Fragments d'une histoire des systèmes linéaires

Anne Michel-Pajus

Dans les cours de mathématiques des années 60, les systèmes d'équations linéaires n'avaient rien de bien passionnant. Leur résolution se présentait comme une **application des déterminants**, condensée dans un théorème dit de **Rouché-Fontené**, aussi ennuyeux à énoncer qu'à utiliser. Ensuite est apparue dans les programmes la méthode dite "du **pivot de Gauss**", et je me suis demandé pourquoi l'on attachait le nom de ce prestigieux mathématicien à ce qui ne semblait rien d'autre qu'une résurgence systématisée de la bonne vieille méthode des combinaisons linéaires. En même temps affluaient dans les problèmes de concours des méthodes de résolution par **approximations successives** qui semblaient construites sur mesure pour les ordinateurs. J'ai appris alors que la résolution des très grands systèmes était toujours d'actualité, et loin d'être close.

Sur ces entrefaites j'ai rencontré, dans le cadre de la Commission Inter-Irem d'Histoire et Epistémologie, un petit groupe de gens qui s'intéressaient à l'histoire des algorithmes, et nous avons décidé d'écrire un livre, qui doit paraître l'an prochain [7]. Je tiens à dire tout ce que je dois à ce groupe, et particulièrement à Jean-Luc Chabert et Michel Guillemot.

L'histoire des systèmes linéaires peut se découper en trois phases. Pendant la première, l'objet mathématique **"système linéaire"** n'existe pas en tant que tel. On rencontre seulement des problèmes particuliers, souvent habillés de scénarios aussi improbables que les histoires de robinets du Certificat d'études, assortis d'une solution numérique, et que nous interprétons, nous, comme une résolution de systèmes. Ce qui n'empêche pas les mathématiciens de cette époque de chercher **des méthodes générales** pour résoudre ces problèmes.

La deuxième phase commence après l'invention du symbolisme algébrique, à la fin du 16ème siècle. C'est quand on représente non seulement les inconnues, mais aussi les coefficients, par des lettres, que peut se dégager la structure de "système d'équations". Les efforts portent alors sur la recherche de formules générales donnant les solutions en fonction des coefficients. L'expression, puis la démonstration, de ces formules va donner naissance à la théorie des déterminants, puis des matrices, contribuant ainsi largement à **la naissance de l'algèbre linéaire.**

Au début du 19ème siècle enfin, vont apparaître des méthodes destinées à résoudre numériquement, de façon approchée, des systèmes de grande taille, dont les coefficients sont eux-mêmes donnés avec une certaine approximation. Ces méthodes visent à résoudre des problèmes issus d'autres domaines scientifiques, l'astronomie et la géodésie, considérés d'ailleurs à l'époque comme des branches des mathématiques. La généralisation et la théorisation de ces méthodes va constituer après la deuxième guerre mondiale une part importante de **l'analyse numérique.**

I• Un problème babylonien

Les tablettes babyloniennes nous présentent des problèmes numériques plus ou moins "concrets", suivis d'une **liste d'instructions** qui conduit à la solution. Les problèmes que nous ramenons à des systèmes linéaires sont souvent résolus par élimination successive des inconnues, parfois assez nombreuses. (Il existerait, d'après [22], un problème à 10 équations et 10 inconnues.)

Le problème de la tablette VAT 8389, gravée vers 1800 av. J.C., ne comporte que deux inconnues: les surfaces de deux champs, connaissant la somme des surfaces (que nous noterons x et y), le rendement de chacune (a et b), et la différence du grain récolté. La traduction est faite à partir de celle de HØYRUP[18]. J'ai ajouté les termes entre crochets et les symboles d'unités pour la rendre plus lisible.

Rappelons que les babyloniens de cette période utilisent une méthode de numération de position et sexagésimale , mais sans symbole qui indiquerait l'absence d'une unité, et différentes mesures de surface et de capacités, sans les préciser non plus. Ici les unités de surface sont le bur et le sar (1 bur = 30x60 sar) et celles de capacité le gur et le sila (1 gur = 5x60 sila). De plus, les instructions mêlent les opérations qui servent à la résolution et les calculs intermédiaires qui servent à exécuter les opérations arithmétiques de base. Les divisions, par exemple, se font par multiplication par l'inverse, que le scribe peut trouver, ou non, dans une table, ainsi que les conversions d'unité. Certaines fractions, comme $\frac{1}{2}$, $\frac{1}{3}$ sont considérées au même titre que les unités.

Tropfke[36], Van der Waerden[38], Høyrup[18], ont proposé des explications à cet enchaînement de calculs. Nous ne les détaillerons pas ici, mais ce premier exemple montre déjà le rôle que jouent en mathématiques les moyens dont on dispose pour les exprimer (numération, notations.).

$x + y = S$

$ax - by = d$
après conversion en sar et sila (cf. plus loin) on obtient:
$a = \frac{40}{60}$ et $b = \frac{30}{60}$
$d = 8(60) + 20$ $S = 30x60$

Par 1 bur [de surface d'un premier champ], j'ai récolté 4 gur de grain
Par 1 bur [de surface d'un autre champ], j'ai récolté 3 gur de grain.
Le grain [du premier] excède l'autre de 8;20 [en sila]
J'ai additionné mes champs: 30[x60 sar]

Que sont mes champs?

calcul des rendements en sila/bur
4 gur/ bur = 20x60 sila/ bur
3 gur / bur = 15x60 sila/bur

$S/2 = 15x60$ ➥M_1

Pose 30[sar] le bur [de la première parcelle]: pose 20[sila] le grain récolté. Pose 30[sar] le bur [de la] second[e parcelle]: pose 15[sila] le grain récolté. Pose 8;20 ce dont le [premier] grain excède l'autre et pose 30[x60sar] l'accumulation des surfaces des champs.
Et 30 l'accumulation des surfaces des champs fractionnée en deux est 15. Pose 15 et 15 deux fois.

calcul des rendements en sila/sar
$1 \text{ sar} = \frac{1}{30x60} \text{ bur} = \frac{2}{60x60} \text{ bur}$
$20x60 \times \frac{2}{60x60} = \frac{40}{60} = a$
$M_1 a = S/2 = 10$ ➥M_2

Cherche l'inverse de 30, la [surface de la première] parcelle: 2[soixantièmes]. [Multiplie] 2 par 20, le grain qu'on y récolte: il vient 40, la fausse quantité de grain. [Multiplie] par 15 que tu as posé deux fois: il vient 10. Que ta tête le retienne.

$15x60x\frac{2}{60x60} = \frac{30}{60} = b$

Cherche l'inverse de 30, la seconde parcelle: 2. [Multiplie] par 15 , le grain qu'on y récolte: 30, le grain faux. Multiplie par 15 que tu as posé deux fois: 7;30.

$M_1 b = 7 + \frac{30}{60} = 7;30 \rightarrow M_3$

10, que ta tête retient, de quoi excède-t'il 7;30? Il l'excède de 2;30.Enlève cet excès de 2;30 à 8;20 dont un grain excède l'autre tu laisses 5;50. Que ta tête le retienne.

$M_2 - M_1 = 2;30$

$d - (M_2 - M_3) = 5;50 \rightarrow M_4$

Accumule le 40 [grain faux] et le 30: il vient 1;10. [je n'en connais pas] l'inverse. Que doit-on poser par 1;10 pour obtenir 5;50 que ta tête retient? Pose 5, par 1;10, cela te donne 5;50.

$a + b = 1;10$

$\frac{M_4}{a+b} = 5 \rightarrow M_5$

Des 15 que tu as posé deux fois, soustrais de l'un, ajoute à l'autre ce 5 que tu as posé.

$x = M_1 + M_5 = 20$

$y = M_1 - M_5 = 10$

Premièrement, [il vient] 20, deuxièmement: 10. 20 est la superficie du premier champ, 10 est celle du second champ.

Vérification

Si 20 est la surface du premier champ, 10 la surface du second champ, quel est le grain donné par les deux?

Cherche l'inverse de 30, la parcelle:2. Multiplie 2 par 20 le grain qu'on y récolte: 40. Multiplie par 20 la surface du champ: 13;20.

[Cherche]l'inverse de 30, la seconde parcelle:2. [Multiplie] 2 par 15 ,le grain qu'on y récolte: 30. [Multiplie]30 par 10, la superficie du champ:il vient 5:c'est le grain des 10, la surface du second champ.

Le grain 13;20 [du premier champ], de combien excède-t'il le grain 5[du second champ]? [il l'excède de 8;20).

2• Un problème de DIOPHANTE (IIIème siècle après J.C.)

Ce problème n'est pas caractéristique des travaux de Diophante, qui s'intéresse généralement à des problèmes qui admettent une infinité de solutions, et dont il cherche les solutions rationnelles strictement positives. Mais il est intéressant par sa limpidité, et l'introduction explicite d'**une inconnue supplémentaire**. Le traducteur contemporain, Ver Eecke[10] a forgé le néologisme d'"arithme" pour la distinguer des autres, appelées "nombres", mais Diophante utilise le même mot.

Trouver trois nombres qui, pris deux à deux, forment des nombres proposés.

Il faut toutefois que la moitié de la somme des nombres proposés soit plus grande que chacun de ces nombres.

Proposons donc que le premier nombre, augmenté du second, forme 20 unités; que le second, augmenté du troisième, forme 30 unités, et que le troisième, augmenté du premier, forme 40 unités.

Posons que la somme des trois nombres est 1 arithme. Dès lors, puisque le premier nombre plus le second forment 20 unités, si nous retranchons 20 unités de 1 arithme, nous aurons comme troisième nombre 1 arithme moins 20 unités. Pour la même raison, le premier nombre sera 1 arithme moins 30 unités, et le second nombre sera 1 arithme moins 40 unités. Il faut encore que la somme des trois nombres devienne égale à 1 arithme. Mais, la somme des trois nombres forme 3 arithmes moins 90 unités. Egalons-les à 1 arithme, et l'arithme devient 45 unités.

Revenons à ce que nous avons posé: le premier nombre sera 15 unités, le troisième sera 25 unités, et la preuve est claire.

SCHEMA DES CALCULS:				
x+y = a y+z = b z+x = c	avec	a=20 b=30 c=40	x+y+z=S (arithme)	z= S-a x= S-b y= S-c
				x+y+z=3S-(a+b+c) S=3S-(a+b+c) $S=\frac{(a+b+c)}{2}$
	x = S - b		y = S - c	z = S - a

Des questions qui se traduisent par des systèmes linéaires, énoncées de façon rhétorique, c'est-à dire sans notations algébriques, se rencontrent fréquemment au Moyen-Age en Inde, dans les pays d'Islam et en Europe, mais on en trouve en Chine bien avant.

3• Les méthodes chinoises

Les plus étonnantes sont celles de la dynastie Han (206 av.J.C., 220 ap.J.C). Les voici décrites par Jean-Claude Martzloff[27]:

"Les techniques chinoises reposent sur la répartition des nombres issus de tel ou tel problème arithmétique en colonnes parallèles (hang), chaque colonne traduisant une condition linéaire imposée à un ensemble d'inconnues, appelées les "choses' (wu). Ces colonnes de nombres (représentées concrètement par des groupements de baguettes affectées aux diverses unités décimales des nombres) sont posées au sens propre du mot sur une "surface à calculer" et laissent donc apparaître des configurations de nombres correspondant à ce que nous appellerions "la matrice du système augmentée des seconds membres"...A partir de là, les procédés de résolution utilisent tout un arsenal se composant d'opérations telles que la "multiplication partout"(bian cheng), multiplication d'une colonne de nombres par un même facteur, ou bien la "réduction directe"(zhi chu), c'est-à dire la soustraction terme à terme des coefficients de deux colonnes en vue de parvenir à l'élimination du coefficient de l'une des inconnues. L'un des procédés chinois les plus remarquables consiste à réduire la "matrice" du système à la forme triangulaire, puis à calculer les inconnues par substitutions successives..." (on trouve des exemples d'applications jusqu'à 6 inconnues).

4• Des méthodes de fausse position

Les chinois utilisent aussi la *"ying bu zu shu"* (règle du trop ou du pas assez). Cette méthode occupe un chapitre entier de la bible arithmétique chinoise, le *Jiuzhang suanshu (Procédés calculatoires en neuf chapitres)* . Il s'agit d'un cas particulier de la méthode de double fausse position, que nous détaillerons sur un exemple de résolution d'un problème à 3 inconnues, exposé par Clavius[1]. Il est possible de lire comme des méthodes de fausse position (simple) la résolution de certains problèmes égyptiens et babyloniens. Utilisées aussi par les Indiens (Bhaskara, au XII^{ème}), ces méthodes sont introduites par les Arabes en Europe, où on les retrouve à partir du douzième siècle (chez Ben Ezra, par exemple, et Fibonacci)[33].

---

[1]Jésuite du 16ème siècle, qui oeuvra beaucoup pour l'enseignement des mathématiques dans les collèges de Jésuites.

Cette méthode restera en usage dans l'enseignement jusqu'au XX^{ème} siècle. Elle présente l'avantage de ne pas opérer de manière générale sur une inconnue, dans le **cadre algébrique**, ce qui pose des difficultés conceptuelles, mais de rester dans le **cadre arithmétique**, en travaillant sur des valeurs numériques particulières.

Dans la méthode de double fausse position , appliquée à ce que nous notons

(S_1) $ax_1 + bx_2 = c_1$

(S_2) $cx_1 + dx_2 = c_2$

on part d'une valeur de x_1 (posée ou supposée), et on en déduit x_2 à l'aide de (S_1); on calcule alors le **résidu** $e = cx_1 + dx_2 - c_2$ à l'aide de (S_2).

Un calcul immédiat montre que $e = x_1(c - \frac{ad}{b}) + \frac{dc_1}{b} - c_2$. Nous voyons que:

- e est une fonction affine de x_1
- e = 0 correspond à la valeur exacte de x_1.

Il suffit donc de choisir deux valeurs de x_1 (x'_1 et x''_1), de calculer les valeurs correspondantes de e (e' et e''), et d'utiliser une interpolation linéaire pour obtenir la valeur exacte de x_1

La Règle (nommée **Regula Falsi** au Moyen-Age) donne

$$x_1 = \frac{x'_1\, e'' - x''_1\, e'}{e'' - e'}$$

Cette méthode et sa justification se généralisent à un système d'ordre n. ..Soit le système (supposé de Cramer):

(S_1) $a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = c_1$

(S_2) $a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = c_2$

. .

(S_{n-1}) $a_{n-11}x_1 + a_{n-12}x_2 + \ldots + a_{1n\_1n}x_n = c_{n-1}$

(S_n) $a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = c_n$

Pour tout n-uplet $(x_1, x_2, \ldots, x_n)$ vérifiant les (n-1) premières équations, on appelle résidu $e = a_{n-1}x_1 + \ldots + a_{nn}x_{n-} c_n$.

