'abord le cas $n = 1$. Pour $0 \leq a < b$, on a

$$P(a \leq Y_1^2 < b) = P(\sqrt{a} \leq Y_1 < \sqrt{b}) + P(-\sqrt{a} \geq Y_1 > -\sqrt{b})$$
$$= 2 \int_{\sqrt{a}}^{\sqrt{b}} \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}\, dx = \int_a^b \frac{1}{\sqrt{2\pi}} \cdot e^{-t/2} \cdot \frac{dt}{\sqrt{t}},$$

ce qui démontre (7.21) pour $n = 1$ car $\Gamma(1/2) = \sqrt{\pi}$.

Pour le cas général, nous procédons par récurrence. Nous utilisons le résultat du Lemme 7.2 qui affirme que la fonction de densité de $Y_1^2 + \ldots + Y_{n+1}^2$ est la convolution de celle de $Y_1^2 + \ldots + Y_n^2$ avec celle de Y_{n+1}^2. Le calcul

$$(f_n * f_1)(x) = \frac{1}{\sqrt{2} \cdot \Gamma(1/2) \cdot 2^{n/2} \cdot \Gamma(n/2)} \int_0^x (x-t)^{-1/2} e^{-(x-t)/2} t^{n/2-1} e^{-t/2}\, dt$$
$$= \frac{e^{-x/2}}{\sqrt{2} \cdot \Gamma(1/2) \cdot 2^{n/2} \cdot \Gamma(n/2)} \int_0^x (x-t)^{-1/2} t^{n/2-1}\, dt$$
$$= \frac{x^{(n+1)/2-1} e^{-x/2}}{\sqrt{2} \cdot \Gamma(1/2) \cdot 2^{n/2} \cdot \Gamma(n/2)} \int_0^1 (1-s)^{1/2} s^{n/2-1} ds \;=\; f_{n+1}(x)$$

nous permet de conclure. □

Pour les variables aléatoires C_i de (7.16), ce théorème montre que

$$\sum_{i=n+1}^{m} C_i^2 \tag{7.22}$$

est une variable aléatoire ayant comme fonction de densité $f_{m-n}(x)$ (on rappelle qu'après normalisation, on a $\sigma_i = 1$ pour les variables aléatoires B_i).

Appliquons ce résultat à l'exemple du paragraphe IV.5 (voir la formulation (7.11)). Dans ce cas, on a $\|c''\|_2^2 = 25.2$ et $m - n = 18$ degrés de liberté. La fig. IV.7 montre que cette valeur de $\|c''\|_2^2$ est suffisamment petite pour être probable.

Si l'on avait travaillé avec le modèle plus simple

$$U = a + bT \tag{7.23}$$

(à la place de (5.4)) on aurait trouvé $\|c''\|_2^2 = 526.3$ et $m - n = 19$. Cette valeur est trop grande pour être probable. La conclusion est que, pour les données du tableau IV.2, cette "loi" **est à réfuser !**