Zeitschrift: Commentarii Mathematici Helvetici
Herausgeber: Schweizerische Mathematische Gesellschaft
Band: 30 (1956)

Artikel: Erreurs de chute dans la résolution de systèmes algébriques linéaires.
Autor: Blanc, Ch. / Liniger, W.
DOI: https://doi.org/10.5169/seals-23914

Erreurs de chute dans la résolution de systèmes algébriques linéaires

par Ch. Blanc et W. Liniger, Lausanne[1])

Le développement récent des techniques de calcul numérique a mis en évidence l'importance des erreurs de chute[2]), en particulier dans la résolution de grands systèmes algébriques linéaires. La question a été déjà étudiée à divers points de vue ; voir, par exemple, [4] et [6]. Il convient de faire d'emblée la remarque suivante : une étude des erreurs de chute n'a de sens que si l'on précise très exactement les opérations arithmétiques effectuées ; en particulier, il faut indiquer à quels moments il s'introduit dans les calculs des erreurs de chute élémentaires, par abandon de décimales.

Nous supposerons ici que les calculs sont disposés selon la technique indiquée dans [5] (il s'agit d'une heureuse adaptation de la méthode de Gauss, désignée parfois sous le nom de méthode de Cholesky) ; précisons que lorsqu'on effectue une somme de produits (et cas échéant son quotient par un nombre), le résultat final seul est arrondi, mais non les résultats partiels (c'est ce qui se passe en général si l'on calcule avec une machine de bureau). Nous supposerons enfin que les erreurs de chute élémentaires sont des grandeurs aléatoires, équidistribuées dans l'intervalle $(-\frac{1}{2}\cdot 10^{-m}, \frac{1}{2}\cdot 10^{-m})$ si l'on arrondit le résultat à m décimales. Cette hypothèse correspond très bien à ce qui se passe réellement pour la plupart des systèmes linéaires ; elle n'est pratiquement en défaut que pour des systèmes spécialement construits ; elle peut du reste être testée, comme nous le montrerons.

Il sera avantageux de ne pas adopter d'emblée une hypothèse aussi restrictive sur les erreurs de chute : nous commencerons par admettre que ce sont des variables aléatoires distribuées d'une manière quelconque, avec toutefois un support assez petit. Nous montrerons que les erreurs de

[1]) Institut de mathématiques appliquées de l'Ecole Polytechnique de l'Université de Lausanne. Recherche subventionnée par le Fonds national suisse de la recherche scientifique.

[2]) Rundungsfehler, round off errors.

chute sur les inconnues satisfont à un système linéaire qui ne se distingue du système donné que par les seconds membres. Leur écart-type sera dès lors facile à calculer, ainsi que celui des résidus du système.

§ 1. Résolution du système avec erreurs de chute

Considérons le système linéaire

$$\sum_{j=1}^{n} a_{ij} x_j = c_i \ , \qquad i = 1, \ldots, n \tag{1.1}$$

et sa transformation par les relations (appliquées dans l'ordre convenable)

$$\left.\begin{aligned}
&\text{(a)} \quad b_{i1} = a_{i1} \ , \\
&\text{(b)} \quad b_{ik} = a_{ik} - \sum_{j=1}^{k-1} b_{ij} b_{jk} \ , && i \geqslant k > 1 \ , \\
&\text{(c)} \quad b_{1k} = \frac{1}{b_{11}} a_{1k} \ , && k > 1 \ , \\
&\text{(d)} \quad b_{ik} = \frac{1}{b_{ii}} \left(a_{ik} - \sum_{j=1}^{i-1} b_{ij} b_{jk}\right) \ , && 1 < i < k \ , \\
&\text{(e)} \quad d_1 = \frac{1}{b_{11}} c_1 \ , \\
&\text{(f)} \quad d_i = \frac{1}{b_{ii}} \left(c_i - \sum_{j=1}^{i-1} b_{ij} d_j\right) \ , && i > 1 \ ,
\end{aligned}\right\} \tag{1.2}$$

en un système triangulaire

$$\left.\begin{aligned}
&\text{(a)} \quad x_i + \sum_{j=i+1}^{n} b_{ij} x_j = d_i \ , && i < n \ , \\
&\text{(b)} \quad x_n = d_n \ ;
\end{aligned}\right\} \tag{1.3}$$

la résolution progressive de ce système (1.3) donne les solutions du système (1.1).