Si l'on fixe x_1 par exemple, on en déduit $(x_2, \ldots, x_n)$ comme solution du système

(S_1) $a_{12}x_2 + \ldots + a_{1n}x_n = c_1 - a_{11}x_1$

(S_2) $a_{22}x_2 + \ldots + a_{2n}x_n = c_2 - a_{21}x_1$

. .

(S_{n-1}) $a_{n-12}x_2 + \ldots + a_{1n\_1n}x_n = c_{n-1} - a_{n-1}x_1$

$x_2, \ldots, x_n$ sont donc des fonctions affines de x_1, et donc e également. Ceci est évidemment valable pour chaque x_i. Il suffit donc de disposer pour chaque variable de 2 couples de valeurs $(x'_i, e(x'_i))$ et $(x''_i, e(x''_i))$ pour obtenir la valeur exacte de x_i à l'aide de la regula falsi .

Il est intéressant de noter qu'il n'est pas nécessaire de recommencer tout le processus pour chaque variable. En effet, les calculs intermédiaires exécutés pour la première variable fournissent 2 n-uplets solutions des (n-1) premières équations et les 2 valeurs correspondantes du résidu.

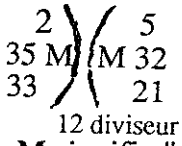
CLAVIUS *Aritmetica prattica,* Rome 1586 (traduction Maryvonne Spiesser et Anne Michel-Pajus[33])

On cherche trois nombres tels que : le premier augmenté de 73 égale le double des deux autres; le second augmenté de 73 égale le triple des deux autres , le troisième augmenté de 73 le quadruple des deux autres.

Tu poses le premier nombre égal à 1 (..). Puisque 1 avec 73 fait 74, lequel doit selon la question proposée être le double des deux autres, il faut que la somme des deux autres soit 37. Et parce que le second avec 73 doit faire un nombre triple du premier(qui est 1) et du troisième ensemble, il faudra partager (comme dans la question enseignée précédemment) le nombre 37 en deux parties, dont la première avec 73 fait un nombre triple du nombre composé de la seconde partie et du 1. Ainsi, avant de résoudre la question proposée, il est nécessaire d'en résoudre une autre qui apparaît dans cette opération.

Tu poses donc que la première partie de 37 est 2 et par suite la seconde est 35. La première partie avec 73 fait 75 et la seconde avec 1 fait 36, dont le triple n'est pas 75, mais 108. Donc nous avons manqué la vérité de 33 unités.

Tu poses alors la première partie égale à 5 et donc la seconde égale à 32; la première avec 73 fait 78 et la seconde avec 1 fait 33, duquel le triple n'est pas 78 , mais 99. Donc on a manqué de nouveau la vérité de 21 unités.



12 diviseur

On fait alors, selon le précepte de la règle du faux et on trouvera que la première partie est 10 1/4, et la deuxième partie est 26 3/4. (...)

N.B. **M** signifie "manque" , **P** signifie "(dé)passe"

Résumons ce passage et le suite avec des notations algébriques:

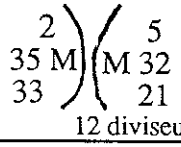
Système à résoudre:

x+73=2(y+z)
y+73=3(z+x)
z+73=4(x+y)

Si x=1 alors y+z = 37
y+73 = 3(z+1)

Si y = 2 alors z=35
(e'= -33) 33 Manque

Si y=5 alors z=32
(e"=-21) 21 Manque



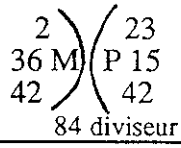
12 diviseur

y=10 1/4
z=26 3/4

Si x=3 alors y+z = 38
y+73 = 3(z+3)

Si y = 2 alors z=36
(e'= -42) 42 Manque

Si y=23 alors z=15
(e"= 42) 42 Pass

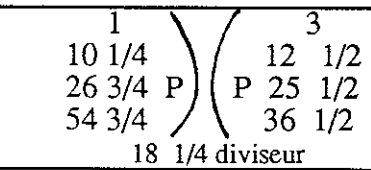


84 diviseur

y=12 1/2
z=25 1/2

z+73=99 3/4 4(x+y)=45
54 3/4 Passe

z+73=98 1/2 4(x+y)=62
36 1/2 Pass



18 1/4 diviseur

x = 7
y = 17
z = 23

II Les formules générales

C'est à la fin du 16ème qu'à la suite de **Viète** (*l'Introduction à l'Art Analytique* est publiée en 1591), les mathématiciens désignent par des lettres, non seulement les inconnues (voyelles) et leurs puissances, mais les coefficients indéterminés (consonnes). Ce mouvement se poursuit chez les mathématiciens du $17^{ème}$, en particulier avec **Descartes** qui désire utiliser l'algèbre en Géométrie, et il semble donc que ce soit à la fin du $17^{ème}$ qu'apparaît l'objet mathématique que nous appelons système linéaire, c'est à dire un système d'équations général, avec des coefficients littéraux. Les notations adoptées ne sont pas très pratiques, et ne mettent pas toujours en évidence la structure des formules. Leibniz, génial précurseur, restera longtemps ignoré.

1• Leibniz

Il faut attendre la publication par Gerhardt, entre 1849 et 1863, de nombreux manuscrits de Leibniz pour découvrir ses travaux sur les équations. Depuis cette date, les chercheurs continuent à étudier ses innombrables manuscrits . J'ai traduit le texte suivant d'après une édition de 1980([23] p 5-6). Il est daté de juin 1678. On y trouve les formules généralement attribuées à Cramer, avec des "déterminants" développés selon une colonne, celle qui est justement différente au numérateur et au dénominateur.

Pour retrouver nos notations habituelles, il suffit de considérer les doubles chiffres comme les doubles indices des coefficients, et les chiffres simples (2,3,4,5) comme les indices des inconnues. Ainsi,
12,2 + 13,3 + 14,4 + 15,5 - A égal à 0 se traduit par :
$a_{12}x_2 + a_{13}x_3 + a_{14}x_4 + a_{15}x_5 - A = 0$
12,23,34,45 par $a_{12} \cdot a_{23} \cdot a_{34}\ a_{45}$
La dernière formule du texte ci-après correspond à

$$x_5 = \frac{\begin{vmatrix} a_{12} & a_{13} & a_{14} & A \\ a_{21} & a_{23} & a_{24} & B \\ a_{32} & a_{33} & a_{34} & C \\ a_{42} & a_{43} & a_{44} & D \end{vmatrix}}{\begin{vmatrix} a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{23} & a_{24} & a_{25} \\ a_{32} & a_{33} & a_{34} & a_{35} \\ a_{42} & a_{43} & a_{44} & a_{45} \end{vmatrix}}$$

lorsque l'on développe les déterminants suivant la dernière colonne.

Specimen d'Analyse nouvelle par laquelle on évite les erreurs, l'esprit est conduit comme par la main, et l'on trouve facilement des développements.

(...)

L'une des remarques qui peuvent produire en calcul de très grands effets est celle-ci: de même que les théorèmes remarquables, et les développements, se dévoilent facilement et comme spontanément, de même peuvent-ils s'écrire le plus souvent sans calculs, pour autant que l'on ait effleuré les prémisses.

Assurément, si quelqu'un, dans le présent exemple qui comporte quatre lettres [inconnues] données, reliées par un nombre égal d'équations simples, cherche la valeur de l'une d'elles et choisit indistinctement comme il est d'usage les lettres a b x y comme il lui plaît, il sera confronté à une horrible confusion et à un labeur immense. Avec notre manière, la chose est accomplie presque sans tracas, comme ce sera évident.

La règle qui résume donc cet art de la caractéristique est que tous les caractères qui se trouvent dans l'objet désigné expriment ce qui, au moyen de nombres, apportera le maximum de facilité de copie et de calcul. Et de même cela sera d'un grand usage en géométrie, pour exprimer l'emplacement. [rature]

12,2 + 13,3 + 14,4 + 15,5 - A égal à 0
22,2 + 23,3 + 24,4 + 25,5 - B égal à 0
32,2 + 33,3 + 34,4 + 35,5 - C égal à 0
42,2 + 43,3 + 44,4 + 45,5 - D égal à 0

Le dénominateur de la fraction exprimant la valeur de [l'inconnue] 5 est:

12,23,34,45	*12,23,35,44*	*12,24,33,45*	*12,24,35,43*	*12,25,33,44*	*13,25,34,43*
- 13,22,34,45 +	*13,22,35,44 -*	*13,24,32,45 +*	*13,24,35,42 -*	*13,25,32,44 +*	*13,25,34,42*
14, 23, 32,45	*14,23,35,42*	*14,22,35,43*	*14,22,33,45*	*14,25,32,43*	*14,25,33,42*
15,22,33,44	*15,22,34,43*	*15,23,32,44*	*15,23,34,42*	*15,24,32,43*	*15,24,33,42*

que nous écrivons très facilement au moyen de nombres comme l'on voit, puisque'il faut prendre tour à tour parmi les quantités ou plutôt les nombres , 1.2.3.4. pour les premiers caractères, et pour les suivants,d'abord 2.3.4.5./ 2.3.5.4./ 2.4.3.5./ 2.4.5.3./ 2.5.3.4./ 2.5.4.3./ avec 2 placé en premier,puis 3.2.4.5./3.2.5.4./3.4.2.5./3.4.5.2./3.5.2.4./3.5.4.2./ et ainsi de suite en suivant les transpositions possibles des nombres 2.3.4.5. Où il est clair aussi que l'on obtient la ligne inférieure à partir de la supérieure en gardant les mêmes nombres excepté deux qui s'échangent, comme dans

2.3.4.5

3.2.4.5.

évidemment la ligne inférieure est obtenue à partir de la supérieure en changeant 2 en 3 et inversement. Ceci étant entendu pour les caractères qui suivent ou plutôt qui sont à droite. D'où, en ordonnant, on aura le numérateur et le dénominateur ou plutôt la fraction:

- 12,23,34	*+12,23,44*	*- 12,33,44*	*+ 22,33,44*
+ 12,24,33	*- 12,24,43*	*+12,34,43*	*- 22,34,43*
+ 13,22,34	*- 13,22,44*	*+ 13,32,44*	*- 23,32,44*
- 13,24,32	*+ 13,24,42*	*- 13,34,42*	*+ 23,34,42*
- 14,22,33	*+ 14,22,43*	*- 14,32,43*	*+ 24,32,43*
+ 14,23,32	*- 14,23,42*	*+ 14,33,42*	*- 24,33,42*

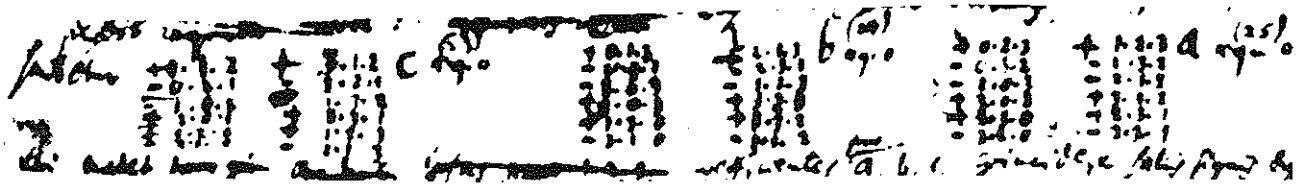
[var] 5 égale

+ 45	*+ 35*	*- 25*	*+.....15*
-	-	+	-.....
+	+	-	+.....
-	-	+	-.....
+	+	-	+.....
-	-	+	-.....

Où il apparaît également facilement la façon dont il faut écrire. Si je n'avais pas utilisé des nombres au lieu de lettres, je n'aurais pas pu l'écrire ainsi facilement. (...)

Fragment de manuscrit
LH XXXV 14,1 Bl. 261
(Stück 8, Zeile 1-116)

Gottfried Wilhelm Leibniz



2• MACLAURIN

Nous avons dit que, faute de publication, les travaux de Leibniz resteront longtemps ignorés, et c'est en 1748 qu'est publié, à titre posthume, l'ouvrage de Colin Maclaurin,*"A treatise of algebra"* datant sans doute de 1729. Voici les "théorèmes généraux pour trouver les valeurs des inconnues", dans une traduction de 1753. [26]

THÉOREMES GÉNÉRAUX

Pour trouver les valeurs des inconnues.

204. J'appellerai coëfficiens du même ordre, les quantités qui accompagnent la même inconnue dans différentes équations, ou celles qui n'en accompagnent aucune.

205. J'appellerai coëfficiens oppoſés, ceux qui accompagnent différentes inconnues dans différentes équations.

THÉOREME PREMIER.

206. Deux équations & deux inconnues étant données, chacune de ces inconnues eſt égale à une fraction dont le numérateur eſt la différence des produits des coëfficiens oppoſés des ordres où cette inconnue ne ſe trouve point, & le dénominateur eſt la différence des produits des coëfficiens oppoſés des deux inconnues.

DEMONSTRATION.

Soit ..... $a x + b y = c$

$d x + e y = f$

je dis que ..... $y = \frac{a f - d c}{a e - d b}$

& ..... $x = \frac{c e - b f}{a e - d b}$

car par la premiere équation

$$x = \frac{c - b y}{a}$$

par la deuxiéme ... $x = \frac{f - e y}{d}$

donc ..... $\frac{c - b y}{a} = \frac{f - e y}{d}$

$$c d - d b y = a f - a e y \quad \Big| \quad y = \frac{a f - c d}{a e - d b}$$
$$a e y - d b y = a f - c d$$

ſubſtituant cette valeur de y dans une des valeurs de x, on trouvera

$$x = \frac{c e - b f}{a e - d b}$$

THÉOREME II.

207. Trois équations & trois inconnues étant données, chaque inconnue ſera égale à une fraction dont le numérateur contiendra tous les produits qu'on peut faire de trois coëfficiens oppoſés, pris dans les ordres où cette inconnue ne ſe trouve point, & le dénominateur contiendra les différens produits qu'on peut former de trois coëfficiens oppoſés, pris dans les ordres qui renferment les trois inconnues.

DEMONSTRATION.

