Chapitre IV

Systèmes d'Equations Linéaires

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

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

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

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

Rappelons que le système (0.2) possède une solution unique si et seulement si $\det A \neq 0$.

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 et par Lagrange (en 1781 dans ses calculs astronomiques). 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 :

Soit donné le système (0.1) et supposons que $\det A \neq 0$. 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

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

où

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

(si $a_{11} = 0$, on échange la première ligne de (0.1) avec une autre ligne pour arriver à $a_{11} \neq 0$; ceci est toujours possible car $\det A \neq 0$).

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"

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

Théorème 1.1 *Soit* $\det A \neq 0$. *L'élimination de Gauss donne*

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

où P *est une matrice de permutation et*

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

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

Remarque. Les colonnes et les lignes d'une matrice de permutation P sont des vecteurs unité. On a $\det P = \pm 1$. Un exemple est

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad P \begin{pmatrix} \text{ligne } 1 \\ \text{ligne } 2 \\ \text{ligne } 3 \end{pmatrix} = \begin{pmatrix} \text{ligne } 2 \\ \text{ligne } 1 \\ \text{ligne } 3 \end{pmatrix}.$$

Démonstration. Supposons que toutes les permutations nécessaires soient déjà faites avant que l'on commence l'élimination des variables (par abus de notation, nous écrivons A au lieu de PA dans cette démonstration). En utilisant les matrices

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

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

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

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 $\det P \cdot \det A = \det L \cdot \det R$. Comme $\det P = (-1)^\sigma$, où σ est le nombre de permutations dans l'élimination de Gauss, on obtient

$$\det A = (-1)^\sigma \cdot r_{11} \cdot \ldots \cdot r_{nn}. \tag{1.9}$$

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 $Lc = Pb$, 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.

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).

Le calcul de $b^{(1)}$ nécessite $n-1$ opérations (voir (1.3)). Par conséquent, on obtient c avec $\approx (n-1) + \ldots + 2 + 1 \approx n^2/2$ opérations. Similairement, la résolution du système (1.4) se fait en $n^2/2$ opérations.

En résumé, l'appel au sous-programme DEC nécessite $\approx n^3/3$ opérations, tandis que SOL a seulement besoin de $\approx n^2$ opérations (sur des ordinateurs sériels). A titre de comparaison, la formule habituelle pour le déterminant d'une matrice $n \times n$ contient $n! \approx n^n/e^n$ termes.

IV.2 Le choix du pivot

Dans l'élimination de Gauss, il faut au début choisir une équation (avec $a_{i1} \neq 0$; cet élément s'appelle "le pivot") à l'aide de laquelle on élimine x_1 dans les autres équations. Le choix de cette équation (choix du pivot) peut-il influencer la précision du résultat numérique, si l'on fait le calcul sur ordinateur en virgule flottante?

Exemple 2.1 (Forsythe) Considérons le système

$$\begin{aligned} 1.00 \cdot 10^{-4} \cdot x_1 \;+\; 1.00 \cdot x_2 \;&=\; 1.00 \\ 1.00 \cdot x_1 \;+\; 1.00 \cdot x_2 \;&=\; 2.00 \end{aligned} \tag{2.1}$$

avec pour solution exacte

$$x_1 = \frac{1}{0.9999} = 1.00010001\ldots, \qquad x_2 = \frac{0.9998}{0.9999} = 0.99989998\ldots. \tag{2.2}$$

Appliquons l'élimination de Gauss et simulons un calcul en virgule flottante avec 3 chiffres significatifs (en base 10).

a) Si l'on prend $a_{11} = 1.00 \cdot 10^{-4}$ comme pivot, on obtient $\ell_{21} = a_{21}/a_{11} = 1.00 \cdot 10^4$, $a_{22}^{(1)} = 1.00 - 1.00 \cdot 10^4 = -1.00 \cdot 10^4$ et $b_2^{(1)} = 2.00 - 1.00 \cdot 10^4 = -1.00 \cdot 10^4$. Par conséquent, $x_2 = b_2^{(1)}/a_{22}^{(1)} = 1.00$ (exacte!), 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.

b) Si l'on échange les deux équations de (2.1), le pivot est 1.00 et l'élimination de Gauss donne: $\ell_{21} = 1.00 \cdot 10^{-4}$, $a_{22}^{(1)} = 1.00 - 1.00 \cdot 10^{-4} = 1.00$ et $b_2^{(1)} = 1.00 - 2.00 * 1.00 \cdot 10^{-4} = 1.00$. De nouveau, on obtient $x_2 = b_2^{(1)}/a_{22}^{(1)} = 1.00$. Mais cette fois le résultat pour x_1 est

$$x_1 = (b_1 - a_{12}x_2)/a_{11} = (2.00 - 1.00 * 1.00)/1.00 = 1.00.$$

Les deux valeurs numériques (pour x_2 et aussi pour x_1) sont correctes.

Pour mieux comprendre dans quelle partie de l'élimination de Gauss on a perdu une information essentielle, considérons les sous-problèmes (addition, soustraction, multiplication, division) séparément et étudions leur "condition".

Condition d'un problème. Considérons une application $\mathcal{P} : I\!R^n \to I\!R$, le problème consistant à calculer $\mathcal{P}(x)$ pour les données $x = (x_1, \ldots, x_n)$. Il est intéressant d'étudier l'influence de perturbations dans x sur le résultat $\mathcal{P}(x)$.

Définition 2.2 *La condition κ d'un problème $\mathcal{P}$ est le plus petit nombre tel que*

$$\frac{|\widehat{x}_i - x_i|}{|x_i|} \leq \textit{eps} \qquad \Longrightarrow \qquad \frac{|\mathcal{P}(\widehat{x}) - \mathcal{P}(x)|}{|\mathcal{P}(x)|} \leq \kappa \cdot \textit{eps}. \tag{2.3}$$

On dit que le problème $\mathcal{P}$ est bien conditionné, si κ n'est pas trop grand. Sinon, il est mal conditionné.

Dans cette définition, *eps* représente un petit nombre. Si *eps* est la précision de l'ordinateur (voir le paragraphe II.5) alors, $\widehat{x}_i$ peut être interprété comme l'arrondi de x_i. Remarquons encore que la condition κ dépend des données x_i et du problème $\mathcal{P}$, mais qu'elle ne dépend pas de l'algorithme avec lequel on calcule $\mathcal{P}(x)$.

Exemple 2.3 (multiplication de deux nombres réels) Soient donnés les nombres x_1 et x_2, considérons le problème de calculer $\mathcal{P}(x_1, x_2) = x_1 \cdot x_2$. Pour les deux valeurs perturbées

$$\widehat{x}_1 = x_1(1 + \epsilon_1), \qquad \widehat{x}_2 = x_2(1 + \epsilon_2), \qquad |\epsilon_i| \leq \textit{eps} \tag{2.4}$$

on a

$$\frac{\widehat{x}_1 \cdot \widehat{x}_2 - x_1 \cdot x_2}{x_1 \cdot x_2} = (1 + \epsilon_1)(1 + \epsilon_2) - 1 = \epsilon_1 + \epsilon_2 + \epsilon_1 \cdot \epsilon_2.$$

Comme *eps* est un petit nombre, le produit $\epsilon_1 \cdot \epsilon_2$ est négligeable par rapport à $|\epsilon_1| + |\epsilon_2|$ et on obtient

$$\Big|\frac{\widehat{x}_1 \cdot \widehat{x}_2 - x_1 \cdot x_2}{x_1 \cdot x_2}\Big| \leq 2 \cdot \textit{eps}. \tag{2.5}$$

On a donc $\kappa = 2$ et ce problème est bien conditionné.