Supposons maintenant que l'on modifie ce procédé de résolution de la manière suivante : On continue à se servir des équations (1.2) et (1.3), mais chaque fois que l'on a calculé une des expressions (à l'exception de (1.2.a) et (1.3.b)), on lui ajoute une variable aléatoire dont les propriétés seront précisées ; les calculs suivants utilisent toujours la quantité ainsi modifiée ; désignons alors par β_{ik}, δ_i, ξ_i les quantités obtenues à la place des b_{ik}, d_i et x_i ; les variables introduites à l'occasion des divers

calculs seront désignées par μ_{ik}, ν_i, π_i. Nous avons donc, à la place des relations (1.2) et (1.3) :

$$\left.\begin{array}{lll}
\text{(a)} & \beta_{i1} = a_{i1} \ , & \\
\text{(b)} & \beta_{ik} = a_{ik} - \sum_{j=1}^{k-1} \beta_{ij}\beta_{jk} + \mu_{ik} \ , & i \geqslant k > 1 \ , \\
\text{(c)} & \beta_{1k} = \dfrac{1}{\beta_{11}} a_{1k} + \mu_{1k} \ , & k > 1 \ , \\
\text{(d)} & \beta_{ik} = \dfrac{1}{\beta_{ii}} (a_{ik} - \sum_{j=1}^{i-1} \beta_{ij}\beta_{jk}) + \mu_{ik} \ , & 1 < i < k \ , \\
\text{(e)} & \delta_1 = \dfrac{1}{\beta_{11}} c_1 + \nu_1 \ , & \\
\text{(f)} & \delta_i = \dfrac{1}{\beta_{ii}} (c_i - \sum_{j=1}^{i-1} \beta_{ij}\delta_j) + \nu_i \ , & i > 1
\end{array}\right\} \qquad (1.4)$$

et

$$\left.\begin{array}{lll}
\text{(a)} & \xi_i + \sum_{j=i+1}^{n} \beta_{ij}\xi_j = \delta_i + \pi_i \ , & i < n \ , \\
\text{(b)} & \xi_n = \delta_n \ . &
\end{array}\right\} \qquad (1.5)$$

Nous supposons que les variables aléatoires μ_{ik}, ν_i et π_i sont de moyenne nulle, indépendantes entre elles et de support intérieur à un intervalle $(-\varepsilon, \varepsilon)$. Nous dirons qu'une variable ζ, fonction des μ, ν et π, est $O(\varepsilon^k)$ si son support est intérieur à un intervalle $(-\varepsilon', \varepsilon')$, avec $\varepsilon' = O(\varepsilon^k)$, pour $\varepsilon \to 0$. Nous négligerons dans la suite, dans une somme, une variable $O(\varepsilon^2)$ vis-à-vis d'une variable $O(\varepsilon)$. Nous supposons enfin que ε est assez petit pour que le support de β_{ii} ne contienne pas l'origine (ce qui est du reste une condition pour que la méthode de résolution considérée donne un résultat convenable).

Introduisons les variables aléatoires

$$\left.\begin{array}{ll}
\text{(a)} & \eta_i = \xi_i - x_i \ , \\
\text{(b)} & \varrho_i = \sum_{j=1}^{n} a_{ij}\xi_j - c_i \ ;
\end{array}\right\} \qquad (1.6)$$

les η_i sont les erreurs sur les x_i, les ϱ_i les résidus des équations pour les solutions approchées ; nous nous proposons de déterminer leurs moments d'ordres *un* et *deux*, à partir des moments correspondants des μ, ν, π.

Posons encore

$$\left.\begin{aligned}
&\alpha_{i1} = a_{i1} \\
&\alpha_{ik} = a_{ik} + \mu_{ik} \ , && i \geqslant k > 1 \ , \\
&\alpha_{ik} = a_{ik} + \beta_{ii}\mu_{ik} \ , && i < k \ , \\
&\gamma_i = c_i + \beta_{ii}\nu_i + \sum_{j=1}^{i} \beta_{ij}\pi_j \ , \\
&\tau_i = \delta_i + \pi_i \ , && i < n \ , \\
&\tau_n = \delta_n \ .
\end{aligned}\right\} \qquad (1.7)$$