Soit les trois inconnues x, y, z, & les trois équations

$$ax + by + cz = m$$
$$dx + ey + fz = n$$
$$gx + hy + kz = p$$

Ne regardant d'abord comme inconnue que x & y, & n'ayant égard qu'aux deux premieres équations, on aura par le Théorême précédent

$$y = \frac{an - afz - dm + dcz}{ae - db}$$

n'ayant égard qu'à la premiere & à la derniere

$$y = \frac{ap - akz - gm + gcz}{ah \quad gb}$$

donc $\frac{an - afz - dm + dcz}{ae - db} = \frac{ap - akz - gm + gcz}{ah - gb}$

& ayant fait évanouir les fractions, détruit les termes ſemblables, & diviſé les deux membres de l'équation par a, on trouvera

$$aekz - afhz + cdhz - bdkz + bfgz - cegz = aep - ahn + dhm - dbp + gbn - gem$$

& par conſéquent $z = \frac{aep - ahn + dhm - dbp + gbn - gem}{aek - afh + cdh - ceg + bdk - bfg}$

pour trouver la valeur de y, au lieu de ſubſtituer celle de z, ce qui ſeroit une opération fort embarraſſante, on recommence entierement l'opération, dans laquelle on traite y comme on a traité z dans la précédente, & on trouve

$$y = \frac{afp - akn + dkm - dep + cgn - fgm.}{aek - afh + cdh - bdk + bfg - ceg.}$$

S'il manquoit quelque termes dans quelqu'une des trois équations données, on trouveroit des valeurs plus ſimples des inconnues; je ſuppoſe, par exemple, $f = 0$, $k = 0$; alors le terme fz s'évanouira dans la ſeconde équation, & kz dans la troiſiéme, & l'on aura

$$z = \frac{aep - ahn + dhm - bdp + bgn - egm}{cdh - ceg}$$

$$y = \frac{cgn - dep.}{cdh - ceg.}$$

3• CRAMER

Cramer est amené à rechercher ces formules dans le cadre de ses travaux de géométrie, en particulier la détermination des coniques passant par cinq points donnés.. Il les publie dans *l'Introduction à l'analyse des lignes courbes algébriques* en 1750[8].

Ses notations sont un peu meilleures que celles de Maclaurin, mais pas aussi pratiques que celles de Leibniz ! Il démontre les résultats pour les systèmes de 2 et 3 équations puis généralise par induction la formule pour n équations. Il étudie ensuite des systèmes impossibles ou indéterminés.

Soient pluſieurs inconnues z, y, x, v, &c. & autant d'équations

$$\begin{aligned} A' &= Z'z + Y'y + X'x + V'v + \&c. \\ A^{2} &= Z^{2}z + Y^{2}y + X^{2}x + V^{2}v + \&c. \\ A^{3} &= Z^{3}z + Y^{3}y + X^{3}x + V^{3}v + \&c. \\ A^{4} &= Z^{4}z + Y^{4}y + X^{4}x + V^{4}v + \&c. \\ &\&c. \end{aligned}$$

où les lettres A', A^{2}, A^{3}, A^{4}, &c. ne marquent pas, comme à l'ordinaire, les puiſſances d'A, mais le prémier membre, ſuppoſé connu, de la prémiére, ſeconde, troiſiéme, quatriéme &c. équation. De même Z', Z^{2}, &c. ſont les coëfficients de z; Y', Y^{2}, &c. ceux de y; X', X^{2}, &c. ceux de x; V', V^{2}, &c. ceux de v; &c. dans la prémiére, ſeconde, &c. équation.

Cette Notation ſuppoſée, s'il n'y a qu'une équation & qu'une inconnue z; on aura $z = \frac{A'}{Z'}$. S'il y a deux équations & deux inconnues z & y; on trouvera $z = \frac{A'Y^{2} - A^{2}Y'}{Z'Y^{2} - Z^{2}Y'}$, & $y = \frac{Z'A^{2} - Z^{2}A'}{Z'Y^{2} - Z^{2}Y'}$. S'il y a trois équations & trois inconnues z, y, & x; on trouvera

$$z = \frac{A'Y^{2}X^{3} - A'Y^{3}X^{2} - A^{2}Y'X^{3} + A^{2}Y^{3}X' + A^{3}Y'X^{2} - A^{3}Y^{2}X'}{Z'Y^{2}X^{3} - Z'Y^{3}X^{2} - Z^{2}Y'X^{3} + Z^{2}Y^{3}X' + Z^{3}Y'X^{2} - Z^{3}Y^{2}X'}$$

$$y = \frac{Z'A^{2}X^{3} - Z'A^{3}X^{2} - Z^{2}A'X^{3} + Z^{2}A^{3}X' + Z^{3}A'X^{2} - Z^{3}A^{2}X'}{Z'Y^{2}X^{3} - Z'Y^{3}X^{2} - Z^{2}Y'X^{3} + Z^{2}Y^{3}X' + Z^{3}Y'X^{2} - Z^{3}Y^{2}X'}$$

$$x = \frac{Z'Y^{2}A^{3} - Z'Y^{3}A^{2} - Z^{2}Y'A^{3} + Z^{2}Y^{3}A' + Z^{3}Y'A^{2} - Z^{3}Y^{2}A'}{Z'Y^{2}X^{3} - Z'Y^{3}X^{2} - Z^{2}Y'X^{3} + Z^{2}Y^{3}X' + Z^{3}Y'X^{2} - Z^{3}Y^{2}X'}$$

L'examen de ces Formules fournit cette Règle générale. Le nombre des équations & des inconnues étant n, on trouvera la valeur de chaque inconnue en formant n fractions dont le dénominateur commun a autant de termes qu'il y a de divers arrangements de n choſes différentes. Chaque terme eſt compoſé des lettres $ZYXV$ *&c.* toujours écrites dans le même ordre, mais auxquelles on diſtribue, comme expoſants, les n prémiers chiffres rangés en toutes les maniéres poſſibles. Ainſi, lorſqu'on a trois inconnues, le dénominateur a [$1\times2\times3=$] 6 termes, compoſés des trois lettres ZYX, qui reçoivent ſucceſſivement les expoſants 123, 132, 213, 231, 312, 321. On donne à ces termes les ſignes + ou —, ſelon la Règle ſuivante. Quand un expoſant eſt ſuivi dans le même terme, médiatement ou immédiatement, d'un expoſant plus petit que lui, j'appellerai cela un *dérangement*. Qu'on compte, pour chaque terme, le nombre des dérangements: s'il eſt pair ou nul, le terme aura le ſigne +; s'il eſt impair, le terme aura le ſigne —. Par ex. dans le terme $Z^1Y^2V^3$ il n'y a aucun dérangement: ce terme aura donc le ſigne +. Le terme $Z^2Y^1X^3$ a auſſi le ſigne +, parce qu'il a deux dérangements, 3 avant 1 & 3 avant 2. Mais le terme $Z^2Y^3X^1$, qui a trois dérangements, 3 avant 2, 3 avant 1, & 2 avant 1, aura le ſigne —.

Le dénominateur commun étant ainſi formé, on aura la valeur de z en donnant à ce dénominateur le numérateur qui ſe forme en changeant, dans tous ſes termes, Z en A. Et la valeur d'y eſt la fraction qui a le même dénominateur & pour numérateur la quantité qui réſulte quand on change Y en A, dans tous les termes du dénominateur. Et on trouve d'une maniére ſemblable la valeur des autres inconnues.

Généralement parlant, le Probléme eſt déterminé. Mais il peut y avoir des Cas particuliers, où il reſte indeterminé; & d'autres où il devient impoſſible. C'eſt lorſque le dénominateur commun ſe trouve égal à zéro

Ces résultats sont immédiatement diffusés et enseignés.*"Cette méthode était tellement en faveur que les examens aux écoles des sevices publics ne roulaient, pour ainsi dire, que sur elle; on était admis ou rejeté suivant qu'on la possédait bien ou mal"* aurait dit Gergonne[1].

[1]Recteur de l'Académie de Montpellier dans la première moitié du 19éme siècle(Muir I, p14)

4• CAUCHY

Malgré de nombreuses tentatives d'expression ou de démonstration des formules générales, comme celle de Bézout (1764 [2]), Vandermonde (1771[37]), Laplace (1772[21]), ce n'est qu'en 1812, dans le fameux Mémoire de Cauchy [3] *"Sur les fonctions qui ne peuvent obtenir que 2 valeurs,égales et de signes contraires par suite des transpositions opérées entre les variables qu'elles renferment"*, que figure la première démonstration générale des formules pour n inconnues. Cauchy y baptise les déterminants (en empruntant le mot à Gauss chez qui il désigne le discriminant d'une forme quadratique) et y inaugure aussi le double indice et la disposition en tableau (sans les barres verticales, qui seront ajoutées par Cayley en 1841).

Le déterminant de ce que nous appelons la matrice A, de terme général $[a_{i,j}]$, est noté $S(+a_{1,1}a_{2,2}a_{3,3}\cdots a_{n,n})$. Après avoir démontré les formules de développement suivant une rangée, et défini le cofacteur b_{ij} de a_{ij}, qu'il appelle quantité adjointe, Cauchy démontre les formules (appelées équations(10) dans l'extrait suivant), et que nous résumons par

$$\sum_{k=1}^{n} a_{i,k} b_{j,k} = \delta^{j}_{i} \det(A)$$

La résolution des systèmes n'est plus qu'une application des déterminants et n' occupe que 4 pages sur 82.

Les équations dont il s'agit seront de la forme

$$(20)\qquad \begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \ldots + a_{1,n}x_n = m_1, \\ a_{2,1}x_1 + a_{2,2}x_2 + \ldots + a_{2,n}x_n = m_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ a_{n,1}x_1 + a_{n,2}x_2 + \ldots + a_{n,n}x_n = m_n. \end{cases}$$

Cela posé, les relations établies par les équations (10) entre les termes du système $(a_{1,n})$ et ceux du système adjoint $(b_{1,n})$ fournissent un moyen facile d'obtenir la valeur d'une inconnue prise à volonté. Ainsi, par exemple, si l'on multiplie la première des équations (20) par $b_{1,1}$, la deuxième par $b_{2,1}$, ..., enfin la dernière par $b_{n,1}$, et qu'on les ajoute entre elles en ayant égard aux équations (10), on aura, pour déterminer la valeur de x_1, l'équation suivante :

$$D_n x_1 = m_1 b_{1,1} + m_2 b_{2,1} + \ldots + m_n b_{n,1}.$$

En général, si l'on multiplie la première des équations (20) par $b_{1,\mu}$, la deuxième par $b_{2,\mu}$, ..., enfin la dernière par $b_{n,\mu}$, et qu'ensuite on les ajoute entre elles, on aura, en vertu des équations (10),

$$D_n x_\mu = m_1 b_{1,\mu} + m_2 b_{2,\mu} + \ldots + m_n b_{n,\mu}. \tag{21}$$

Par suite, la valeur générale de l'une quelconque des inconnues sera

$$x_\mu = \frac{m_1 b_{1,\mu} + m_2 b_{2,\mu} + \ldots + m_n b_{n,\mu}}{D_n} = \frac{m_1 b_{1,\mu} + m_2 b_{2,\mu} + \ldots + m_n b_{n,\mu}}{a_{1,\mu} b_{1,\mu} + a_{2,\mu} b_{2,\mu} + \ldots + a_{n,\mu} b_{n,\mu}}.$$

Cette valeur se présente donc sous la forme d'une fraction qui a pour dénominateur le déterminant

$$D_n = S(\pm a_{1,1} a_{2,2} \ldots a_{n,n})$$

et pour numérateur ce que devient ce déterminant quand on y remplace les coefficients de l'inconnue pris dans les équations (20) par les seconds membres de ces mêmes équations.

Les compléments sur la résolution exacte des systèmes vont préciser les cas d'indétermination et d'impossibilité. Le théorème complet est connu en France sous le nom de théorème de ROUCHE-FONTENE. Ces deux professeurs, dont le premier, polytechnicien, était professeur au Lycée Charlemagne, et le second , agrégé de mathématiques, au Collège Rollin, ont publié plusieurs notes entre 1875 et 1880 sur le sujet. On doit à Eugène Rouché les dénominations de déterminant principal et déterminants caractéristiques. Mais Muir[29] montre que C.L.DODGSON , plus connu sous le nom de Lewis Carroll, avait déjà publié ces résultats en 1867[11].

Notons que ces recherches vont conduire Frobenius, en 1879, à dégager le concept de rang. La résolution des systèmes linéaires joue ainsi un rôle fondamental dans l'émergence des concepts de l'algèbre linéaire.[12]

1. LA METHODE DES MOINDRES CARRéS

Nous avons vu que les mathématiciens disposent donc à partir de la deuxième moitié du XVIII^{ème} de méthodes tout à fait efficaces pour résoudre les systèmes à 2,3, à la rigueur 4, inconnues, mais les recherches dans d'autres branches des mathématiques, telles l'astronomie ou la géodésie, vont conduire à des systèmes **d'une nature toute différente**. Non seulement le nombre d'inconnues est beaucoup plus grand, mais les coefficients des équations sont incertains, puisque résultant de mesures physiques. L'on dispose en revanche d'un nombre pratiquement illimité d'équations...Tous les systèmes obtenus ainsi sont incompatibles ! comment obtenir malgré tout une solution acceptable?

La réponse est la méthode des moindres carrés.

Elle est publiée pour la première fois par Legendre[25] en appendice de ses *Nouvelles méthodes pour la Détermination des orbites des Comètes*, en 1805, mais Gauss l'a utilisée déjà au moins dès 1801, pour déterminer à la stupéfaction générale l'orbite de Cérès, malgré des observations très courtes. La revendication de paternité de la méthode provoquera d'ailleurs des controverses assez violentes ! [6]

Dans le cadre conceptuel actuel, le principe est très simple. Notons $MX+N = 0$, ou $MX = (-N)$ le système à résoudre. Il est incompatible en général, car (-N) n'appartient pas à l'image de M. On va choisir comme solution un vecteur Xo tel que la distance de MXo à (-N) soit minimale; dans une structure euclidienne, c'est donc que MXo est la projection orthogonale de (-N) sur l'image de M. On a donc pour tout vecteur X, $\langle MXo + N , MX\rangle = \langle {}^tMMXo + {}^tMN , X\rangle = 0$. Ce qui équivaut à: Xo est solution de ${}^tMMXo = -{}^tMN$. On se ramène ainsi à un système qui a autant d'équations que d'inconnues, de matrice $A = {}^tMM$, symétrique, positive et généralement définie positive.

Ce n'est évidemment pas ainsi que les mathématiciens du XIX^{ème} justifient la méthode. Prenons les notations de Gauss:

La transposition matricielle du système à résoudre est $N + MX = 0$,et, à tout vecteur X, on associe le vecteur $W= N + MX$, avec

$$M=\begin{pmatrix} a.\ b.\ c\ \dots\ \dots \\ a'\ b'\ c'\ \dots\ \dots \\ a''\ b''\ c''\ \dots\ \dots \\ \dots\ \dots\ \dots\ \dots\ \dots \end{pmatrix} \qquad X=\begin{pmatrix} p \\ p \\ r \\ \dots \end{pmatrix} \qquad N=\begin{pmatrix} n \\ n' \\ n'' \\ \dots \end{pmatrix} \qquad W=\begin{bmatrix} w \\ w' \\ w'' \\ \dots \end{bmatrix}$$

w,w',w" sont appelés les erreurs ou les écarts. La quantité à minimiser est donc $W=w^2+w'^2+w''^2\dots$, qui est fonction des inconnues p,q,r...Legendre (en 1805) obtient le nouveau système en annulant les dérivées partielles de W par rapport à chacune des inconnues. Gauss donne plusieurs justifications, entre 1809 et 1821 en s'appuyant sur un raisonnement de probabilité. On peut aussi utiliser une autre norme pour la distance, c'est ce que fait Euler, en 1749, qui minimise le maximum de la valeur absolue des écarts, ou Laplace, en 1799, qui minimise la somme des valeurs absolues des écarts. Mais la méthode des moindre carrés aboutit à des calculs plus agréables, car elle conserve la linéarité. Les nouvelles équations obtenues sont désignées par le terme d'équations normales.