Exemple 2.4 (soustraction) Pour le problème $\mathcal{P}(x_1, x_2) = x_1 - x_2$, un calcul analogue donne

$$\Big|\frac{(\widehat{x}_1 - \widehat{x}_2) - (x_1 - x_2)}{x_1 - x_2}\Big| = \Big|\frac{x_1\epsilon_1 - x_2\epsilon_2}{x_1 - x_2}\Big| \leq \underbrace{\frac{|x_1| + |x_2|}{|x_1 - x_2|}}_{\kappa} \cdot \textit{eps}. \tag{2.6}$$

Si sign$x_1 = -$signx_2 (ceci correspond à une addition et non pas à une soustraction) on a $\kappa = 1$; le problème est bien conditionné.

Par contre, si $x_1 \approx x_2$ la condition $\kappa = (|x_1| + |x_2|)/|x_1 - x_2|$ devient très grande et on est confronté à un problème qui est extrêmement mal conditionné. Pour mieux illustrer l'effet de cette grande condition, considérons l'exemple numérique

$$x_1 = \frac{1}{51}, \qquad x_2 = \frac{1}{52} \qquad \text{pour lequel} \qquad \kappa \approx \frac{2/50}{(1/50)^2} = 100.$$

En faisant le calcul avec 3 chiffres significatifs (en base 10), on obtient $\widehat{x}_1 = 0.196 \cdot 10^{-1}$, $\widehat{x}_2 = 0.192 \cdot 10^{-1}$ et $\widehat{x}_1 - \widehat{x}_2 = 0.400 \cdot 10^{-3}$. Comme les deux premiers chiffres sont les mêmes pour $\widehat{x}_1$ et $\widehat{x}_2$, la soustraction les fait disparaître et on n'a plus qu'un chiffre qui est significatif (le résultat exact est $1/(51 \cdot 52) = 0.377 \cdot 10^{-3}$). On parle d'*extinction de chiffres*.

Explication du choix du pivot. Si ℓ_{21} est très grand (ce qui est le cas dans la situation (a) de l'exemple de Forsythe) alors,

$$\left.\begin{array}{rcrcl} a_{22}^{(1)} & = & a_{22} - \ell_{21}a_{12} & \approx & -\ell_{21}a_{12} \\ b_2^{(1)} & = & b_2 - \ell_{21}b_1 & \approx & -\ell_{21}b_1 \end{array}\right\} \qquad x_2 = \frac{b_2^{(1)}}{a_{22}^{(1)}} \approx \frac{b_1}{a_{12}}. \tag{2.7}$$

Cette valeur, obtenue pour x_2, est en général correcte. Mais le calcul de x_1

$$x_1 = (b_1 - a_{12}x_2)/a_{11}$$

nécessite une soustraction qui est très mal conditionnée car $a_{12}x_2 \approx b_1$ et, à cause de l'extinction de chiffres, on perd de la précision.

La conclusion de cette étude est qu'il faut éviter des ℓ_{ij} trop grands.

Recherche partielle de pivot. L'idée est de ne pas se contenter d'un pivot qui soit différent de zéro ($a_{11} \neq 0$), mais d'échanger les équations de (0.1) afin que a_{11} soit le plus grand élément (en valeur absolue) de la première colonne de A. De cette manière, on a toujours $|\ell_{i1}| \leq 1$. La même stratégie est répétée pour les sous-systèmes apparaissant dans l'élimination de Gauss.

Expérience numérique. 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

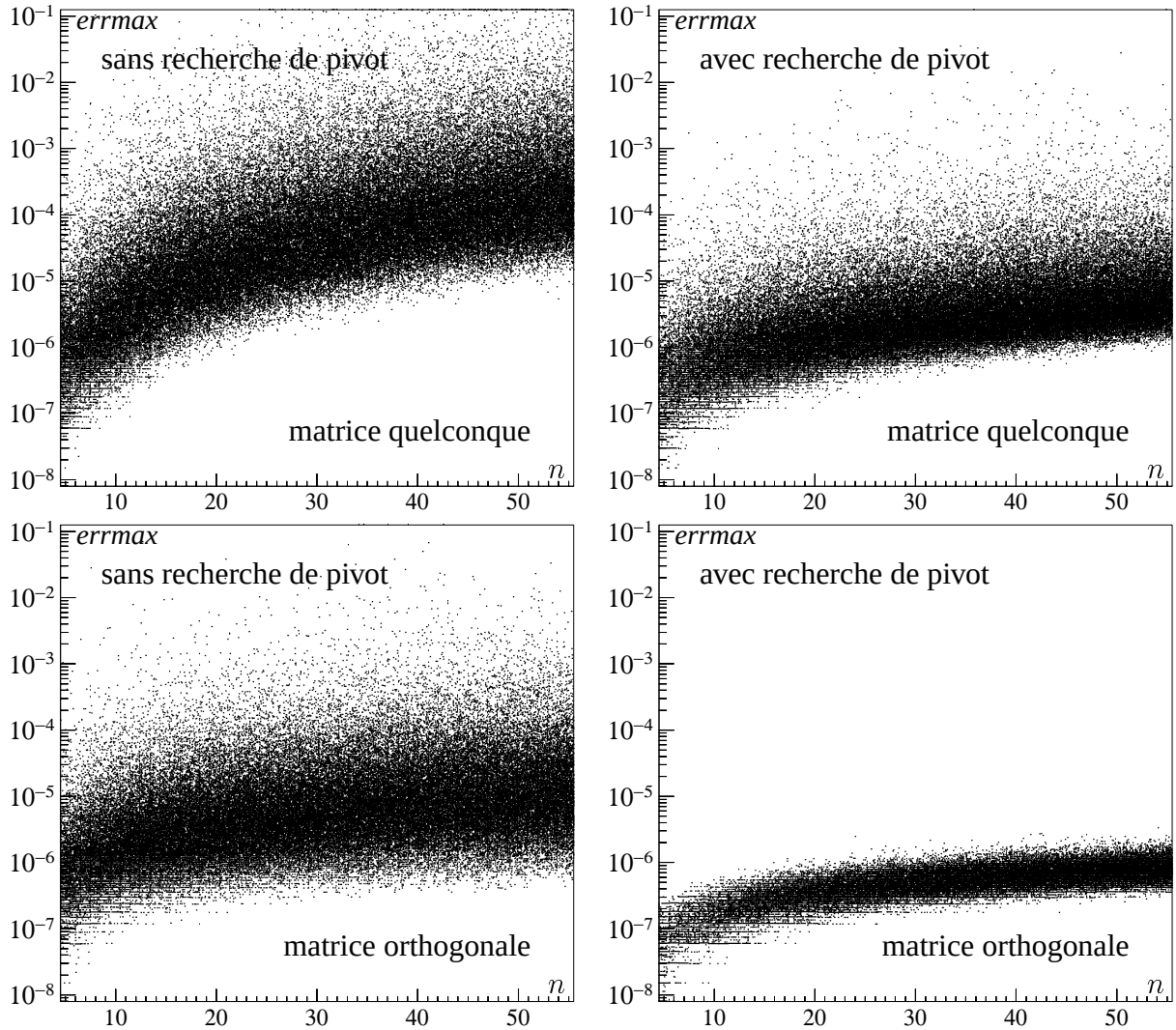


FIG. IV.1: Erreurs pour 1 million de systèmes linéaires de dimensions 5×5 à 55×55

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.1. 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. Cette matrice est calculée en double précision, le reste de l'expérience continue comme auparavant (voir les résultats au bas de la figure IV.1). Cette fois-ci il n'y a pas d'exception dans la bonne performance de l'algorithme de Gauss avec recherche de pivot. Nous allons démontrer ces observations dans les paragraphes suivantes.

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

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

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

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

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 $\|A \cdot B\| \leq \|A\| \cdot \|B\|$.

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

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

où ϵ_A et ϵ_b spécifient la précision des données (par exemple $\epsilon_A \leq$ *eps*, $\epsilon_b \leq$ *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

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

Notre premier résultat donne une estimation de $\|\widehat{x} - x\|$, en supposant que (3.7) soit vrai.