Les relations (1.4) et (1.5) prennent alors la forme

$$\left.\begin{aligned}
&\text{(a)} \quad \beta_{i1} = \alpha_{i1} \\
&\text{(b)} \quad \beta_{ik} = \alpha_{ik} - \sum_{j=1}^{k-1} \beta_{ij}\beta_{jk} \ , && i \geqslant k > 1 \ , \\
&\text{(c)} \quad \beta_{1k} = \frac{1}{\beta_{11}}\alpha_{1k} \ , && k > 1 \ , \\
&\text{(d)} \quad \beta_{ik} = \frac{1}{\beta_{ii}}\left(\alpha_{ik} - \sum_{j=1}^{i-1} \beta_{ij}\beta_{jk}\right) \ , && 1 < i < k \ , \\
&\text{(e)} \quad \tau_1 = \frac{1}{\beta_{11}}\gamma_1 \ , \\
&\text{(f)} \quad \tau_i = \frac{1}{\beta_{ii}}\left(\gamma_i - \sum_{j=1}^{i-1} \beta_{ij}\tau_j\right) \ , && i > 1
\end{aligned}\right\} \qquad (1.8)$$

et

$$\left.\begin{aligned}
&\text{(a)} \quad \xi_i + \sum_{j=i+1}^{n} \beta_{ij}\xi_j = \tau_i \ , && i < n \ , \\
&\text{(b)} \quad \xi_n = \tau_n \ .
\end{aligned}\right\} \qquad (1.9)$$

Or les relations (1.8) et (1.9) fournissent exactement la solution du système

$$\sum_{j=1}^{n} \alpha_{ij}\xi_j = \gamma_i \ , \qquad i = 1, \ldots, n \ , \qquad (1.10)$$

d'où, en tenant compte de (1.6),

$$\sum_{j=1}^{n} \alpha_{ij}\eta_j = \gamma_i - \sum_{j=1}^{n} \alpha_{ij}x_j \ ;$$

en tenant compte des valeurs tirées des équations (1.7), on a ensuite

$$\sum_{j=1}^{n} \alpha_{ij}\eta_j = -\sum_{j=2}^{i} \mu_{ij}x_j - \beta_{ii}\sum_{j=i+1}^{n} \mu_{ij}x_j + \beta_{ii}\nu_i + \sum_{j=1}^{i} \beta_{ij}\pi_j \ ; \qquad (1.11)$$

en négligeant des variables $O(\varepsilon^2)$, on peut remplacer ici α_{ij} par a_{ij} et β_{ij} par b_{ij}, d'où, avec cette approximation,

$$\varrho_i = \sum_{j=1}^{n} a_{ij}\eta_j = -\sum_{j=2}^{i} \mu_{ij}x_j - b_{ii}\sum_{j=i+1}^{n} \mu_{ij}x_j + b_{ii}\nu_i + \sum_{j=1}^{i} b_{ij}\pi_j \ ; \qquad (1.12)$$

on voit ainsi que les résidus peuvent se calculer linéairement à partir des erreurs de chute élémentaires ; ensuite, les erreurs η_i sur les inconnues se calculent en résolvant un système linéaire obtenu en remplaçant dans (1.1) les seconds membres par les résidus ϱ_i.

Comme on a supposé que les erreurs de chute élémentaires sont de moyenne nulle, on a immédiatement

$$E\,\varrho_i = 0 \ , \qquad (1.13)$$

$$E\,\eta_i = 0 \ . \qquad (1.14)$$

Passons maintenant aux moments d'ordre deux. On a

$$\sum_{j,p} a_{ij}a_{kp}E\eta_j\eta_p = E\,\varrho_i\varrho_k \ , \qquad (1.14)$$

ce qui permet de déterminer les covariances des η_i à partir de celles des ϱ_i. Pour ces dernières, on a simplement, en tenant compte de l'hypothèse de l'indépendance des erreurs élémentaires :

$$E\,\varrho_i\varrho_k = \begin{cases} \displaystyle\sum_{j=1}^{i} b_{ij}b_{kj}E\pi_j^2 \ , & i<k \ , \\ \displaystyle\sum_{j=1}^{i} b_{ij}^2\,E\pi_j^2 + b_{ii}^2\,E\nu_i^2 + \sum_{j=2}^{i} x_j^2\,E\mu_{ij}^2 & \\ \qquad + b_{ii}^2\displaystyle\sum_{j=i+1}^{n} x_j^2\,E\mu_{ij}^2 \ , & i=k<n \ , \\ \displaystyle\sum_{j=1}^{n-1} b_{nj}^2\,E\pi_j^2 + b_{nn}^2\,E\nu_n^2 + \sum_{j=2}^{n} x_j^2\,E\mu_{nj}^2 \ , & i=k=n \ . \end{cases} \qquad (1.15)$$