2. Le "Pivot" de Gauss

C'est dans le *De ellipticis Palladis* (1810) [14] que j'ai trouvé le texte qui se rapproche le plus de ce que nous désignons ainsi[1] . Après le succès remporté avec Cérès, Gauss se penche sur l'orbite de Pallas, découverte par Olbers. Avec 6 inconnues et 11 équations, il obtient 6 équations normales par la méthode des moindres carrés. Pour éviter les "pénibles" calculs de l'élimination classique, Gauss propose une autre méthode, qui consiste en fait à exécuter ce que nous appelons la réduction en carrés de Gauss de la forme quadratique W, avec un carré supplémentaire qui correspond aux termes constants. Il obtient le minimum en annulant tous les carrés, ce qui donne les n équations d'un système triangulaire qui coincide avec celui obtenu par la méthode dite "du pivot".

(Traduction de Joseph Bertrand 1855)

Dans l'impossibilité où nous sommes de satisfaire exactement aux onze équations proposées, c'est-à-dire d'annuler tous les seconds membres, nous chercherons à rendre la somme de leurs carrés aussi petite que possible.

Traduction matricielle

On aperçoit facilement que si l'on considère les fonctions linéaires

$$N + AX = W$$

$$\begin{aligned} &n + ap + bq + cr + ds + \ldots = w,\\ &n' + a'p + b'q + c'r + d's + \ldots = w',\\ &n'' + a''p + b''q + c''r + d''s + \ldots = w'',\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

les équations qu'il faut résoudre pour rendre

$$\Omega = \|AX + N\|^2$$

$$\Omega = ww + w'w' + w''w'' \ldots$$

un minimum, sont

$$\frac{\partial\Omega}{\partial p} = \frac{\partial\Omega}{\partial q} = \ldots = 0$$

$$\begin{aligned} &aw + a'w' + a''w'' + \ldots = 0,\\ &bw + b'w' + b''w'' + \ldots = 0,\\ &cw + c'w' + c''w'' + \ldots = 0,\\ &\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

ou, en posant, pour abréger,

$$\begin{aligned} &an + a'n' + a''n'' + \ldots = (an),\\ &aa + a'a' + a''a'' + \ldots = (aa),\\ &ab + a'b' + a''b'' + \ldots = (ab),\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &bb + b'b' + b''b'' + \ldots = (bb),\\ &bc + b'c' + b''c'' + \ldots = (bc),\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

p, q, r, s, etc., devront se déterminer par les équations suivantes :

$$\begin{aligned} &(an) + (aa)p + (ab)q + (ac)r + \ldots = 0,\\ &(bn) + (ab)p + (bb)q + (bc)r + \ldots = 0,\\ &(cn) + (ac)p + (bc)q + (cc)r + \ldots = 0,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

$$^tAN + {}^tAAX = 0$$

[1] Gauss formule des remarques à ce sujet à la fin de la Théorie du Mouvement des corps celestes de 1809.et reprend sa méthode dans la deuxième partie de la Théorie de la Combinaison des Observations parue en 1823

L'élimination, très-pénible lorsque le nombre des inconnues est considérable, peut se simplifier notablement de la manière suivante. Outre les coefficients (an), (aa), etc., [dont le nombre est $\frac{1}{2}(i^2+3i)$, si le nombre des inconnues est i], supposons que l'on ait calculé la somme

$$n^2+n'^2+n''^2+\ldots=(nn);$$

on voit facilement que l'on a

$$\begin{aligned}\Omega=&(nn)+2(an)p+2(bn)q+2(cn)r+\ldots\\&+(aa)p^2+2(ab)pq+2(ac)pr+\ldots\\&+(bb)q^2+2(bc)qr+2(bd)qs+\ldots\\&+(cc)r^2+2(cd)rs+\ldots;\end{aligned}$$

et, en désignant

$$(nn)+(aa)p+(ab)q+\ldots$$

par A, tous les termes de $\frac{A^2}{(aa)}$ qui contiennent le facteur p, se trouvent dans l'expression Ω, et, par suite,

$$\Omega-\frac{A^2}{(aa)}$$

est une fonction indépendante de p. C'est pourquoi, en posant

$$\begin{aligned}(nn)-\frac{(an)^2}{(aa)}&=(nn,1),\\(bn)-\frac{(an)(bn)}{(aa)}&=(bn,1),\\(cn)-\frac{(an)(cn)}{(aa)}&=(cn,1),\\&\ldots\ldots\ldots\ldots\\(bb)-\frac{(ab)^2}{(aa)}&=(bb,1),\\(bc)-\frac{(ab)(ac)}{(aa)}&=(bc,1),\\(bd)-\frac{(ab)(ad)}{(aa)}&=(bd,1),\\&\ldots\ldots\ldots\ldots\end{aligned}$$

on aura

$$\begin{aligned}\Omega-\frac{A^2}{(aa)}=&(nn,1)+2(bn,1)q+2(cn,1)r+2(dn,1)s\ldots\\&+(bb,1)q^2+2(bc,1)qr+2(bd,1)qs\\&+(cc,1)r^2+2(cd,1)rs\\&+\ldots\ldots\ldots\end{aligned}$$

nous désignerons cette fonction par Ω'.

$\Omega = \langle AX+N, AX+N\rangle$
$= \|N\|^2 + 2\langle {}^tAN, X\rangle + {}^tX\,{}^tAAX$

* erreur du traducteur : lire:
(bn) - $\frac{(an)(ab)}{(aa)}$
(cn) - $\frac{(an)(ac)}{(aa)}$

De même, en posant

$$(bn,1)+(bb,1)q+(bc,1)r\ \ldots=\mathrm{B},$$

la différence

$$\Omega'-\frac{\mathrm{B}^2}{(bb,1)}$$

sera indépendante de q; nous la représenterons par Ω''.

En posant, de même,

$$(nn,1)-\frac{(bn,1)^2}{(bb,1)}=(nn,2),$$

$$(cn,1)-\frac{(bn,1)(bc,1)}{(bb,1)}=(cn,2),$$

$$(cc,1)-\frac{(bc,1)^2}{(bb,1)}=(cc,2),$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

et

$$(cn,2)+(cc,2)r+(cd,2)s+\ldots=\mathrm{C},$$

la différence

$$\Omega''-\frac{\mathrm{C}^2}{(cc,2)}$$

sera une fonction indépendante de r.

En continuant ainsi, nous formerons une suite d'expressions Ω, Ω', Ω'', etc., dont la dernière sera indépendante des diverses inconnues, et représentée par (nn,μ), si μ désigne le nombre de ces inconnues; nous aurons alors

$$\Omega=\frac{\mathrm{A}^2}{(aa)}+\frac{\mathrm{B}^2}{(bb,1)}+\frac{\mathrm{C}^2}{(cc,2)}+\frac{\mathrm{D}^2}{(dd,3)}+\ldots+(nn,\mu).$$

On prouvera facilement que Ω étant une somme de carrés

$$w^2+w'^2+w''^2+\ldots,$$

et ne pouvant devenir négative, les diviseurs (aa), $(bb,1)$, $(cc,2)$, etc., sont tous positifs. (Nous supprimons, pour abréger, le détail de la démonstration.) D'après cela, la valeur minimum de Ω correspond évidemment aux valeurs des inconnues, pour lesquelles

$$\mathrm{A}=0,\quad \mathrm{B}=0,\quad \mathrm{C}=0,\ldots,$$

et, en commençant à résoudre le système par la dernière équation, qui ne contient qu'une inconnue, on trouvera les valeurs de p, q, r, s, etc., sans avoir aucune élimination à effectuer. La méthode donne, en même temps, la valeur minimum de Ω, qui est (nn,μ).

Les problèmes posés par les arrondis des calculs successifs seront abordés par Moulton en 1913[28], qui donne le premier exemple connu de matrice "mal conditionnée". La première définition du "conditionnement d'une matrice" sera introduite par Todd en 1950[35].

3• La méthode indirecte de GAUSS

Vers 1817, Gauss achève ses travaux d'Astronomie et se consacre à la Géodésie. Son programme de triangulation du Hanovre va l'occuper jusqu'en 1825. Il utilise une chaîne de 26 triangles. Les systèmes deviennent monstrueux, et de plus, il est fatigué par les travaux d'arpentage sur le terrain. Il invente alors une méthode qui supporte les erreurs de calcul et qu'il peut pratiquer, comme il l'écrit à son ami Gerling, le 26 décembre 1823, "à moitié endormi ou en pensant à autre chose".

Elle peut être considérée comme l'ancêtre des méthodes que Southwell a décrites entre 1935 et 1949, et baptisées : **méthodes de relaxation.** Il ne semble pas que ce soit pour leurs qualités relaxantes , mais par analogie avec le problème mécanique qu'il étudie, un cadre chargé de poids:

Les méthodes indirectes de résolution des systèmes linéaires

Nous reprenons la notation matricielle du système : MX + B = 0, avec

$$M = \begin{pmatrix} m_1^1 m_1^2 \dots \dots m_1^n \\ m_2^1 m_2^2 \dots \dots m_2^n \\ \dots \dots \dots \dots \dots \dots \\ m_n^1 m_n^2 \dots \dots m_n^n \end{pmatrix} \qquad X = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \qquad B = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{pmatrix}$$

Pour un vecteur Y donné, on appelle résidu relatif au vecteur Y le vecteur W = MY + B.

Les méthodes consistent en la construction par récurrence des suites $X^{(p)}$ et $W^{(p)} = M X^{(p)} + B$, $X^{(0)}$ étant choisi arbitrairement. Il est clair que , si M est inversible, et si $W^{(p)}$ tend vers le vecteur nul, $X^{(p)}$ a une limite , qui est la solution du système.

Les méthodes de relaxation:

Pour passer de $X^{(p)}$ à $X^{(p+1)}$, il faut choisir une équation, soit la i-ème:

$$\sum_{j=1}^{n} m_i^j x_j + b_i = 0$$

et une composante: la k-ème, qui sera la seule à être modifiée, de sorte que la i-ème composante du résidu $W^{(p+1)}$ soit nulle.

$X^{(p+1)}$ est donc définie par:

$$\forall j \neq k \dots\dots x_j^{(p+1)} = x_j^{(p)}$$

$$x_k^{(p+1)} = \frac{1}{m_i^k}\left[b_i - \sum_{j \neq k} m_i^j x_j^{(p)}\right] = x_k^{(p)} - \frac{w_i^{(p)}}{m_i^k}$$

Il faut évidemment que m_i^k soit non nul. Dans les méthodes que nous allons voir, l'auteur a toujours choisi k=i, ce qui exige des coefficients diagonaux non nuls.

Ces méthodes se différencient donc par le critère de choix de i.

Pour GAUSS[16], i est tel que $\frac{|w_i^{(p)}|}{m_i^i}$ soit maximal

Pour SEIDEL[31], i est tel que $\frac{|w_i^{(p)}|^2}{m_i^i}$ soit maximal

Pour SOUTHWELL[32], i est tel que $|w_i^{(p)}|$ soit maximal.

Les méthodes itératives :

Ce sont des méthodes de relaxation cycliques : à chaque pas, i augmente d'une unité (modulo n). On peut alors exprimer (comme c'est clairement expliqué dans le texte de Nekrasov pour la méthode de Seidel) $X^{(p+1)}$ en fonction de $X^{(p)}$ sous forme matricielle:

si M = T+A, la suite est définie par $TX^{(p+1)} = -AX^{(p)} - B$, ou $X^{(p+1)} = -T^{-1}AX^{(p)} - B$.

Les méthodes diffèrent par le choix de T. Pour Jacobi, T est la diagonale de M, pour la version cyclique de Seidel, T est le triangle inférieur de M.

On voit facilement sur cette expression que si la suite $X^{(p)}$ converge, sa limite est solution du système.

Dans sa lettre à Gerling - que nous présentons intégralement en annexe, comme les autre textes de cette troisième partie inédits ou difficiles à trouver - Gauss donne seulement un exemple d'application de sa méthode à la résolution d'un système numérique de 4 équations normales à 4 inconnues, obtenues à partir de la méthode des moindres carrés .

A la suite des techniques de compensation de la géodésie, la matrice du système a tous ses éléments diagonaux positifs, et opposés à la somme des autres éléments de la même rangée, et le vecteur cherché est supposé très petit.

Gauss part du vecteur nul, et à chaque étape modifie la composante dont l'indice donne un maximum pour la valeur absolue de la modification. Il s'arrête quand cette modification est inférieure à 0,5. Notons que le système est à coefficients entiers, et qu'il arrondit toujours à l'entier le plus proche, qui mesure le millième de seconde. Les vecteurs résidus, disposés en colonne, ont une norme strictement décroissante. Le processus s'arrête forcément en un nombre fini d'étapes. On peut remarquer aussi que Gauss utilise le fait que la somme des composantes du résidu est toujours nulle pour vérifier l'exactitude des calculs. Ceci fait penser aux invariants de boucle de nos algorithmes.

Bien que Gerling publie en 1843 un ouvrage [17] où il détaille cette méthode avec de nombreux exemples, il semble qu'elle ne se diffuse pas dans le milieu mathématique. La plus ancienne référence que j'ai trouvée se trouve dans WHITTAKER et ROBINSON (1924)[39]

4• La méthode de JACOBI (1845)

C'est dans un article de Jacobi que nous trouvons la première description systématique d'une **méthode itérative**. Dans le cadre de ses recherches sur les perturbations séculaires des planètes , il utilise la méthode des moindres carrés, et pour simplifier les calculs, remarque que dans ce cas "les coefficients diagonaux sont prépondérants parce qu'ils sont sommes de carrés , tandis que les autres résultent de l'addition de nombres positifs et négatifs qui s'annulent en partie."

Pour le cas où la diagonale n'est pas assez dominante, Jacobi propose une méthode de transformation qui revient en fait à diagonaliser la matrice, puisqu'elle donne les coefficients hors diagonale aussi petits que l'on veut. "Mais à partir d'un certain point , laissé à l'appréciation du calculateur, il sera avantageux d'introduire la méthode d'approximation. Si on le fait trop tôt, cette méthode elle-même mettra en évidence les coefficients qui compromettent son succès et qui doivent par conséquent être éliminés par de nouvelles transformations. "Il donne également des invariants de contrôle. Il sent bien que l'aspect dominant de la diagonale joue un rôle dans la convergence mais il ne donne pas de justifications de celle-ci.