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*

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

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), $\|\widehat{x}\| \leq \|x\| + \|\widehat{x} - x\|$ et $\|b\| = \|Ax\| \leq \|A\| \cdot \|x\|$. Nous obtenons ainsi

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

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) $\kappa(A) \geq 1$ *pour toute A,*
b) $\kappa(\alpha A) = \kappa(A)$ *pour $\alpha \neq 0$,*
c) $\kappa(A) = \max\limits_{\|y\|=1} \|Ay\| \Big/ \min\limits_{\|z\|=1} \|Az\|$.

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

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

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

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.

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).

TAB. IV.1: Condition 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	$5.6 \cdot 10^2$	$3.7 \cdot 10^4$	$2.4 \cdot 10^6$	$1.6 \cdot 10^8$	$1.0 \cdot 10^{10}$

Concernant l'interpolation avec des fonctions splines, nous avons rencontré la matrice (voir le paragraphe II.10, cas équidistant)

$$A = \frac{1}{h} \begin{pmatrix} 4 & 1 & & \\ 1 & 4 & 1 & \\ & 1 & \ddots & \ddots \\ & & \ddots & \ddots \end{pmatrix} \quad \Big\} \; n \tag{3.10}$$

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 La stabilité d'un algorithme

Le but de ce paragraphe est d'étudier l'influence des erreurs d'arrondi sur le résultat pour l'élimination de Gauss. Commençons par la définition de la stabilité d'un algorithme et avec quelques exemples simples.

Un *algorithme* pour résoudre le problème $\mathcal{P}(x)$ est une suite d'opérations élémentaires $f_1, \ldots, f_n$ (addition, soustraction, multiplication, division, évaluation d'une racine, d'une fonction élémentaire, ...) telle que

$$\mathcal{P}(x) = f_n(f_{n-1}(\ldots f_2(f_1(x)) \ldots)). \tag{4.1}$$

En général, il existe beaucoup d'algorithmes différents pour résoudre le même problème $\mathcal{P}(x)$.

L'amplification de l'erreur, en faisant l'opération f_i, est décrite par la condition $\kappa(f_i)$ (voir la définition dans le paragraphe IV.2). L'estimation

$$\kappa(\mathcal{P}) \leq \kappa(f_1) \cdot \kappa(f_2) \cdot \ldots \cdot \kappa(f_n). \tag{4.2}$$

est une conséquence simple de la définition de la condition d'un problème.

Définition 4.1 *Un algorithme est numériquement stable (au sens de "forward analysis") si*

$$\kappa(f_1) \cdot \kappa(f_2) \cdot \ldots \cdot \kappa(f_n) \leq \mathit{Const} \cdot \kappa(\mathcal{P}) \tag{4.3}$$

où Const n'est pas trop grand (par exemple, Const $= \mathcal{O}(n)$).

La formule (4.3) exprime le fait que l'influence des erreurs d'arrondi durant le calcul de $\mathcal{P}(x)$ n'est pas beaucoup plus grande que l'influence d'erreurs dans les données (qui sont inévitables).

Exemple 4.2 Soit $x = 10^4$ et considérons le problème de calculer $1/(x(1+x))$. Examinons les deux algorithmes suivants :

a) $$x \quad \begin{array}{c} \nearrow \\ \searrow \end{array} \quad \begin{array}{c} x \\ x+1 \end{array} \quad \begin{array}{c} \searrow \\ \nearrow \end{array} \quad x(x+1) \quad \longrightarrow \quad \frac{1}{x(x+1)}.$$

Toutes ces opérations sont très bien conditionnées (voir le paragraphe IV.3). Ainsi, cet algorithme est numériquement stable.

b) $$x \quad \begin{array}{c} \nearrow \\ \searrow \end{array} \quad \begin{array}{c} 1/x \\ x+1 \end{array} \quad \longrightarrow \quad 1/(x+1) \quad \begin{array}{c} \searrow \\ \nearrow \end{array} \quad \frac{1}{x} - \frac{1}{x+1} = \frac{1}{x(x+1)}.$$

Dans cet algorithme, seules les trois premières opérations sont bien conditionnées. La soustraction, à la fin, est très mal conditionnée car $1/x \approx 1/(x+1)$. Ainsi, cet algorithme est numériquement *instable*.

La vérification, si un algorithme (non-trivial) est stable (au sens de "forward analysis"), est souvent très complexe et difficile. Pour cette raison, Wilkinson (1961, J. Ass. Comp. Mach. 8) a introduit une autre définition de la stabilité d'un algorithme.

Définition 4.3 *Un algorithme pour résoudre le problème $\mathcal{P}(x)$ est numériquement stable (au sens de "backward analysis") si le résultat numérique $\widehat{y}$ peut être interprété comme un résultat exact pour des données perturbées $\widehat{x}$ (c.-à-d., $\widehat{y} = \mathcal{P}(\widehat{x})$) et si*

$$\frac{|\widehat{x}_i - x_i|}{|x_i|} \leq \textit{Const} \cdot \textit{eps} \tag{4.4}$$

où Const n'est pas trop grand et eps est la précision de l'ordinateur.

Remarque. Pour l'étude de cette stabilité, il ne faut pas connaître la condition du problème.

Exemple 4.4 Considérons le problème de calculer le produit scalaire $x_1 \cdot x_2 + x_3 \cdot x_4$. On utilise l'algorithme

$$(x_1, x_2, x_3, x_4) \quad \begin{array}{c} \nearrow \\ \searrow \end{array} \quad \begin{array}{c} x_1 \cdot x_2 \\ x_3 \cdot x_4 \end{array} \quad \begin{array}{c} \searrow \\ \nearrow \end{array} \quad x_1 \cdot x_2 + x_3 \cdot x_4. \tag{4.5}$$

Le résultat numérique (sous l'influence des erreurs d'arrondi) est

$$\Big(x_1(1+\epsilon_1) \cdot x_2(1+\epsilon_2)(1+\eta_1) + x_3(1+\epsilon_3) \cdot x_4(1+\epsilon_4)(1+\eta_2)\Big)(1+\eta_3)$$

où $|\epsilon_i|, |\eta_j| \leq \textit{eps}$. Ce résultat est égal à $\widehat{x}_1 \cdot \widehat{x}_2 + \widehat{x}_3 \cdot \widehat{x}_4$ si l'on pose

$$\begin{aligned} \widehat{x}_1 &= x_1(1+\epsilon_1)(1+\eta_1), \qquad & \widehat{x}_3 &= x_3(1+\epsilon_3)(1+\eta_2), \\ \widehat{x}_2 &= x_2(1+\epsilon_2)(1+\eta_3), \qquad & \widehat{x}_4 &= x_4(1+\epsilon_4)(1+\eta_3). \end{aligned}$$

Ainsi, (4.4) est vérifié pour $\textit{Const} = 2$ (on néglige les produits $\epsilon_i \cdot \eta_j$). En conséquence, l'algorithme (4.5) est toujours numériquement stable (au sens de "backward analysis").

Cet exemple montre bien qu'un algorithme peut être stable, même si le problème est mal conditionné. Ainsi, il faut bien distinguer les notions "stabilité numérique" et "condition d'un problème".

La stabilité de l'élimination de Gauss. Soit donnée une matrice A ($\det A \neq 0$) ayant la décomposition $A = LR$ (on suppose que les permutations nécessaires sont déjà effectuées). En appliquant l'élimination de Gauss, nous obtenons deux matrices $\widehat{L}$ et $\widehat{R}$, qui représentent la décomposition exacte de la matrice $\widehat{A} := \widehat{L} \cdot \widehat{R}$. Pour montrer la stabilité numérique (au sens de "backward analysis") de l'élimination de Gauss, on a besoin de trouver une estimation de la forme $|\widehat{a}_{ij} - a_{ij}| \leq |a_{ij}| \cdot \textit{Const} \cdot \textit{eps}$. En tous cas, il faut estimer la différence $\widehat{a}_{ij} - a_{ij}$.

Théorème 4.5 (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,*

$$|\widehat{a}_{ij} - a_{ij}| \leq 2 \cdot a \cdot \min(i-1, j) \cdot \mathit{eps} \tag{4.6}$$

où $a = \max_{i,j,k} |a_{ij}^{(k)}|$.

Démonstration. Lors de la $k^{\text{ème}}$ étape de l'élimination de Gauss, on calcule $\widehat{a}_{ij}^{(k)}$ à partir de $\widehat{a}_{ij}^{(k-1)}$. Si l'on prend en considération les erreurs d'arrondi, on obtient

$$\begin{aligned} \widehat{a}_{ij}^{(k)} &= (\widehat{a}_{ij}^{(k-1)} - \widehat{\ell}_{ik} \cdot \widehat{a}_{kj}^{(k-1)} \cdot (1 + \epsilon_{ijk}))(1 + \eta_{ijk}) \\ &= \widehat{a}_{ij}^{(k-1)} - \widehat{\ell}_{ik} \cdot \widehat{a}_{kj}^{(k-1)} + \mu_{ijk} \end{aligned} \tag{4.7}$$

où (en négligeant les termes $\mathcal{O}(\mathit{eps}^2)$)

$$|\mu_{ijk}| \leq |\widehat{a}_{ij}^{(k)}| \cdot |\eta_{ijk}| + |\widehat{\ell}_{ik}| \cdot |\widehat{a}_{kj}^{(k-1)}| \cdot |\epsilon_{ijk}| \leq 2 \cdot a \cdot \mathit{eps}. \tag{4.8}$$

Par définition de $\widehat{A}$, on a $\widehat{a}_{ij} = \sum_{k=1}^{\min(i,j)} \widehat{\ell}_{ik} \cdot \widehat{r}_{kj} = \sum_{k=1}^{\min(i,j)} \widehat{\ell}_{ik} \cdot \widehat{a}_{kj}^{(k-1)}$ et, en utilisant la formule (4.7), on obtient pour $i > j$

$$\widehat{a}_{ij} = \sum_{k=1}^{j} (\widehat{a}_{ij}^{(k-1)} - \widehat{a}_{ij}^{(k)} + \mu_{ijk}) = a_{ij} + \sum_{k=1}^{j} \mu_{ijk} \tag{4.9a}$$

car $\widehat{a}_{ij}^{(j)} = 0$ dans cette situation. Pour $i \leq j$, on a

$$\widehat{a}_{ij} = \sum_{k=1}^{i-1} (\widehat{a}_{ij}^{(k-1)} - \widehat{a}_{ij}^{(k)} + \mu_{ijk}) + \widehat{\ell}_{ii} \cdot \widehat{a}_{ij}^{(i-1)} = a_{ij} + \sum_{k=1}^{i-1} \mu_{ijk} \tag{4.9b}$$

car $\widehat{\ell}_{ii} = 1$. Les formules (4.9) ensemble avec l'estimation (4.8) démontrent l'estimation (4.6). □

Conséquence. L'élimination de Gauss est stable (au sens de "backward analysis") si le quotient

$$\max_{i,j,k} |a_{ij}^{(k)}| \;\Big/\; \max_{i,j} |a_{ij}| \tag{4.10}$$

n'est pas trop grand. En fait, il existe des matrices pathologiques pour lesquelles ce quotient atteint 2^{n-1}, où n est la dimension de la matrice A (voir exercice 12). Mais heureusement cette constante est en général beaucoup plus petite. Pour illustrer ceci, en suivant une idée de Trefethen & Schreiber (1990, SIAM J. Matrix Anal. Appl. 11) nous avons pris un grand nombre de matrices de dimensions allant de 2 à 24 dont les éléments sont des nombres aléatoires dans $[-1, 1]$. Nous avons dessiné dans la figure IV.2 le quotient (4.10) en fonction de la dimension de la matrice (chaque point représente la moyenne de 30 échantillons). En échelle doublement logarithmique, le résultat semble mystérieusement devenir une droite.

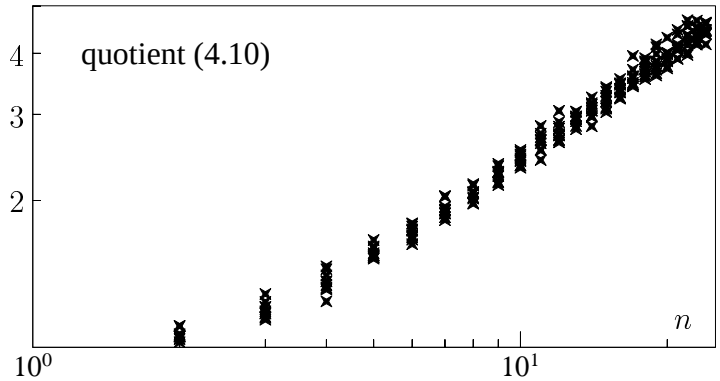


FIG. IV.2: Stabilité numérique de l'élimination de Gauss

IV.5 L'algorithme de Cholesky

Etudions l'élimination de Gauss pour le cas important où

A est *symétrique* ($A^T = A$) et

A est *définie positive* ($x^T Ax > 0$ pour $x \neq 0$).

Le théorème suivant montre que, dans cette situation particulière, il n'est pas nécessaire d'effectuer une recherche de pivot.

Théorème 5.1 *Soit A une matrice symétrique et définie positive.*
a) L'élimination de Gauss est faisable sans recherche de pivot.
b) La décomposition $A = LR$ satisfait

$$R = DL^T \qquad \textit{avec} \qquad D = \operatorname{diag}(r_{11}, \ldots, r_{nn}). \tag{5.1}$$

Démonstration. a) On a $a_{11} = e_1^T A e_1 > 0$ (avec $e_1 = (1, 0, \ldots, 0)^T$) car la matrice A est définie positive. Alors, on peut choisir a_{11} comme pivot dans la première étape de l'élimination de Gauss. Ceci donne

$$A = \begin{pmatrix} a_{11} & a^T \\ a & C \end{pmatrix} \qquad \longrightarrow \qquad A^{(1)} = \begin{pmatrix} a_{11} & a^T \\ 0 & C^{(1)} \end{pmatrix} \tag{5.2}$$

où $c_{ij}^{(1)} = a_{ij} - \dfrac{a_{i1}}{a_{11}} \cdot a_{1j}$ pour $i, j = 2, \ldots, n$, ce qui est équivalent à

$$C^{(1)} = C - \frac{1}{a_{11}} \cdot a \cdot a^T. \tag{5.3}$$

La matrice $C^{(1)}$ est symétrique (trivial). Montrons qu'elle est aussi définie positive. Pour ceci, nous prenons un $y \in I\!R^{n-1}$, $y \neq 0$. Il faut montrer que $y^T C^{(1)} y > 0$. La partition de (5.2) et le fait que A soit définie positive impliquent

$$(x_1, y^T) \begin{pmatrix} a_{11} & a^T \\ a & C \end{pmatrix} \begin{pmatrix} x_1 \\ y \end{pmatrix} = a_{11} x_1^2 + 2x_1 \cdot y^T a + y^T C y > 0. \tag{5.4}$$

En posant $x_1 = -y^T a / a_{11}$ dans (5.4), on obtient de (5.3) que

$$y^T C^{(1)} y = y^T C y - \frac{1}{a_{11}} (y^T a)^2 > 0.$$

Par récurrence, on voit que la deuxième et aussi les autres étapes de l'élimination de Gauss sont faisables sans recherche de pivot.

b) La formule (5.1) est une conséquence de l'unicité de l'élimination de Gauss pour des matrices inversibles. En effet, on peut écrire $R = D\widehat{L}^T$ et on obtient $A = A^T = R^T L^T = \widehat{L}(DL^T)$, d'où $\widehat{L} = L$.