"La difficulté de résoudre exactement un grand nombre d'équations linéaires (comme c'est souvent le cas) quand on applique la méthode des moindres carrés, a conduit à penser à des méthodes d'approximation. Il en est une qui se présente d'elle même quand, dans les différentes équations, ce nest jamais la même variable qui est affectée d'un coefficient nettement supérieur aux autres. Soient en effet les équations:

$$(00)x + (01)x_1 + (02)x_2 + \ . = (0m)$$
$$(10)x + (11)x_1 + (12)x_2 + \ .... = (1m)$$
$$(20)x + (21)x_1 + (22)x_2 + \ .... = (2m)$$
$$\text{etc.} \qquad \text{etc.} \qquad \text{etc.}$$

dans lesquelles tous les coefficients (ik) sont très petits par rapport à ceux (jj) de la diagonale; on obtiendra une valeur approchèe des inconnues x, x_1, x_2, etc. par:

$$(00)x = (0m),\ (11)x_1 = (1m),\ (22)x_2 = (2m)\ \text{etc...}$$

Si on désigne ces valeurs par a, a_1, a_2 etc..., on obtient leurs premières corrections Δ, Δ_1, Δ_2 etc. par:

$$(00)\ \Delta = -((01)a_1 + (02)a_2 + ...)$$
$$(11)\ \Delta_1 = -((10)a + (12)a_2 + ...)\ \text{etc....}$$

et, si on pose en général:

$$x = a + \Delta + \Delta^2 + \Delta^3 ...$$
$$x_1 = a_1 + \Delta_1 + \Delta_1^2 + \Delta_1^3 ...$$
$$x_2 = a_2 + \Delta_2 + \Delta_2^2 + \Delta_2^3 ... \text{ etc}$$

où les indices supérieurs désignent les corrections successives, de plus en plus petites, on déduira des Δ^{i+1} des Δ^i par les égalités:

$$(00)\Delta^{i+1} = -((01)\ \Delta_1^i + (02)\ \Delta_2^i + ...),$$
$$(11)\Delta_1^{i+1} = -((10)\ \Delta^i + (12)\ \Delta_2^i + ...),$$
$$(22)\Delta_2^{i+1} = -((20)\ \Delta^i + (21)\ \Delta_1^i + (23)\ \Delta_3^i + ...),$$
$$\text{etc.} \qquad \text{etc.}$$

Dans les équations auxquelles conduit la méthode des moindres carrés, les coefficients diagonaux sont effectivement prépondérants parce qu'ils sont des sommes de carrés, tandis que les autres résultent de l'addition de nombres positifs et négatifs qui s'annulent en partie....."

[traduction : Colette BLOCH]

Appelons D la diagonale de la matrice M du système (avec M = D+C).

Puisqu'elle est prépondérante, on obtient une première solution approchée $X^{(o)}$ en résolvant $DX^{(o)} = -N$, puis on ajoute $D^{(1)}$, vérifiant $DD^{(1)} = CX^{(o)}$ On obtient ainsi $X^{(i+1)} = X^{(i)} + D^{(i+1)}$, puis par récurrence, $DX^{(i+1)} = CX^{(i)} - N$.

5• La méthode de SEIDEL (1874)

Seidel est un élève de Jacobi et il effectue beaucoup de calculs pour lui. En particulier il doit résoudre des systèmes à 72 inconnues pour calculer les valeurs les plus probables des log des luminosités des étoiles. Pour accélérer le procédé , il revient à la méthode de Gauss (celle de Pallas, car il semble ignorer l'autre, pourtant publiée par Gerling en 1845) et reprend la quantité W à minimiser.

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n + b_1 = w_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n + b_1 = w_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n + b_n = w_n \end{cases}$$

Première étape de la réduction de Gauss:

$$\Omega = \frac{w_1^2}{a_{11}} + f\,(x_2, x_3, \ldots, x_n)$$

Remplaçons x_1 par x_1+Dx_1, avec $Dx_1 = -\frac{W_1}{a_{11}}$:

W_1 devient $W'_1 = 0$

W_i devient $W'_i = W_i + a_{1i}\,Dx_1$ pour $i \neq 1$

Ω diminue de $\frac{W_1^2}{a_{11}}$ et devient Ω'

Maintenant, choisissons k tel que $\frac{W_k^2}{a_{kk}}$ soit maximal :

$$\Omega' = \frac{w'^2_k}{a_{kk}} + f\,'(x_1, \ldots x_{k-1}, x_{k+1}, \ldots, x_n)$$

En remplaçant x_k par x_k+Dx_k, avec $Dx_k = -\frac{W'_k}{a_{kk}}$:

W'_i devient $W''_i = W'_i + a_{ki}\,Dx_k$ pour $i \neq k$, et Ω diminue de $\frac{W'^2_k}{a_{kk}}$

La variante cyclique de Gauss-Seidel a été étudiée en 1885 par Nekrasov[30], qui la baptise méthode de Seidel, démontre sa convergence et relie la vitesse de convergence aux valeurs propres de la matrice. La traduction du texte de Nekrasov est en annexe (annexe 2), comme celle de Seidel (annexe 3).

6• La méthode de CHOLESKY (1923)

Ce polytechnicien, né en 1875, était chargé de recherches géodésiques. Il a participé en particulier à la triangulation de la Crète. Tué au combat en 1918, il avait imaginé une méthode très performante de résolution des équations normales, que son ami le Commandant Benoit a publiée en 1923.

Soit à résoudre: (1) $AX+K=0$

et soit (2) $X = {}^tAL$ (l'équation corrélative)

(1) et (2) $\Rightarrow$ (3) $A{}^tAL + K = 0$

Cholesky détermine par identification les relations qui permettent de calculer les coefficients de la matrice T triangulaire inférieure telle que $T{}^tT = A{}^tA$.

On peut alors considérer que (3) provient de :

(1') $TY+K = 0$ et

(2') $Y = {}^tTL$

Il reste à résoudre successivement:

(1') $TY+K = 0$ qui donne Y par un système triangulaire

(2') $Y = {}^tTL$ qui donne L

(2) $X = {}^tAL$

Si l'on part de $BL+K = 0$, où B est une matrice symétrique définie positive, il faut déterminer T, et l'on obtient Y par (1'), puis L par (2').

Cholesky semble avoir ignoré les méthodes itératives. On n'en trouve d'ailleurs aucune trace dans les cours de géodésie de l'Ecole Polytechnique (donnés par Callandreau, puis Poincaré à l'époque), alors qu'il y a trente lignes de conseils sur l'art de disposer ses calculs.

On peut y voir une marque du mépris dans lequel ont longtemps été tenues les Mathématiques Appliquées en France. La méthode des moindres carrés y a d'ailleurs été également oubliée jusqu'après la Guerre de 1870 [21]. A ma connaissance, la méthode de Seidel a été publiée en français pour la première fois dans l'Encyclopédie de Molk, traduite de l'Encyclopédie allemande, entre 1911 et 1916.

Les méthodes itératives, comme les Mathématiques Appliquées en général, connaîtront un grand développement aux Etats-Unis pendant la deuxième guerre mondiale. Un de leurs promoteurs est G.E. Forsythe, et c'est un de ses articles [13] qui m'a donné les pistes pour cette dernière partie. Il s'intitule **"Solving linear algebraic equations can be interesting "**.

J'espère vous en avoir convaincu(e) !

Bibliographie

[1] **BENOIT** Ct.,1923 Note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrés à un système d'équations linéaires en nombre inférieur à celui des inconnues(Procédé du Ct Cholesky), *Bulletin Géodésique*, 2, p 67-77.

[2] **BEZOUT** E.,1764.*Histoire de l'Académie Royale des Sciences de Paris.*

[3] **CAUCHY** L.A.,1815.Mémoire sur les fonctions qui ne peuvent obtenir que deux valeurs égales et de signes contraires par suite des transpositions opérées entre les variables qu'elles renferment. *Journal de l'Ecole Polytechnique*, 10,p 29-112. Oeuvres 2°série ,p 191-169.

[4] **CAUCHY** L.A.,1847.Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes Rendus de l'Académie des Sciences de Paris*,25,p 536-538.

[5] **CAYLEY** A.,1858. A memoir on the theory of Matrices. *Philosophical Transactions*,148,p 17-37.

[6] **CHABERT** J.L.,1989. Gauss et la méthode des moindres carrés, *Revue d'Histoire des Sciences*,T42, p 5-26.

[7] **CHABERT** et alii,1993. *Histoire d'algorithmes: du caillou à la puce*, Belin,Paris.

[8] **CRAMER** G., 1750. *Introduction à l'analyse des lignes courbes algébriques.* Genève.

[9] **DAHAN-DALMEDICO** A. et **PEIFFER** J.,1986. *Une histoire des mathématiques.* Points Sciences, Le Seuil, Paris.

[10] **DIOPHANTE** d'Alexandrie, *Les six livres d'arithmétique...* Trad Paul Ver Eecke, Bruges 1926. Réimpression Blanchard Paris, 1959.

[11] **DODGSON** C.L.(Lewis Caroll) ,1867 *.An elementary theory of déterminants.*

[12] **DORIER** J.L.,1990. *Analyse historique de l'emergence des concepts élémentaires d'algèbre linéaire*, Cahier Didirem 7,IREM Paris VII.

[13] **FORSYTHE** G.E.,1953,Solving linear algebraic equations can be intersting. *Bulletin of the American Mathematical Society*, 59,299-329

[14] **GAUSS** C.F.,1810 Disquisitio de elementis ellipticis Palladis. Mémoire présenté le 25 Novembre 1810 à la Soc Roy Sciences de Göttingen. *Oeuvres* VI,1874,3-24

[15] **GAUSS** C.F.,1821 Theoria Combinationis observationum erroribus minimis obnoxiae. 1° partie présentée à la Soc Roy des Sciences de Göttingen en 1821, 2° en 1823. Supplément en 1826. *Oeuvres*,1V,1873-, 1-108. (Trad. J. Bertrand. Méthodes des moindres carrés.Paris 1855)

[16] **GAUSS** C.F.1823. Lettre du 26 Décembre 1823. *Briefwechsel zwischen Carl Friedrich Gauss und Christian Ludwig Gerling.* Berlin 1927

[17] **GERLING** C.L.,1843. *Die Ausleichungs-Rechnungen des practischen Geometrie, oder die Methode der kleinsten Quadrate mit Anschwenchungen für geödatische Aufgaben* (Appendice à Application du Calcul des compensations à la géométrie pratique ,ou la méthode des moindres carrés appliquée aux problèmes géodésiques. Appendice intitulé : Sur l'élimination des équations normales)

[18] **HØYRUP** J.,1990. Algebra and naive geometry. An Investigation of Some Basic Aspects of Old Babylonian Mathematical Thought(II) *Altorientalische Forshungen* 17,1990,2.

[19] **HOUSEHOLDER** A.S.,1953, *Principles of Numerical analysis*, New York

[20] **JACOBI** C.G.,1845. Uber eine neue Auflosungsart der bei der Methode der kleinsten Quadrate vorkommenden linearen Gleichungen. Astronomische Nachrichten, 22, p 297-306,*Gesammelte Werke*, 3, 1884,p 469-478.

[21] **JOZEAU** M.F.,1992. *Comparaison de méthodes géodésiques en France et en Allemagne dans la première moitié du XIXe siècle*, Univ. Paris XIII.

[22] **KLINE** M. 1972. *Mathematical thought from Ancient to Modern Times*, Oxford University Press.

[23] **KNOBLOCH** E., 1980, *Der Begin der Determinantentheorie,Leibnizens...* Hildersheim, Gerstenberg Verlag.

[24] **LAPLACE** P.S.,1772. Recherches sur le calcul intégral et sur le système du monde. Mémoires de l'Académie des Sciences de Paris. *Oeuvres*,vol VIII.

[25] **LEGENDRE** A.M. 1805. *Nouvelles méthodes pour la détermination des orbites des comètes*, Paris.

[26] **MACLAURIN** C.,1748. *A treatise on algebra*, Londres. Trad fr. Paris 1753

[27] **MARTZLOFF** J.C.,1987, *Histoire des mathématiques chinoises*, Masson,Paris.

[28] **MOULTON** F.R.,1913. On the solutions of linear equations having small determinants, *The American Mathematical Monthly*, XX,p 242-249.

[29] **MUIR** T.,1906-1923. *The theory of determinants in the historical order of development*, Londres.

[30] **NEKRASOV** P.A. 1885. Détermination des inconnues par la méthode des moindres carrés quand le nombre d'inconnues est très grand. *Matematiceskij Sbornik* T 12,189-203 (en russe)

[31] **SEIDEL** L.,1874. Ueber ein Verfahren die Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineare Gleichungen uberhaupt, durch successive Annaherung aufsulosen. *Communication à la Section math-physique de L'Académie Royale des Sciences de Berlin, séance du 7 février 1874.*

[32] **SOUTHWELL** R.V.,1940. *Relaxation methods in engineering science, a treatise on approximate computation*, Oxford University Press.

[33] **SPIESSER** M.,1982. *Equations du premier degré. Méthode de fausse position.* IREM de Toulouse.

[34] **TEMPLE** G.,1939 (ou 1949), The general theory of relaxation methods applied to linear systems,*Proceedings of the Royal Society of London*

[35] **TODD**.J.,1949, The condition of a certain matrix, *Proceedings of the Cambridge Philos.Soc.*, vol 46,p 116-118.

[36] **TROPFKE** J.,1902-1903.Geschichte der Elementarmathematik. Leipzig.

[37] **VANDERMONDE** A.T., 1772. Mémoire sur l'élimination. *Histoire de l'Académie Royale des Sciences de Paris.*

[38] **VAN DER WAERDEN** B.L.,1983. *Geometry and algebra in ancient civilisations*,Springer Verlag.

[39] **WHITTAKER** E.T. et **ROBINSON** G.,1924.*The calculus of observations* (a treatise on numerical mathematics).

ANNEXE 1

Lettre de Gauss à Gerling du 26 décembre 1823,

Briefwechsel zwischen Carl Friedriech Gauss und Christian Ludwig Gerling, Berlin, 1927 Traduit par A. Michel-Pajus

Göttingen, le 26 décembre 1823,

Ma lettre est parvenue trop tard à la poste et elle m'est revenue. Aussi je la décachette pour ajouter des indications pratiques relatives à l'élimination. A vrai dire, il y a là beaucoup de petits avantages particuliers que l'on n'apprend qu'à l'usage.

Je prends par exemple vos mesures d'Orber-Reisig[1].