Pour montrer l'unicité de l'élimination de Gauss, supposons $A = LR = \widehat{L}\widehat{R}$ et considérons l'identité $\widehat{L}^{-1} L = \widehat{R} R^{-1}$. Le produit de deux matrices triangulaires inférieures reste une matrice triangulaire inférieure; de même pour les matrices triangulaires supérieures. Comme les éléments de la diagonale de $\widehat{L}^{-1} L$ sont tous égaux à 1, on a

$$\widehat{L}^{-1} L = \widehat{R} R^{-1} = I, \tag{5.5}$$

ce qui implique l'unicité de l'élimination de Gauss. □

La décomposition

$$A = LDL^T \tag{5.6}$$

s'appelle *décomposition rationnelle de Cholesky*[1]. Comme $r_{ii} > 0$ (A est définie positive), on peut considérer la racine $D^{1/2} = \mathrm{diag}(\sqrt{r_{11}}, \ldots, \sqrt{r_{nn}})$, et la décomposition (5.6) devient $A = (LD^{1/2})(D^{1/2}L^T) = (LD^{1/2})(LD^{1/2})^T$. Par abus de notation, en écrivant L pour $LD^{1/2}$, nous obtenons la *décomposition de Cholesky*

$$A = LL^T \qquad \text{où} \qquad L = \begin{pmatrix} \ell_{11} & & 0 \\ \vdots & \ddots & \\ \ell_{n1} & \ldots & \ell_{nn} \end{pmatrix}. \tag{5.7}$$

Une comparaison des coefficients dans l'identité $A = LL^T$ donne pour

$$\begin{aligned} i = k: &\qquad a_{kk} = \ell_{k1}^2 + \ell_{k2}^2 + \ldots + \ell_{kk}^2 \\ i > k: &\qquad a_{ik} = \ell_{i1}\ell_{k1} + \ldots + \ell_{i,k-1}\ell_{k,k-1} + \ell_{ik}\ell_{kk} \end{aligned}$$

et on en déduit l'algorithme suivant:

Algorithme de Cholesky.

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

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

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

Ceci correspond à la moitié du coût de la décomposition LR.

Pour résoudre le système $Ax = b$, on calcule d'abord la décomposition de Cholesky (5.7). Puis, on résoud successivement les deux systèmes $Lc = b$ et $L^Tx = c$, dont les matrices sont triangulaires.

Comme pour l'élimination de Gauss, on peut étudier la stabilité de l'algorithme de Cholesky.

Théorème 5.2 *Soit A une matrice symétrique et définie positive. Notons $\widehat{A} = \widehat{L}\cdot\widehat{L}^T$, où $\widehat{L}$ est la matrice triangulaire obtenue par l'algorithme de Cholesky. Alors,*

$$|\widehat{a}_{ij} - a_{ij}| \le a_0 \cdot \min(i,j) \cdot \mathit{eps} \tag{5.8}$$

où $a_0 = \max_{i,j}|a_{ij}| = \max_i |a_{ii}|$. □

Ce résultat démontre que l'algorithme de Cholesky est toujours numériquement stable. Il n'est donc pas nécessaire de faire une recherche de pivot, si A est symétrique et définie positive.

[1]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).

IV.6 Systèmes surdéterminés – méthode des moindres carrés

Considérons un système d'équations linéaires

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \vdots \qquad \vdots \qquad\qquad \vdots \quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m \end{array} \tag{6.1}$$

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 (6.1) ne possède, en général, pas de solution. L'idée est de chercher un vecteur x tel que

$$\|Ax - b\|_2 \to \min \tag{6.2}$$

pour la norme euclidienne. Une justification probabiliste de cette condition sera donnée dans le paragraphe IV.8. Le nom "méthode des moindres carrés" indique le choix de la norme dans (6.2) (la somme des carrés des erreurs doit être minimale).

Théorème 6.1 *Soit A une matrice $m \times n$ (avec $m \geq n$) et soit $b \in \mathbb{R}^m$. Le vecteur x est solution de (6.2) si et seulement si*

$$A^T Ax = A^T b. \tag{6.3}$$

Les équations du système (6.3) 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 (6.2) 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 (6.3).

Exemple 6.2 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 figure IV.3). Les données du tableau IV.2 sont prises du livre de P.R. Bevington[2].

On suppose que cette dépendance obéit à la loi

$$U = a + bT + cT^2 \tag{6.4}$$

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

$$U_i = a + bT_i + cT_i^2, \qquad i = 1, \ldots, 21. \tag{6.5}$$

En résolvant les équations normales (6.3) pour ce problème, on obtient $a = -0.886$, $b = 0.0352$ et $c = 0.598 \cdot 10^{-4}$. Avec ces paramètres, la fonction (6.4) est dessinée dans la figure IV.3. On observe une très bonne concordance avec les données.

Remarque. Les équations normales (6.3) possèdent toujours au moins une solution (la projection sur E existe toujours). La matrice $A^T A$ est symétrique et non-négative ($x^T A^T Ax = \|Ax\|^2 \geq 0$).

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

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

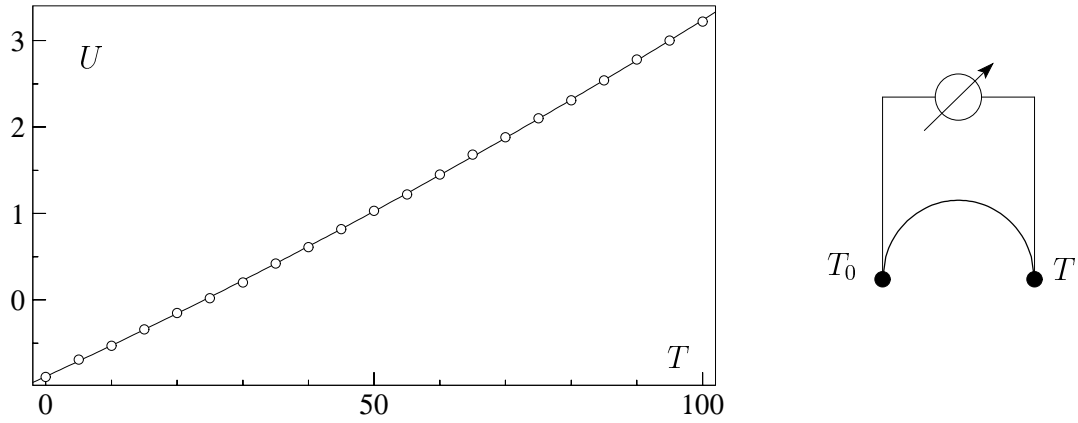


FIG. IV.3: Tension en fonction de la température et schéma de l'expérience

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 (6.3). Mais, souvent, il est préférable de calculer la solution directement de (6.2) sans passer par les équations normales (6.3).

IV.7 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 (6.2), nous cherchons une matrice orthogonale Q telle que

$$Q^T(Ax - b) = Rx - c = \begin{pmatrix} R' \\ 0 \end{pmatrix} x - \begin{pmatrix} c' \\ c'' \end{pmatrix} \tag{7.1}$$

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

$$\|Ax - b\|_2^2 = \|Q^T(Ax - b)\|_2^2 = \|Rx - c\|_2^2 = \|R'x - c'\|_2^2 + \|c''\|_2^2. \tag{7.2}$$

On obtient alors la solution de (6.2) en résolvant le système

$$R'x = c'. \tag{7.3}$$

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

$$A = QR. \tag{7.4}$$

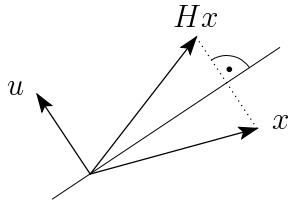
Cette factorisation s'appelle la "décomposition QR" de la matrice A. Pour arriver à ce but, on peut se servir des rotations de Givens (voir exercice 11 du chapitre V) ou des réflexions de Householder.

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

$$H = I - 2uu^T \qquad \text{où} \qquad u^Tu = 1 \tag{7.5}$$

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

$$H_1A = \begin{pmatrix} \alpha_1 & \times & \cdots & \times \\ 0 & \times & \cdots & \times \\ \vdots & \vdots & & \vdots \\ 0 & \times & \cdots & \times \end{pmatrix}. \tag{7.6}$$

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,

$$u_1 = C \cdot v_1 \qquad \text{où} \qquad v_1 = A_1 - \alpha_1e_1 \tag{7.7}$$

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

$$\alpha_1 = -\text{sign}(a_{11}) \cdot \|A_1\|_2 \tag{7.8}$$

pour éviter une soustraction mal conditionnée dans le calcul de $v_1 = A_1 - \alpha_1e_1$.