Je prends d'abord [pour la direction de][2] 1 = 0

d'où, avec 1.3, 3 = 77°57'53",107

(je choisis celui-ci car 1.3 a plus de poids que 1.2); donc avec

13	1.2	2 =	26°44' 7",423	2 = 26°44' 6",696
50	2.3	2 =	6, 507	

finalement avec

26	1.4	4 =	136°21'13",481	4 = 136°21'11",641
6	2.4	4 =	8, 529	
78	3.4	4 =	11,268	

Pour améliorer encore l'approximation, j'obtiens, avec

13	1.2	1 = -0",727	1 = -0",855
28	1.3	1 = 0	
26	1.4	1 = -1,840	

Puisque l'on ne s'occupe que de positions relatives, je peux ajouter +0",855 aux quatre valeurs et je pose

1 = 0° 0' 0",000 + a
2 = 26 44 7, 551 + b
3 = 77 57 53,962 + c
4 = 136 21 12,496 + d

Dans la méthode indirecte, il est très avantageux d'ajouter une modification à chaque direction. Vous pouvez vous en convaincre aisément en calculant sur le même exemple sans cette astuce, en outre, vous vous privez de la grande commodité d'avoir toujours un contrôle au moyen de la somme des termes absolus = 0. Maintenant , je

---

[1] Les mesures d'angle données par Gerling (trouvées sur une feuille dans les papiers de Gauss) étaient, sachant que 1 représente Berger Warte, 2 Johannisberg, 3 Taufstein, 4 Milseburg

Répétitions	Angle
13	1.2 = 26°44' 7",423
28	1.3 = 77 57 53,107
26	1.4 =136 21 13,481
50	2.3 = 51 13 46,600
6	2.4 =109 37 1,833
78	3.4 = 58 23 18,161

(Note de Kruger qui a préparé le volume IX des *Werke*)

[2] ajouté par Kruger

forme les équations normales suivant le schéma suivant (en fait, si les termes sont trop nombreux, je sépare les négatifs des positifs[1].):

ab - 1664	ba + 1661	ca + 23940	da - 25610
ac - 23940	bc + 9450	cb - 9450	db + 8672
ad + 25610	bd - 18672	cd - 29094	dc +29094

Les équations de condition sont donc :

0 = + 6 + 67a - 13b - 28c - 26d
0 = - 7558 - 13a + 69b - 50c - 6d
0 = -14604 - 28a - 50b +156c - 78d
0 = +22156 - 26a - 6b - 78c +110d
Somme = 0

Pour l'élimination indirecte, je remarque que, si 3 des quantités a,b,c,d sont prises égales à 0, la quatrième prend la valeur la plus grande quand d est choisie comme quatrième. Naturellement chaque quantité doit être déterminée à partir de sa propre équation, et par conséquent, d à partir de la quatrième. Je pose donc d = -201 et substitue cette valeur. Les termes constants deviennent: +5232, -6352, +1074, +46; le reste est inchangé.

Maintenant c'est au tour de b, je trouve b = +92, je substitue et je trouve les termes absolus : +4036, -4, -3526, -506. Je continue ainsi jusqu'à ce qu'il n'y ait plus rien à corriger. Mais de tout ce calcul je n'écris en réalité que le tableau suivant :

	d =-201	b =+92	a =-60	c =+12	a =+5	b =-2	a =-1
+6	+5232	+4036	+16	-320	+15	+41	-26
-7558	-6352	-4	+776	+176	+111	-27	-14
-14604	+1074	-3526	-1846	+26	-114	-14	+14
+22156	+46	-506	+1054	+118	-12	0	+26

Dans la mesure où je ne pousse le calcul qu'au 2000ème de seconde près, je vois qu'il n'y a maintenant plus rien à corriger. J'en rassemble:

a =-60	b =+92	c =+12	d= -201
+5	-2		
-1			
-56	+90	+12	-201

en ajoutant à chacun la correction +56, j'obtiens:

a = 0	b =+146	c =+68	d =-145

ainsi les valeurs [des directions] sont:

1	0° 0' 0",000
2	26 44 7,697
3	77 57 54,030
4	136 21 12,351

Presque chaque soir, je fais une nouvelle édition du tableau, qu'il est facile d'améliorer n'importe où. Contre la monotonie du travail d'arpentage, c'est toujours une plaisante distraction; on peut aussi voir immédiatement si quelque chose de douteux s'est glissé, ce qui reste à obtenir, etc. Je vous recommande cette méthode comme modèle. Il ne vous arrivera presque plus jamais de pratiquer une élimination directe, du moins quand vous avez plus de deux inconnues. La procédure indirecte peut se faire à moitié endormi, ou en pensant à autre chose.

---

[1](Note de Krüger) l'unité des constantes est le millième de seconde

ANNEXE 2

DETERMINATION DES INCONNUES PAR LA METHODE DES MOINDRES CARRES QUAND LE NOMBRE DES INCONNUES EST TRES GRAND.

P.A. Nekrassov

Extrait de *Matematiceskij Sbornik,* p 189-203,Tome 12, Moscou, 1885.

Traduit du Russe par Jacques SIMON, Professeur de Mathématiques Spéciales.

§ 1. L'astronomie et la géodésie présentent des situations où il faut déterminer par la méthode des moindres carrés un grand nombre d'inconnues. Dans ces cas la résolution par un système normal d'équations servant à définir les inconnues, présente d'énormes difficultés. Ainsi un système normal contient parfois jusqu'à 70 inconnues. Les déterminants au moyen desquels il faut exprimer ces inconnues contiennent chacun 70 lignes et sont constitués d'un nombre énorme de termes exprimé par le produit 1.2.3. ... .70. L'évaluation d'un tel déterminant est impensable sans extrêmes difficultés[1] . Pour contourner ces difficultés, l'astronomie fait un calcul approché des inconnues. Parmi les procédés de ce genre l'un des plus efficaces est la méthodes de Ludwig Seidel, qui consiste à faire le calcul approché successivement des solutions cherchées, et dont la description détaillée est donnée par l'auteur dans son mémoire appelé:"Uber ein Verfahren die Gleichungen ,.... München"

Examinant la méthode de Seidel, à la demande du respecté astronome V.K. Tséraski, qui est concerné par les difficultés évoquées, j'ai remarqué que dans son mémoire, Seidel ne s'interesse pas au problème, important dans ses implications pratiques, de la rapidité d'approximation des solutions. Pour combler cette lacune, je montrerai que dans les cas favorables la méthode de Seidel converge assez vite mais il est des cas où elle converge lentement et même infiniment lentement.

§2. Les fondements de la méthode de Seidel.

Soit x, y, z....t formant m inconnues. Posant:

$$f_i = a_i x + b_i y + c_i z + \ldots + p_i t - q_i$$

Supposons que les équations obtenues à la suite d'observations sont:

$$f1=0,\ f2=0,\ f3=0,\ \ldots\ fm=0,$$

où $m \geq \mu$. Le système normal d'équations déduites de la méthode des moindres carrés, est, comme il est connu:

$$\begin{aligned}
&(aa)x+(ab)y+(ac)z+\ldots.\ +(ap)t-(aq) = 0 \quad (1)\\
&(ba)x+(bb)y+(bc)z+\ldots.+(bp)t-(bq) = 0\\
&(ca)x+(cb)y+(cc)z+\ldots.\ +(cp)t-(cq) = 0\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&(pa)x+(pb)y+(pc)z+\ldots.+(pp)t-(pq) = 0
\end{aligned}$$

---

[1] Souvent grâce à une certaine symétrie et grâce aux observations déduites de la pratique, on peut alléger de tels calculs de déterminants dont dépend la résolution des équations. J'ai indiqué comment procéder dans une communication faite à la Société des Amateurs des Sciences Naturelles et à la Société Mathématique. Je développerai cette méthode dans l'un des articles suivants.

où: $(kl) = k_1l_1+k_2l_2+....+k_ml_m$

Les grandeurs x, y, z, ... t. qui satisfont à l'équation (1), comme il est bien connu, rendent minimum le polynôme:

$$Q = \frac{1}{2}(f^2_1+f^2_2+...+f^2_m) \quad (2)$$

En ordonnant le deuxième membre par rapport à x, on obtient:

$$Q = \frac{1}{2}((aa)x^2+2(ab)xy+2(ac)xz+....+2(ap)xt-2(aq)x) + R \quad (2')$$

où R est un polynôme ne contenant pas x. L'égalité (2') peut encore s'écrire ainsi:

$$Q = \frac{1}{2\,(aa)}\{(aa)\,x + (ab)\,y + (ac)\,z + ...+ (ap)\,t - (aq)\}^2 + P \quad (3)$$

où P est un polynôme ne contenant pas x.

Soit $x_o, ...,t_o$ un système de valeurs particulières remplaçant $x,...,t$, et ne vérifiant pas les équations (1). Alors la grandeur Q correspondant à ces valeurs sera:

$$Q_o = \frac{1}{2\,(aa)}\{(aa)\,x_o + (ab)\,y_o + (ac)\,z_o + ... + (ap)\,t_o - (aq)\}^2 + P_o \quad (3')$$

Soit x_1 une grandeur vérifiant l'équation:

$(aa)x_1+(ab)y_o+(ac)z_o+....+(ap)t_o-(aq) = 0$

Il est clair que pour le système de valeurs particulières $x_1, y_o, ...t_o$ remplaçant $x, y, ..., t$ la grandeur Q prendra une valeur Q_1, égale à Po, et en outre nous aurons: $Q_1 < Qo$.

Ainsi pour les valeurs $x_1, y_o, ...t_o$ la grandeur Q est plus proche de son minimum, que pour les valeurs $x_o, y_o, ...t_o$ des variables $x, y, ...t$.

Soit y_1 une grandeur satisfaisant à l'équation:

$(ba)x_1+(bb)y_1+(bc)z_o+....+(bp)t_o-(bq) = 0$

Par la méthode indiquée, il est facile de se convaincre que Q qui peut se présenter sous la forme:

$$Q = \frac{1}{2\,(bb)}\{(ba)\,x + (bb)\,y + (bc)\,z + ... + (bp)\,t - (bq)\}^2 + L$$

pour les valeurs $x_1, y_1, z_o ...t_o$ des variables $x, y, ...t$ sera plus proche de son minimum que la grandeur Q correspondant aux valeurs $x_1, y_o, ...t_o$.

Poursuivant ainsi par la même méthode on peut successivement remplacer toutes les grandeurs $x_o, y_o, ...t_o$ par $x_1, y_1, ...t_1$. puis de la même façon par $x_2, y_2, ...t_2$ etc...A la limite après un nombre infini d'opérations Q tendra vers son minimum et le

système des valeurs prises par les variables $x,\ldots t$ tendra vers un système de valeurs satisfaisant au système d'équations (1).

§ 3. Expression de l'erreur sur les valeurs approchées obtenues par la méthode de Seidel.

La méthode de Seidel converge nécessairement vers les solutions cherchées du système (1), mais pour évaluer la rapidité de convergence, il est nécessaire d'avoir l'expression générale de l'erreur commise sur les grandeurs approchées par cette méthode. Cette expression à laquelle nous allons nous intéresser manque dans les travaux de Seidel.

Pour évaluer cette expression nous supposons que l'approximation successive des grandeurs se fait dans l'ordre suivant: d'abord successivement $x_1, y_1, \ldots t_1$ puis les grandeurs $x_2, y_2, \ldots t_2$ etc...

Soit $x_n, y_n, \ldots t_n$ un système de grandeurs approchées par la méthode de Seidel après les n opérations décrites. Ensuite et par la même méthode, on obtiendra le système $x_{n+1}, y_{n+1}, \ldots t_{n+1}$.

Ces grandeurs satisfont aux équations:

$$\begin{aligned}
&(aa)x_{n+1} + (ab)y_n + (ac)z_n + \ldots + (ap)t_n - (aq) = 0\\
&(ba)x_{n+1} + (bb)y_{n+1} + (bc)z_n + \ldots + (bp)t_n - (bq) = 0\\
&(ca)x_{n+1} + (cb)y_{n+1} + (cc)z_{n+1} + \ldots + (cp)t_n - (cq) = 0\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&(pa)x_{n+1} + (pb)y_{n+1} + (pc)z_{n+1} + \ldots + (pp)t_{n+1} - (pq) = 0
\end{aligned} \qquad (4)$$

Ces équations peuvent être regardées comme des équations différentielles aux différences finies par rapport à la variable n. Une des fonctions de n, $x_n, y_n, \ldots t_n$ peut facilement être éliminée à l'aide de la dernière des relations (4), il en résultera un système de μ-1 équations avec μ-1 fonctions du premier ordre avec des coefficients constants. La résolution de ce système, qui ne présente pas de difficulté théorique conduit à l'expression cherchée sur les grandeurs $x_n, y_n, \ldots t_n$.

Soient $x', y', z', \ldots, t'$ les grandeurs exactes satisfaisant aux équations (1). Les erreurs correspondant à l'approximation $x_n, y_n, \ldots t_n$ sont:

$$\xi_n = x_n - x',\ \eta_n = y_n - y',\ \zeta_n = z_n - z',\ \ldots, \tau_n = t_n - t'.$$

Ces grandeurs satisfont aux égalités suivantes:

$$\begin{aligned}
&(aa)\xi_{n+1} + (ab)\eta_n + (ac)\zeta_n + \ldots + (ap)\tau_n - (aq) = 0\\
&(ba)\xi_{n+1} + (bb)\eta_{n+1} + (bc)\zeta_n + \ldots + (bp)\tau_n - (bq) = 0\\
&(ca)\xi_{n+1} + (cb)\eta_{n+1} + (cc)\zeta_{n+1} + \ldots + (cp)\tau_n - (cq) = 0\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&(pa)\xi_{n+1} + (pb)\eta_{n+1} + (pc)\zeta_{n+1} + \ldots + (pp)\tau_{n+1} - (pq) = 0
\end{aligned} \qquad (4')$$

On va chercher une intégrale particulière de ce système en posant

$\xi_n = A\alpha^n, \eta_n = AB\alpha^n, \zeta_n = AC\alpha^n, \ldots, \tau_n = AP\alpha^n$.