Calcul de H_1A. Notons par A_j et $(H_1A)_j$ les $j^{\text{èmes}}$ colonnes de A et H_1A respectivement. Alors, on a

$$(H_1A)_j = A_j - 2u_1u_1^TA_j = A_j - \beta \cdot v_1^TA_j \cdot v_1 \qquad \text{où} \qquad \beta = \frac{2}{v_1^Tv_1}. \tag{7.9}$$

Le facteur β peut être calculé à l'aide de

$$\beta^{-1} = \frac{v_1^Tv_1}{2} = \frac{1}{2}\big(A_1^TA_1 - 2\alpha_1a_{11} + \alpha_1^2\big) = -\alpha_1(a_{11} - \alpha_1). \tag{7.10}$$

Dans une *deuxième étape*, on applique la procédure précédente à la sous-matrice de dimension $(m-1)\times(n-1)$ de (7.6). 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 (7.6) par la matrice $H_2 = I - 2u_2u_2^T$ donne

$$H_2H_1A = H_2\begin{pmatrix} \alpha_1 & \times & \cdots & \times \\ 0 & & & \\ \vdots & & C & \\ 0 & & & \end{pmatrix} = \begin{pmatrix} \alpha_1 & \times & \cdots & \times \\ 0 & & & \\ \vdots & & \bar{H}_2C & \\ 0 & & & \end{pmatrix} = \begin{pmatrix} \alpha_1 & \times & \times & \cdots & \times \\ 0 & \alpha_2 & \times & \cdots & \times \\ 0 & 0 & \times & \cdots & \times \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \times & \cdots & \times \end{pmatrix}.$$

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

$$\underbrace{H_n \cdot \ldots \cdot H_2H_1}_{Q^T} A = R = \begin{pmatrix} R' \\ 0 \end{pmatrix}.$$

Ceci donne la décomposition (7.4) avec $Q^T = H_n \cdot \ldots \cdot H_2H_1$.

Coût de la décomposition QR. La première étape exige le calcul de α_1 par la formule (7.8) ($\approx m$ opérations), le calcul de $2/v_1^Tv_1$ par la formule (7.10) (travail négligeable) et le calcul de $(H_1A)_j$ pour $j = 2, \ldots, n$ par la formule (7.9) ($\approx (n-1)\cdot 2\cdot m$ 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^TA et $\approx n^3/6$ opérations pour la décomposition de Cholesky de A^TA), 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^Tb et la solution du système triangulaire $R'x = c'$ (voir (7.3)). Le calcul de $Q^Tb = H_n \cdot \ldots \cdot H_2H_1b$ se fait avec une formule analogue à (7.9).

Exemple 7.1 Si les colonnes de A sont "presque" linéairement dépendantes, la résolution du problème (6.2) 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 $\epsilon^2 < \textit{eps}$. Avec un calcul exact, on obtient

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

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 (7.3) donne une bonne approximation de la solution exacte.

IV.8 Etude de l'erreur de la méthode des moindres carrés

Comme c'est le cas dans l'exemple du paragraphe IV.6, nous cherchons les paramètres $x_1, \ldots, x_n$ d'une loi

$$\sum_{j=1}^{n} c_j(t)\, x_j = b \tag{8.1}$$

qui relie les variables t et b (les fonctions $c_j(t)$ sont données, p. ex. $c_j(t) = t^{j-1}$). Supposons que pour plusieurs valeurs de t (disons $t_1, \ldots, t_m$, $m \gg n$) l'on puisse mesurer les quantités $b_1, \ldots, b_m$. On obtient ainsi le système surdéterminé

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \qquad i = 1, \ldots, m \tag{8.2}$$

où $a_{ij} = c_j(t_i)$. 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 $f : \mathbb{R} \to \mathbb{R}$, c.-à-d., la probabilité de l'événement que la valeur de X se trouve dans l'intervalle $[a, b)$ est donnée par

$$P(a \le X < b) = \int_a^b f(x)\, dx \tag{8.3}$$

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

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

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x f(x)\, dx, \tag{8.4}$$

et *variance* la valeur

$$\sigma_X^2 = \mathit{Var}\,(X) = \int_{-\infty}^{\infty} (x-\mu_X)^2 f(x)\, dx = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu_X^2. \tag{8.5}$$

Exemple 8.1 Si une variable aléatoire satisfait (8.3) avec (voir la figure IV.4)

$$f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\exp\Big(-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big) \tag{8.6}$$

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.

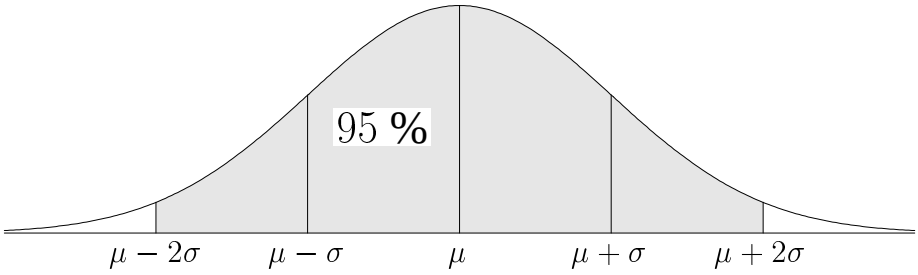


FIG. IV.4: Fonction de densité pour la loi normale

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

$$P(a_i \le X_i < b_i,\ i=1,\dots,n) = \prod_{i=1}^{n} P(a_i \le X_i < b_i). \tag{8.7}$$

Lemme 8.2 *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*

$$\frac{1}{|\alpha|}f\Big(\frac{x-\beta}{\alpha}\Big) \qquad \textit{et} \qquad (f*g)(z) = \int_{-\infty}^{\infty} f(z-y)g(y)\,dy. \tag{8.8}$$

Leur espérance mathématique est

$$E(\alpha X+\beta) = \alpha E(X)+\beta, \qquad E(X+Y) = E(X)+E(Y) \tag{8.9}$$

et leur variance satisfait

$$\textit{Var}\,(\alpha X+\beta) = \alpha^2\textit{Var}\,(X), \qquad \textit{Var}\,(X+Y) = \textit{Var}\,(X)+\textit{Var}\,(Y). \tag{8.10}$$

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 (8.9) et (8.10) 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 16).

Revenons maintenant au problème (8.2). 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 (8.2) 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 $\beta = (\beta_1,\ldots,\beta_m)^T$.

Motivation de la méthode des moindres carrés. Par l'hypothèse H1, la probabilité que B_i soit dans l'intervalle $[b_i, b_i+db_i)$ avec db_i (infinitésimalement) 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 (8.7) implique que

$$\begin{aligned} P(b_i \le B_i < b_i+db_i,\ i=1,\ldots,m) &\approx \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\cdot\sigma_i}\cdot\exp\Big(-\frac{1}{2}\Big(\frac{b_i-\beta_i}{\sigma_i}\Big)^2\Big)\cdot db_i \qquad (8.11)\\ &= C\cdot\exp\Big(-\frac{1}{2}\sum_{i=1}^{m}\Big(\frac{b_i-\beta_i}{\sigma_i}\Big)^2\Big) = C\cdot\exp\Big(-\frac{1}{2}\sum_{i=1}^{m}\Big(\frac{b_i-\sum_{j=1}^{n}a_{ij}\xi_j}{\sigma_i}\Big)^2\Big). \end{aligned}$$

Selon une idée de Gauss (1812), la "meilleure" réponse x_i pour les ξ_i (inconnus) est celle pour laquelle la probabilité (8.11) est maximale ("maximum likelihood"). Alors, on calcule $x_1,\ldots,x_n$ de façon à ce que

$$\sum_{i=1}^{m}\Big(\frac{b_i}{\sigma_i} - \sum_{j=1}^{n}\frac{a_{ij}}{\sigma_i}\cdot x_j\Big)^2 \to \min. \qquad (8.12)$$

Si l'on remplace b_i/σ_i par b_i et a_{ij}/σ_i par a_{ij}, la condition (8.12) est équivalente à (6.2). 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 (8.12) 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

$$X_i = \sum_{j=1}^{m}\alpha_{ij}B_j \qquad \text{ou} \qquad X_i-\xi_i = \sum_{j=1}^{m}\alpha_{ij}(B_j-\beta_j). \qquad (8.13)$$

Théorème 8.3 *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 (8.13), satisfait*

$$E(X_i) = \xi_i \qquad \textit{et} \qquad \textit{Var}\,(X_i) = \epsilon_{ii} \tag{8.14}$$

où ϵ_{ii} est le $i^{ème}$ 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 (8.9) donne $E(X_i) = \xi_i$. Pour calculer la variance de X_i, nous utilisons le fait que *Var* $(B_i) = 1$ et la formule (8.10). 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 8.4 Pour l'expérience sur la thermo-électricité (voir le paragraphe IV.6), 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

$$(A^TA)^{-1} = \begin{pmatrix} 0.356 \cdot 10^{-4} & -0.139 \cdot 10^{-5} & 0.113 \cdot 10^{-7} \\ -0.139 \cdot 10^{-5} & 0.765 \cdot 10^{-7} & -0.713 \cdot 10^{-9} \\ 0.113 \cdot 10^{-7} & -0.713 \cdot 10^{-9} & 0.713 \cdot 10^{-11} \end{pmatrix} \tag{8.15}$$

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 (t_i, b_i) sont compatibles avec la loi (8.1). Ceci revient à justifier l'hypothèse H2.

En utilisant la décomposition QR de la matrice A, le problème surdéterminé $Ax = b$ se transforme en (voir (7.1))

$$\begin{pmatrix} R' \\ 0 \end{pmatrix} x = \begin{pmatrix} c' \\ c'' \end{pmatrix} \qquad \text{où} \qquad \begin{pmatrix} c' \\ c'' \end{pmatrix} = Q^Tb. \tag{8.16}$$

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 $c = Q^Tb$ 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

$$C_i = \sum_{j=1}^{m} q_{ji}(B_j - \beta_j), \qquad i = n+1, \ldots, m. \tag{8.17}$$

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

Lemme 8.5 *Soient $B_1, \dots, B_m$ des variables aléatoires indépendantes satisfaisant la loi normale $N(\beta_i, 1)$. Alors, les variables aléatoires $C_{n+1}, \dots, C_m$, définies par (8.17), sont indépendantes et satisfont aussi la loi normale avec*

$$E(C_i) = 0, \qquad \textit{Var}\,(C_i) = 1. \tag{8.18}$$

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

$$\begin{aligned} P(a_i \le C_i < b_i, i = n+1, \dots, m) &= P(C \in S) = P(Q^T(B-\beta) \in S) \\ &= P(B - \beta \in Q(S)) \overset{\text{(a)}}{=} \iint_{Q(S)} \frac{1}{(\sqrt{2\pi})^m} \exp\Big(-\frac{1}{2}\sum_{i=1}^m y_i^2\Big)\, dy_1 \dots dy_m \\ &\overset{\text{(b)}}{=} \iint_S \frac{1}{(\sqrt{2\pi})^m} \exp\Big(-\frac{1}{2}\sum_{i=1}^m z_i^2\Big)\, dz_1 \dots dz_m = \prod_{i=n+1}^m \int_{a_i}^{b_i} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{z_i^2}{2}\Big)\, dz_i. \end{aligned} \tag{8.19}$$

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

$$P(a_i \le C_i < b_i) = P(C \in S_i) = \dots = \int_{a_i}^{b_i} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{z_i^2}{2}\Big)\, dz_i. \tag{8.20}$$

Une comparaison de (8.19) avec (8.20) démontre l'indépendance de $C_{n+1}, \dots, C_m$ (voir la définition (8.7)). Le fait que les C_i satisfont la loi normale $N(0,1)$ est une conséquence de (8.20). □

Théorème 8.6 (Pearson) *Soient $Y_1, \dots, Y_n$ des variables aléatoires indépendantes qui obéissent à la loi normale $N(0,1)$. Alors, la fonction de densité de la variable aléatoire*

$$Y_1^2 + Y_2^2 + \dots + Y_n^2 \tag{8.21}$$

est donnée par (voir figure IV.5)

$$f_n(x) = \frac{1}{2^{n/2} \cdot \Gamma(n/2)} \cdot x^{n/2-1} \cdot e^{-x/2} \tag{8.22}$$

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 $2n$.

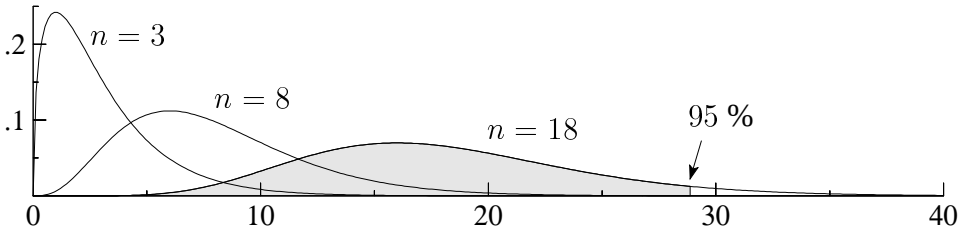


FIG. IV.5: Fonction de densité (8.22)

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

$$\begin{aligned} P(a \le Y_1^2 < b) &= P(\sqrt{a} \le Y_1 < \sqrt{b}) + P(-\sqrt{a} \ge Y_1 > -\sqrt{b}) \\ &= 2\int_{\sqrt{a}}^{\sqrt{b}} \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}\, dx = \int_a^b \frac{1}{\sqrt{2\pi}} \cdot e^{-t/2} \cdot \frac{dt}{\sqrt{t}}, \end{aligned}$$

ce qui démontre (8.22) pour $n = 1$ car $\Gamma(1/2) = \sqrt{\pi}$.

Pour le cas général, nous procédons par récurrence. Nous utilisons le résultat du Lemme 8.2 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

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

nous permet de conclure. □

Pour les variables aléatoires C_i de (8.17), ce théorème montre que

$$\sum_{i=n+1}^{m} C_i^2 \tag{8.23}$$

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.6 (voir la formulation (8.12)). Dans ce cas, on a $\|c''\|_2^2 = 25.2$ et $m-n = 18$ degrés de liberté. La figure IV.5 montre que cette valeur de $\|c''\|_2^2$ est suffisamment petite pour être probable.

Si l'on avait travaillé avec le modèle plus simple

$$U = a + bT \tag{8.24}$$

(à la place de (6.4)) 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, la loi (8.24) est à rejeter sur la base de ces mesures.

IV.9 Exercices

1. Pour calculer l'inverse d'une matrice dont la dimension n est très grande, il existe un algorithme qui exige environ n^3 opérations. Donner cet algorithme.

2. Supposons que la décomposition LR de la matrice A est à disposition. Pour u, v, b des vecteurs donnés, trouver un algorithme efficace pour résoudre le système

   $$(A + uv^T)x = b$$

   qui utilise uniquement la résolution des systèmes $A^{-1}b$ et $A^{-1}u$. Cet algorithme est connu sous la formule de Sherman - Morrison - Woodbury.
   Indication. Calculer d'abord une formule pour $v^T x$.

3. Considérons le problème de calculer le produit scalaire

   $$\langle x, y\rangle \;= \sum_{i=1}^{n} x_i y_i \;.$$

   Quelle est sa condition?