En introduisant ces grandeurs dans les égalités (4'), on obtient, après simplification:

$$\begin{aligned}
&(aa)\alpha + (ab)B + (ac)C + \ldots + (ap)P = 0\\
&(ba)\alpha + (bb)\alpha B + (bc)C + \ldots + (bp)P = 0\\
&(ca)\alpha + (cb)\alpha B + (cc)\alpha C + \ldots + (cp)P = 0\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&(pa) + (pb)B + (pc)C + \ldots + (pp)P = 0
\end{aligned} \qquad (6)$$

En éliminant les grandeurs B, C, ..., P, on obtient pour un α déterminé l'égalité suivante:

$$\begin{vmatrix} (aa)\alpha & (ab) & (ac) \ldots\ldots & (ap) \\ (ba)\alpha & (bb)\alpha & (bc) \ldots\ldots & (bp) \\ (ca)\alpha & (cb)\alpha & (cc)\alpha \ldots\ldots & (cp) \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \\ (pa) & (pb) & (pc) \ldots\ldots & (pp) \end{vmatrix} = 0 \quad (7)$$

Cette équation de degré μ-1 par rapport à a admet μ-1 racines $a_1,\ldots,a_{\mu-1}$. A chacune d'entre elles correspond un système de grandeurs $B, C, \ldots P$, déterminées à l'aide des égalités (6). Il y a μ-1 intégrales particulières de la forme (5) avec une constante arbitraire A. A partir de ces intégrales particulières il est facile de former l'intégrale générale qui sera:

$$\begin{aligned} \xi_n &= A_1\alpha_1^n + A_2\alpha_2^n + \ldots + A_{\mu-1}\alpha_{\mu-1}^n \\ \eta_n &= A_1B_1\alpha_1^n + A_2B_2\alpha_2^n + \ldots + A_{\mu-1}B_{\mu-1}\alpha_{\mu-1}^n \\ \zeta_n &= A_1C_1\alpha_1^n + A_2C_2\alpha_2^n + \ldots + A_{\mu-1}C_{\mu-1}\alpha_{\mu-1}^n \qquad (8) \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \tau_n &= A_1P_1\alpha_1^n + A_2P_2\alpha_2^n + \ldots + A_{\mu-1}P_{\mu-1}\alpha_{\mu-1}^n \end{aligned}$$

où $A_1,\ldots,A_{\mu-1}$ sont des constantes arbitraires d'intégration, α_i, B_i, C_i, ...,P_i sont des grandeurs définies par les égalités (6), puis α_i satisfait à l'égalité (7). Les constantes arbitraires $A_1,\ldots,A_{\mu-1}$ sont liées aux erreurs initiales $h_o, x_o, \ldots, t_o$ et aux valeurs choisies arbitrairement $y_o,\ldots,t_o$ par les égalités:

$$\begin{aligned} \eta_o &= y_o - y' = A_1B_1 + A_2B_2 + \ldots + A_{\mu-1}B_{\mu-1} \\ \zeta_o &= z_o - z' = A_1C_1 + A_2C_2 + \ldots + A_{\mu-1}C_{\mu-1} \qquad (9) \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \tau_n &= t_o - t' = A_1P_1 + A_2P_2 + \ldots + A_{\mu-1}P_{\mu-1} \end{aligned}$$

Si l'équation (6) a des racines non toutes distinctes, les seconds membres des égalités (8) doivent être modifiées de façon évidente, mais les conclusions ci-dessous restent valables.

§ 4. Rapidité de convergence de la méthode de Seidel

Comme l'erreur sur les grandeurs $x_n, y_n, \ldots t_n$ doit, comme il a été prouvé au §2, tendre vers 0 quand n tend vers + ∞, alors les racines de l'équation (7) doivent avoir un module inférieur à 1. Seulement à cette condition les seconds membres des égalités (8) tendront vers 0 quand n augmente indéfiniment.

Soit α la racine de (7) de plus grand module, nous l'appellerons la racine principale. Par rapport à la racine α, les erreurs ξ_n.. déterminées par les égalités (8) seront $A\alpha^n$, $AB\alpha^n$,$AC\alpha^n$...La grandeur A, dépendant, comme le montrent les équations (9) de $y_o - y', \ldots t_o - t'$, ne peut être donnée à l'avance car $y',\ldots t'$ sont supposées inconnues. Ainsi pour un choix arbitraire de $y_o \ldots t_o$, A ne sera pas nulle en

général et c'est pourquoi les valeurs $A\alpha^n$, $AB\alpha^n$... des erreurs, quand n augmentera, diminueront lentement en comparaison avec les autres erreurs. Dans la méthode de Seidel la convergence sera rapide si la racine principale a un module petit. Mais si la racine principale a un module voisin de 1, ce qui n'est pas rare, alors la convergence vers les solutions de (1) sera lente sauf si, mais c'est peu probable, le choix des valeurs initiales $y_o...t_o$ entraîne $A = 0$.

De façon générale la rapidité de convergence de la méthode de Seidel est entièrement liée à la valeur de la racine principale de l'équation (7). Si toutes les solutions de (7), y compris la principale sont petites la méthode convergera vite, sinon elle convergera lentement. Nous allons donner un exemple de chacun de ces cas.

Exemple I. Soit $m=\mu=3$

$$f_1 = x - 3y + z - 3 = 0$$
$$f_2 = 2x + y - 4z - 10 = 0$$
$$f_3 = x + 1{,}01y + 7{,}1z - 9{,}1 = 0$$

Le système normal en fonctions de ces données sera:

$$6x + 0{,}01y + 0{,}1z - 12{,}1 = 0$$
$$0{,}01x + 11{,}0201y + 0{,}171z - 0{,}191 = 0$$
$$0{,}1x + 0{,}171y + 67{,}41z - 67{,}61 = 0$$

L'équation (7) prendra la forme:

$$4457{,}189646\alpha^2 - 0{,}28657801\alpha + 0{,}000171 = 0$$

Il y aura deux racines imaginaires conjuguées. Le module de chacune d'elles sera $\rho = 0{,}0000196$. La méthode de Seidel convergera très rapidement.
En prenant par exemple $y_o = 10$ et $z_o = 3$, on trouve:

$$x_1 = 1{,}95, \quad y_1 = -0{,}0309888, \quad z_1 = 1{,}000152$$

Ces grandeurs sont déjà relativement proches des solutions exactes qui sont 2, 0 et 1.

Appliquant la méthode de Seidel une deuxième fois on obtiendra les résultats suivants presque exacts:

$$x_2 = 2{,}0000491, \quad y_2 = -0{,}00000240, \quad z_2 = 0{,}99999993$$

Exemple II. Soit $m = \mu = 3$,

$$f_1 = x + y + z - 6 = 0 \qquad 6x + 5y + 4z - 28 = 0$$
$$f_2 = 2x + y + z - 7 = 0 \qquad 5x + 6y + 4z - 29 = 0$$
$$f_3 = x + 2y + z - 8 = 0 \qquad 4x + 4y + 3z - 21 = 0$$

L'équation (7) prendra la forme: $108\alpha^2 - 187\alpha + 80 = 0$

Les racines de cette équation sont: $\alpha_1 = 0{,}959....$ et $\alpha_2 = 0{,}872...$
La racine principale est proche de 1. La convergence sera lente.

Si on pose par exemple $y_o = 3$, $z_o = 4$, on trouve :

$$x_1 = -0{,}5, \quad y_1 = 2{,}58333, \quad z_1 = 4{,}22222$$
$$x_2 = -0{,}300925, \quad y_2 = 2{,}26929, \quad z_2 = 4{,}37557$$
$$x_3 = -0{,}141455, \quad y_3 = 2{,}04095, \quad z_3 = 4{,}47638$$

Ces valeurs sont sensiblement différentes des valeurs exactes: 1, 2, 3.

Le système d'équations $f_1 = 0$, $f_2 = 0$, $f_3 = 0$, ... , $f_m = 0$, (10)

peut être choisi tel que la racine principale de l'équation (7) soit aussi proche que l'on veut de 1. En effet soit $m-\mu+p$ relations linéaires entre les fonctions $f_1,\ldots,f_m$ de la forme : $\lambda_1 f_1+\lambda_2 f_2+\ldots+\lambda_m f_m$, où les coefficients $\lambda_1 \ldots \lambda_m$ sont des constantes. Si $p \geq 1$ le système (10) sera indéterminé. En même temps le système (1) sera aussi indéterminé et en outre on aura:

$$\begin{vmatrix} (aa) & (ab) & (ac) \ldots\ldots & (ap) \\ (ba) & (bb) & (bc) \ldots\ldots & (bp) \\ (ca) & (cb) & (cc) \ldots\ldots & (cp) \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ (pa) & (pb) & (pc) & (pp) \end{vmatrix} = 0 \quad (11)$$

Par suite l'équation (7) aura une racine a égale à 1

Donnons aux grandeurs $a_1\ldots, b_1\ldots p_1\ldots$ un petit accroissement de sorte que le système (1) devienne déterminé. En même temps le déterminant premier membre de (11) sera non nul bien que très petit et l'équation (7) aura une racine très proche de 1. On peut obtenir de tels systèmes avec racine principale de module proche de 1 par d'autres méthodes. On peut faire en sorte que cette racine soit proche de -1, ou soit un nombre imaginaire de module 1. Par suite les systèmes de cette sorte peuvent se rencontrer très souvent. Dans de telles situations qu'on peut observer effectivement, la méthode de Seidel se montre non adéquate.

Remarquons encore que l'équation (7) peut se présenter sous la forme:

$$(aa)(bb)(cc)\ldots(pp)\alpha^{\mu-1}+\ldots\pm(ab)(bc)(cd)\ldots(pa)=0 \qquad (7')$$

Il en résulte:

$$\alpha_1\alpha_2 \ldots \alpha_{\mu-1} = \pm \frac{(ab)(bc)\ldots(pa)}{(aa)(bb)\ldots(pp)}$$

Cette égalité montre que $\mathrm{mod}(\alpha) > K$ (12)

où α est la racine principale de l'équation (7) et

$$K = \mathrm{mod} \sqrt[\mu-1]{\frac{(ab)(bc)(cd)\ldots(pa)}{(aa)(bb)(cc)\ldots(pp)}} \qquad (13)$$

Si la valeur de K n'est pas assez petite, alors il est clair que la méthode de Seidel convergera lentement.

Remarquons enfin que la méthode de Seidel convergera rapidement quand les grandeurs (aa), ...,(pp) seront très grandes par rapport aux autres coefficients liés aux inconnues dans les équations (1). En effet dans ce cas le coefficient du premier terme de l'égalité (7') sera très grand par rapport aux autres termes de cette égalité et par suite toutes les racines de (7') et de (7) seront très petites.

L'ordre de recherche des valeurs approchées par la méthode de Seidel peut être différent de celui qui a été indiqué au §3. Par suite la rapidité de convergence peut visiblement être quelque peu améliorée. Mais dans le cas où la racine principale de (7) est très proche de 1, un tel procédé ne rend pas la méthode de Seidel suffisamment rapide pour être utilement utilisé dans la pratique. Pour prouver cette affirmation je vais examiner un cas où l'ordre de calcul le plus favorable donne une convergence très lente. etc... Les grandeurs $M_1,\ldots M_m$ étant calculées, choisissons la plus grande. Supposons que ce soit:

$$M_3 = \frac{N_3^{\,2}}{2(cc)}$$

De façon évidente la première approximation optimum concernera z_0. L'ayant réalisée, on aura:

$$z_1 = z_0 - \frac{N_3}{(cc)}$$

Supposons que les premiers membres de (1) après remplacement des grandeurs $x,y...,z$ par $x_0,...,t_0$ prennent les valeurs $N'_1,....,N'_\mu$. Ensuite nous poserons:

$$M'_1 = \frac{(N')^2}{2(aa)} \quad , \quad M'_2 = \frac{(N'_2)^2}{2(bb)} \quad , \ldots , \quad M'_\mu = \frac{(N'_\mu)^2}{2(pp)}$$

La deuxième approximation devra concerner celle des grandeurs $x,y,z_1,...,t_0$ qui correspond à la plus grande des grandeurs:

$M'_1,...., M'_\mu$

etc[2] ....En effectuant ainsi les calculs, la méthode de Seidel permettra d'approcher les inconnues un peu plus rapidement, mais le gain en rapidité n'est pas grand. En effet si le module de la racine principale de l'équation (7) est très proche de 1, alors après quelques opérations par la méthode de Seidel, les grandeurs $M_1,...., M_m$ deviennent trop peu significatives et la plus grande est de très peu inférieure à Q, sans tenir compte des erreurs significatives sur les solutions. Nous allons éclaircir cela par un exemple numérique.

Exemple: Soit $m = \mu = 3$ et

$f1 = 10x + y + z - 15 = 0$
$f2 = 20x + 2y + z - 27 = 0$
$f3 = 31x + 3y + 2z - 43 = 0$

Le système normal en fonctions de ces données sera:

$$\begin{aligned} 1461x + 143y + 92z - 2023 &= 0 \\ 143x + 14y + 9z - 198 &= 0 \\ 92x + 9y + 6z - 128 &= 0 \end{aligned} \qquad (14)$$

L'équation (7) prendra la forme: $122724\,\alpha^2 - 241127\,\alpha + 118404 = 0$
Cette équation admet une racine très proche de 1.
Les solutions des équations (14) sont: $x = 1,\ y = 2,\ z = 3.$

Soit $x_o = 0,9.....\ y_o = 1,9.....z_o = 2,9$.
On trouve: $N_1 = -169,6 \quad N_2 = -16,6 \quad N_3 = -10,7$
$M_1 = 9,843997 \quad M_2 = 9,8411 \quad M_3 = 9,5408$

La première approximation doit concerner la grandeur x_o, on aura:

$x_1 = x_o + 0,1160849 = 1,0160849.$

Ensuite nous trouvons: $N'_1 = 0,0000389 \quad N'_2 = 0,0001407 \quad N'_3 = -0,0201892$
$M'_1 = 0 \quad M'_2 = 0,0000000007 \quad M'_3 = 0,000034$

La deuxième approximation doit concerner z_o, après quoi Q deviendra très petit.
A partir de z_o on obtiendra: $z_1 = 2,9033649.$
Ensuite nous trouvons: $N''_1 = 0,3096097 \quad N''_2 = 0,0304248 \quad N''_3 = 0,0000002$
$M''_1 = 0,00003281 \quad M''_2 = 0,00003306 \quad M''_3 = 0$

La troisième approximation concernera y_o, après quoi Q sera très petit.
Pour y_o , on aura: $y_1 - y_o = -0,0021732$. Cette approximation éloigne même la valeur approchée y de sa valeur exacte.
Ensuite nous trouvons: $N'''_1 = -0,0011579 \quad N'''_2 = 0 \quad N'''_3 = -0,0195586$
$M''_1 = 46\ 10^{-11} \quad M'''_2 = 0 \quad M'''_3 = 0,0000318$
On obtiendra alors: $z_2 - z_1 = 0,0032598.$

La petitesse de l'approximation par rapport à l'importance de l'erreur montre qu'il faut procéder à un calcul long pour obtenir une approximation qui n'est que de 0,01.

---

[2] Dans la pratique un tel ordre de calcul possède l' inconvénient de rendre le calcul plus compliqué puisque en plus de l'ordre il convient de calculer une série de valeurs prises par la lettre M.