4. Soit A une matrice $n \times m$ à coefficients réels. Montrer que

$$\|A\|_1 = \|A^T\|_\infty \qquad \text{et} \qquad \|A\|_2 = \|A^T\|_2 \ .$$

5. Considérons une matrice-bande avec une largeur inférieure p_ℓ et une largeur supérieure p_u (c'est-à-dire, $a_{ij} = 0$ si $i - j > p_\ell$ et si $j - i > p_u$). Montrer que les matrices L et R de la décomposition LR avec et sans la recherche de pivot ont aussi une structure de bande. Pour le cas tridiagonal, $p_\ell = p_u = 1$, donner les largeurs des bandes apparaissants dans les décompositions et estimer le coût en opérations des algorithmes.

6. Pour résoudre le système linéaire

$$\sum_{j=1}^{n} c_j^{i-1} x_j = b_i, \qquad i = 1, \ldots, n \tag{9.1}$$

(matrice du type Vandermonde), dériver un algorithme qui nécessite seulement $\mathcal{O}(n^2)$ opérations.

Indications.

(a) Le système (9.1) est équivalent à

$$\sum_{j=1}^{n} p(c_j) x_j = b(p) \qquad \text{pour} \quad \deg p \le n-1,$$

où $b(p) = \sum_{j=1}^n d_j b_j$ et $p(i) = \sum_{j=1}^n d_j i^{j-1}$.

(b) Choisir pout $p(t)$ les éléments de la base $1, t-c_1, (t-c_1)(t-c_2), (t-c_1)(t-c_2)(t-c_3), \ldots$

7. (a) Pour la matrice

$$A = \begin{pmatrix} 1 & 0 \\ -1 & 4 \\ 4 & 1 \end{pmatrix}$$

calculer $\|A\|_1$, $\|A\|_2$ et $\|A\|_\infty$.

(b) Démontrer que pour des matrices symétriques nous avons toujours $\|A\|_2 \le \|A\|_1$.

8. Les valeurs de la suite $b_k = \exp(k^{2/3} - (k-1)^{2/3})$ peuvent être calculées par les formules:

$$\begin{aligned} b_k &= \exp(k^{2/3} - (k-1)^{2/3}), \\ b_k &= \exp(k^{2/3}) / \exp((k-1)^{2/3})), \\ b_k &= \exp((2k-1)/(k^{4/3} + (k(k-1))^{2/3} + (k-1)^{4/3})). \end{aligned}$$

Calculer à l'aide d'une calculatrice la valeur pour $k = 100000$ (le résultat est $b_{100000} = 1.01446656424210809769528199600$) et pour k grand, quelle formule est préférable pour un calcul en virgule flottante?

9. Les racines du polynôme $x^2 - 2px - q = 0$ peuvent être calculées par

$$\lambda_1 = p + \sqrt{p^2 + q}, \qquad \lambda_2 = p - \sqrt{p^2 + q}.$$

Montrer que pour $p > 0$ (grand) et $q > 0$ (très petit) cet algorithme est numériquement instable. A l'aide de la relation $\lambda_1 \lambda_2 = -q$, trouver un algorithme qui est numériquement stable.

10. Soient donnés $x_1, x_2, \ldots, x_n$. Une estimation de la variance peut être calculée par chacune des deux formules suivantes:

$$\sigma^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\mu^2 \Big) \qquad\qquad \sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} \Big(x_i - \mu \Big)^2$$

où $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$ est l'espérance. Quelle formule est la plus stable?

a) Appliquer les deux algorithmes à l'exemple $n = 2$, $x_1 = 3001$, $x_2 = 3003$ et simuler un calcul en virgule flottante avec 4 chiffres.

b) Etudier l'influence des erreurs d'arrondi pour les deux algorithmes, si $n = 2$ mais que x_1 et x_2 sont arbitraires.

11. a) Calculer la décomposition de Cholesky $A = LL^T$ pour la matrice de Hilbert

$$A = \left(\frac{1}{i+j-1} \right)_{i,j=1,\ldots,n} , \qquad\qquad n = 3, 6, 9, 12, 15.$$

b) Comparer le résultat numérique avec les valeurs exactes

$$\ell_{jk} = \frac{\sqrt{2k-1} \cdot (j-1)! \cdot (j-1)!}{(j-k)! \cdot (j+k-1)!} \, . \tag{9.2}$$

Combien de chiffres sont exacts?

c) Si $\widehat{L}$ dénote le résultat numérique, calculer le résidu $A - \widehat{L}\widehat{L}^T$.

Calculer aussi le résidu $A - LL^T$ pour la matrice L, donnée par (9.2).

12. Pour une matrice $A = (a_{ij})$ notons par $(a_{ij}^{(k)})$ les matrices des étapes intermédiaires de l'élimination de Gauss. Montrer qu'avec une recherche de pivot partielle, on a

$$\max_{i,j,k} |a_{ij}^{(k)}| \le 2^{n-1} \cdot \max_{i,j} |a_{ij}|. \tag{9.3}$$

Pour la matrice suivante, on a égalité dans la formule (9.3):

$$A = \begin{pmatrix} 1 & & & & & 1 \\ -1 & 1 & & & & 1 \\ -1 & -1 & 1 & & & 1 \\ \vdots & \vdots & \vdots & & & \vdots \\ -1 & -1 & -1 & \ldots & -1 & 1 \end{pmatrix}.$$

13. Soit A une matrice à m lignes et n colonnes ($m \ge n$). On définit pour les matrices non-carrées,

$$\kappa(A) := \max_{||x||=1} ||Ax|| \, / \min_{||y||=1} ||Ay||.$$

Pour la norme Euclidienne, montrer que $\kappa_2(A^TA) = (\kappa_2(A))^2$.
Indication. Transformer la matrice symétrique A^TA sous forme diagonale $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_n \ge 0$ et montrer que

$$\max_{\|x\|_2=1} \|Ax\|_2^2 = \lambda_1, \quad \min_{\|x\|_2=1} \|Ax\|_2^2 = \lambda_n.$$

14. Voici quelques valeurs pour la densité ϱ de l'eau en fonction de sa température T.

$T\,[^\circ C]$	0	5	10	15	20
$\varrho(T)$	0.999868	0.999992	0.999728	0.999126	0.998232

(a) Approcher ces valeurs par un polynôme de degré 2 (méthode des moindres carrés).

(b) Pour quelle valeur de T, la densité est-elle maximale et quelle est cette valeur maximale?

Indication. Si vous préférez calculer avec des nombres plus petits, faites la transformation $x = T/5$, $f(T) = 1 - \varrho(T)$.

15. Soit A une matrice inversible de dimension n. Montrer que la décomposition QR (où Q est orthogonale et R triangulaire supérieure) est unique, si l'on suppose que $r_{ii} > 0$ pour $i = 1, \ldots, n$.

16. Soient X et Y deux variables aléatoires indépendantes obéissant à la loi normale $N(\mu_1, \sigma_1)$ et $N(\mu_2, \sigma_2)$ respectivement. Montrer que $\alpha X + \beta$ (pour $\alpha > 0$) et $X + Y$ obéissent aussi à cette loi.

17. Soit X une variable aléatoire qui obéit à la loi χ^2 avec n degrés de liberté (c.-à-d., $f_n(x)$ de (8.22) est sa fonction de densité). Montrer que

$$E(X) = n \qquad \text{et} \qquad \mathit{Var}(X) = 2n.$$

18. Effectuer une étude complète de l'erreur du modèle trouvé à l'exercice 14. Pour cela, trouver les écarts types des coefficients du polynôme et effectuer un test de confiance du modèle.
Indication. $\|c''\|^2 = \|Ax - b\|^2$.

19. Les éléments de la diagonale de $C = (A^T A)^{-1}$ jouent un rôle important pour l'étude de l'erreur de la méthode des moindres carrés. Supposons que nous avons à disposition la décomposition QR de la matrice A.

(a) Démontrer que $C = (R^T R)^{-1}$.

(b) Trouver un algorithme pour calculer la diagonale de C en $n^3/6$ opérations (n = nombre de colonnes de A; 1 opération = 1 multiplication + 1 addition).