ANNEXE 3

SUR UN PROCEDE POUR RESOUDRE PAR APPROXIMATIONS SUCCESSIVES LES EQUATIONS (DITES "NORMALES") FOURNIES PAR LA METHODE DES MOINDRES CARRES OU, EN GENERAL, TOUT SYSTEME D'EQUATIONS LINÉAIRES.

par Ludwig SEIDEL
(traduction Colette BLOCH)

Communication à la section math-physique de l'Académie Royale des Sciences, Berlin,
Séance du 7 février 1874.

I

En sus des hypothèses communément admises dans la théorie de l'ajustement des résultats d'observations, supposons que chacun de ces résultats s'exprime par les équations linéaires suivantes en fonction des inconnues x, y, z.....:

$$
\begin{aligned}
& a_1x + b_1y + c_1z \ldots\ldots\ldots + n_1 = 0 \\
(A) \qquad & a_2x + b_2y + c_2z \ldots\ldots\ldots + n_2 = 0 \\
& a_3x + b_3y + c_3z \ldots\ldots\ldots + n_3 = 0 \qquad \text{etc.....}
\end{aligned}
$$

où le nombre des observations (et par conséquent celui des équations A) est plus grand que le nombre des inconnues x, y, z, ..., mais où chaque équation est altérée par les erreurs d'observation (ou pourrait l'être) et où les seconds membres ne seraient donc en général pas nuls, même si on pouvait substituer à x, y, z, ... leurs valeurs exactes.

On suppose également que les poids éventuellement différents des différentes observations sont déjà inclus dans la forme des équations, selon la règle habituelle, si bien qu'on n'a plus aucune raison à priori de s'attendre à une erreur plus grande dans telle équation que dans telle autre, c'est-à-dire à une plus grande valeur absolue de l'expression qui devrait être nulle de par l'équation.

On suppose encore que les hypothèses relatives aux conditions d'observation, qui justifient rationnellement l'application de la "méthode des moindres carrés" , sont adéquates ou du moins qu'on peut les considérer comme telles. Il s'ensuit que le système le plus probable[1] de valeurs de x, y, z, ... sera celui qui vérifie la condition suivante: rendre minimale la somme des carrés des premiers membres des équations (A).

Employons, comme il est d'usage dans cette théorie, les crochets [ ] comme signe de sommation, si bien que, par exemple

$$[aa] = a_1^2 + a_2^2 + a_3^2 + \ldots\ldots$$

$$[ab] = [ba] = a_1b_1 + a_2b_2 + a_3b_3 + \ldots\ldots$$

chaque somme a autant de termes qu'il y a d'observations), alors la somme des carrés s'écrit

$$\begin{aligned} Q = {} & [aa]x^2 + [bb]y^2 + [cc]z^2 + \ldots\ldots \\ & + 2[ab]xy + 2[ac]xz + 2[bc]yz + \ldots\ldots \\ & + 2[an]x + 2[bn]y + 2[cn]z + \ldots\ldots \\ & + [nn] \end{aligned}$$

et elle est minimale quand les valeurs des inconnues vérifient les équations normales suivantes:

$$\text{(B)} \qquad \begin{aligned} & [aa]\,x + [ab]\,y + [ac]\,z + \ldots\ldots + [an] = 0 \\ & [ab]\,x + [bb]\,y + [bc]\,z + \ldots\ldots + [bn] = 0 \\ & [ac]\,x + [cb]\,y + [cc]\,z + \ldots\ldots + [cn] = 0 \qquad \text{etc.....} \end{aligned}$$

Si simple que soit, par sa nature mathématique, le problème de calculer un nombre quelconque d'inconnues à partir d'un même nombre d'équations linéaires, il n'en reste pas moins très pénible quand le nombre des inconnues est important, et dans de tels cas (par exemple la résolution d'un réseau géodésique quelque peu étendu) on se voit réduit à sacrifier le calcul systématique exact, à former des systèmes partiels d'inconnues et à les recoller tant bien que mal après les avoir résolus un par un. J'ignore si on a jamais intégralement calculé un complexe de plus de soixante-dix inconnues. Le nombre de 70 est atteint dans le réseau de triangulation de la Prusse orientale (et de plus dans ce cas les inconnues sont liées par 31 conditions impératives, circonstance qui augmente encore la difficulté de la résolution classique). Personnellement j'ai eu à faire à 72 inconnues pour calculer les valeurs les plus probables des logarithmes des luminosités des étoiles qui intervenaient dans mon réseau photométrique.

La méthode habituelle de résolution, donnée par Gauss, consiste, comme on sait , à exprimer une inconnue en fonction des autres en la tirant de l'équation où elle a pour coefficient une somme de carrés, autrement dit où elle figure sur la diagonale principale du système, ainsi par exemple la valeur de x sera tirée de la première équation dans (B) et portée dans les autres équations, ce qui fournit le premier système transformé des équations normales :

$$\text{(C)} \qquad \begin{aligned} & [bb.1]\,y + [bc.1]\,z + \ldots\ldots + [bn.1] = 0 \\ & [bc.1]\,y + [cc.1]\,z + \ldots\ldots + [cn.1] = 0 \qquad \text{etc.....} \end{aligned}$$

qui partage avec le système (B) d'origine la propriété de symétrie des coefficients par rapport à la diagonale et dans lequel on a posé:

$$[bc.1] = [cb.1] = [bc] - \frac{[ab][ac]}{[aa]}$$

On procède de même à la substitution d'une deuxième inconnue, etc..... et on obtient des systèmes transformés dont chacun contient une inconnue de moins que le précédent jusqu'à la dernière inconnue qui figure seule et chacune des autres est calculée par l'équation qui a servi à son élimination en remontant dans l'ordre inverse.

Jacobi a élaboré un autre procédé et l'a appliqué aux 7 équations établies par LeVerrier pour calculer une partie des irrégularités séculaires du système planétaire d'après Laplace; encore étudiant j'ai eu l'honneur de faire pour lui

les calculs numériques. Par ce procédé on arrive à annuler les plus gros coefficients non situés sur la diagonale, grâce à une substitution linéaire convenable (correspondant à une rotation d'axes orthogonaux) qui introduit, à la place de deux inconnues à fort coefficient, deux nouvelles inconnues, en préservant la symétrie de l'ensemble ; il est vrai que dans la suite du calcul les nouvelles substitutions réintroduisent un coefficient non nul, mais la somme des carrés des coefficients hors diagonale diminue constamment au profit des coefficients diagonaux et, par ce procédé dont la convergence est rigoureusement démontée, on s'approche autant qu'on veut du but final où (en dehors des termes constants) il ne reste que les termes diagonaux qui fournissent immédiatement les dernières inconnues. Cette méthode est parfaitement adaptée aux difficultés particulières du cas auquel Jacobi l'a appliquée (où les coefficients diagonaux étaient eux-mêmes fonctions linéaires de variables complémentaires), mais par ailleurs elle ne me paraît en aucune façon plus avantageuse que la méthode généralement employée, dans les cas simples qui se présentent habituellement ; et je doute quelle ait jamais été appliquée jusqu'à présent à un autre cas.

J'ai ouvert une troisième voie dans mon travail photométrique cité plus haut; le choix de cette méthode était particulièrement indiqué par la forme simple des équations expérimentales. Dans le présent article j'ai le dessein d'exposer et de démontrer la méthode de résolution sous la forme où elle est universellement applicable et de dégager en même temps quelques particularités qui lui sont propres.

2

Supposons qu' on ait pris d'abord pour les inconnues x, y, z, ... un ensemble de valeurs quelconques qui ne vérifie pas encore le système le plus probable des "équations normales" (B) mais qui donne

$$[aa]x + [ab]y + ...[an] = N_1$$

$$[ab]x + [bb]y + ...[bn] = N_2 \quad \text{etc.....}$$

Par identification, la somme des carrés des écarts peut s'écrire:

$$Q = \frac{1}{[aa]}([aa]\,x + [ab]\,y + [ac]\,z + ...+ [an]^2$$
$$+ [bb.1]\,y^2 + [bc.1]\,z^2 + ...$$
$$+ 2[bc.1]\,yz + ...+ 2[bn.1]\,y + 2[cn.1]\,z + ...+ [nn.1]$$
$$= \frac{1}{[aa]}N_1^2 + [bb.1]\,y^2 + ...+ [nn.1]$$

Sous cette forme l'inconnue x ne figure que dans le premier terme
(c'est à dire dans N_1); d'où l'on voit aussitôt qu'on diminuera Q de la quantité
$\frac{1}{[aa]}N_1^2$

et, sans changer les valeurs prises pour y, z, ..., on modifie la valeur de x de telle sorte que l'expression qui valait N_1 devienne nulle.

Ce sera réalisé en corrigeant x de $\Delta x = -\frac{N_1}{[aa]}$

et la valeur ainsi améliorée $x + \Delta x$ est celle qui, pour la première inconnue, s'accorde le mieux aux valeurs choisies pour les autres inconnues et serait alors la meilleure si les valeurs provisoires des autres variables étaient déjà les

vraies valeurs cherchées.

La modification de x, qui remplace N_1 par $N'_1 = 0$ modifie du même coup les valeurs de N_2, N_3..... en

$$N'_2 = N_2 + [ab]\,\Delta x$$

$$N'_3 = N_3 + [ab]\,\Delta x \qquad \text{etc.....}$$

Si, au lieu de garder y, z ..... et de corriger x de $-\frac{N_1}{[aa]}$, on avait gardé x, z ....

et corrigé y de $-\frac{N_2}{[bb]}$, la somme Q aurait diminué de $\frac{N_2^2}{[bb]}$ la somme Q aurait

diminué de $\frac{N_3^2}{[cc]}$ si on avait corrigé z seul (de $-\frac{N_3}{[cc]}$.)

Une deuxième étape va donc consister à corriger y de $\Delta y = -\frac{N'_2}{[bb]}$

$y + \Delta y$ sera la meilleure valeur s'accordant au système de valeurs $x + \Delta x$, z .....

et réduisant Q de $\frac{N'^2_2}{[bb]}$ alors que Q avait déjà diminué de $\frac{N_1^2}{[aa]}$

Cette modification de y remplace le système de valeurs $N'_1 = 0$, N'_2 , N'_3 .....

par N''_1 , $N''_2 = 0$, N''_3 ..... où

$$N''_1 = N'_1 + [ab]\,\Delta y = [ab]\Delta y$$

$$N''_2 = N'_2 + [bb]\,\Delta y = 0$$

$$N''_3 = N'_3 + [bc]\,\Delta y \qquad \text{etc.....}$$

Si maintenant on corrigeait z de telle façon que sa nouvelle valeur $z + \Delta z$ s'accorde le mieux possible aux valeurs $x + \Delta x$, $y + \Delta y$ et aux valeurs initiales des inconnues suivantes, on diminuerait Q de

$\frac{N''^2_3}{[cc]}$;

par contre Q diminuera de $\frac{N''^2_1}{[aa]}$ si on revient à x (car $x + \Delta x$ n'est plus la meilleure

valeur convenant à $y + \Delta y$, z , ...) et qu'on lui apporte une deuxième correction $-\frac{N''_1}{[aa]}$.

Si donc partant d'un système quelconque de valeurs initiales, on les modifie successivement en les prenant dans un ordre arbitraire (et il n'est pas necessaire de parcourir tout le cycle avant de revenir à une inconnue dont la valeur a déjà été corrigée) , tout en prenant soin de déterminer toujours chaque correction de façon à vérifier celle des équations normales où la variable concernée occupe la position "diagonale" , alors on diminue pas à pas la somme des carrés des écarts (et chaque fois d'une quantité immédiatement connue, de la forme

$\frac{N_2}{[aa]}$ ou $[aa]\,\Delta x^2)$, et ceci aussi logtemps qu'on peut la diminuer.

Car les amoindrissements de Q et les corrections à apporter aux inconnues (la dernière de la forme $-\frac{N}{[aa]}$) ne deviennent négligeables que lorsque tous les N similtanément ont décru jusqu'à des valeurs pratiquement nulles.

Quand ce but est atteint, toutes les équations normales sont vérifiées et, par améliorations successives, les inconnues ont pris les valeurs les plus vraissemblables.

Il faut bien remarquer que la preuve de la décroissance de Q et aussi de la convergence de ce processus d' approximation repose entièrement sur la condition que pour chaque variable on calcule les améliorations successives par rapport à l'équation où elle figure en terme diagonal; car si, dans un système quelconque de n équations linéaires à n inconnues, on prétendait partir d'un ensemble de valeurs arbitraire pour les inconnues et apporter des améliorations successives en déterminant la correction pour x à partir d'une quelconque des équations , puis la correction pour y à partir d'une autre, arbitrairement choisie, etc....., - on ne pourrait pas prouver qu'en général on se rapproche indéfiniment de la solution du système - il pourrait bien plutôt arriver (comme on s'en convaincra facilement) que les valeurs successives des inconnues tendent vers l'infini ou oscillent indéfiniment entre deux valeurs distinctes finies.

Mais tout le système , de n équations linéaires à n inconnues peut être mis sous la forme normale

$$
\begin{aligned}
&[aa]\,x + [ab]\,y + [\,ac]\,z + \ldots + [an] = 0 \\
(B)\qquad &[ab]\,x + [bb]\,y + [\,bc]\,z + \ldots + [bn] = 0 \\
&[ac]\,x + [cb]\,y + [\,cc]\,z + \ldots + [cn] = 0 \qquad \text{etc.....}
\end{aligned}
$$

à partir des équations

$$
\begin{aligned}
&a_1x + b_1y + c_1z \ldots\ldots + n_1 = 0 \\
(A)\qquad &a_2x + b_2y + c_2z \ldots\ldots + n_2 = 0 \\
&a_3x + b_3y + c_3z \ldots\ldots + n_3 = 0 \qquad \text{etc.....}
\end{aligned}
$$

et sous la forme (B) on peut le résoudre par notre méthode, comme on le fait pour les équations normales déduites d'observations expérimentales. L'oscillation sans fin, ou la croissance sans borne des valeurs des variables sont, dans les deux cas, exclues a priori, puisqu'il est prouvé que chaque étape rapproche du but (diminuer la somme des carrés) et qu'on cesse de s'en rapprocher de façon appréciable seulement quand toutes les équations (B) sont vérifiées à ε près et que toutes les inconnues ont alors atteint leurs valeurs définitives. La somme Q existe naturellement tout autant pour un système qui doit être résolu exactement que pour le système déduit d'observations empiriques; la seule différence est que dans le premier cas la valeur minimale qu'elle obtient est 0.

---

(1)La supériorité de ce système le plus probable tient à ce que la probabilité pour que les vraies valeurs des inconnues soient dans des intervalles de longueur donée autour de valeurs calculées est plus grande avec c système qu'avec tout autre (pour les mêmes amplitudes d'intervalles). Ce ne parait pas toujours être compris avec la netteté désirable.