Précisons maintenant le choix des variables aléatoires μ, ν et π ; nous supposons pour cela que les calculs donnés par les relations (1.2) et (1.3) sont faits, pour chacune de ces relations, avec un même nombre de décimales. Alors les variances figurant dans (1.5) ont des valeurs bien définies ; posons

$$\begin{aligned} E\,\mu_{ik}^2 &= M_1 \ , & i \geqslant k \ , \\ E\,\mu_{ik}^2 &= M_2 \ , & i < k \ , \\ E\,\nu_i^2 &= M_3 \ , & \\ E\,\pi_i^2 &= M_4 \ ; & \end{aligned}$$

alors

$$E\varrho_i\varrho_k = \begin{cases} M_4 \sum_{j=1}^{i} b_{ij}b_{kj} , & i<k , \\ M_4 \sum_{j=1}^{i} b_{ij}^2 + M_3 b_{ii}^2 + M_1 \sum_{j=2}^{i} x_j^2 & \\ \qquad + M_2 b_{ii}^2 \sum_{j=i+1}^{n} x_j^2 , & i=k<n , \\ M_4 \sum_{j=1}^{n-1} b_{nj}^2 + M_3 b_{nn}^2 + M_1 \sum_{j=2}^{n} x_j^2 , & i=k=n . \end{cases} \tag{1.16}$$

Remarquons encore que les valeurs exactes des b_{ik} et des x_j restent pratiquement inconnues ; mais on peut encore, en restant dans les limitet de l'approximation envisagée, les remplacer par les valeurs approchées β_{ik} et ξ_j.

Pratiquement, on peut tirer entre autre de ces expressions le renseignement suivant : si l'on a formé le système triangulaire (1.3) et si l'on connaît l'ordre de grandeur des inconnues, il est possible d'adopter pour la résolution de (1.3) le nombre de décimales qui est le plus favorable ; ce nombre fixe la valeur de M_4 seulement.

§ 2. A propos des hypothèses faites sur les erreurs de chute élémentaires

On peut évidemment construire des systèmes pour lesquels les erreurs de chute élémentaires montrent une distribution très différente de ce qui résulterait de nos hypothèses ; mais ceci ne constitue pas une objection très valable en pratique. Par contre, les considérations précédentes permettent de tester les hypothèses formulées. Supposons en effet que tous les calculs sont faits avec m décimales ; on a alors $M_k = \frac{1}{12}\,10^{-2m}$ $(k = 1, 2, 3, 4)$, et les $E\eta_i^2$ et $E\varrho_i^2$ sont de la forme

$$E\eta_i^2 = P_i^2 \cdot 10^{-2m} , \qquad E\varrho_i^2 = Q_i^2 \cdot 10^{-2m} ,$$

P_i et Q_i étant indépendants de m ; pour un même système, dont nous connaissons la solution exacte, nous pouvons dès lors effectuer les calculs pour diverses valeurs de m, puis calculer les moyennes quadratiques (sur m) E_i et F_i des erreurs et résidus effectifs pris en unités de la dernière décimale, et enfin comparer les E_i et F_i avec les P_i et Q_i. Prenons par exemple le système

$$
\begin{aligned}
304\,x_1 - 264\,x_2 + 96\,x_3 - 16\,x_4 + x_5 &= 121\\
-264\,x_1 + 400\,x_2 - 280\,x_3 + 97\,x_4 - 16\,x_5 &= -63\\
96\,x_1 - 280\,x_2 + 401\,x_3 - 280\,x_4 + 96\,x_5 &= 33\\
-16\,x_1 + 97\,x_2 - 280\,x_3 + 400\,x_4 - 264\,x_5 &= -63\\
x_1 - 16\,x_2 + 96\,x_3 - 264\,x_4 + 304\,x_5 &= 121
\end{aligned}
$$

dont la solution exacte est $x_1 = x_2 = x_3 = x_4 = x_5 = 1$; si on le résout successivement pour $m = 4, \ldots, 8$, on obtient les nombres portés au tableau ci-dessous :

i	P_i	E_i	Q_i	F_i
1	3,288	3,8	215	296
2	4,519	4,8	134	88
3	4,251	4,2	105	121
4	3,186	3,3	89	61
5	1,661	1,7	56	44

On constate qu'il y a une bonne concordance, ce qui justifie les hypothèses formulées sur les erreurs de chute élémentaires[3]).

Il faudra cependant prendre certaines précautions dans les cas où la période de la fraction $1/a_{11}$ est courte ; il peut arriver alors que les erreurs de chute sur les b_{1k} soient distribuées d'une manière très particulière. En outre, si a_{k1} est nul, il n'y a pas d'erreur de chute sur le calcul de b_{k2}, donc $\mu_{k2} = 0$, ce dont il faut tenir compte dans la suite des calculs.

§ 3. Systèmes à solutions aléatoires

Supposons maintenant que les x_i et les c_i du système

$$\sum_{j=1}^{n} a_{ik} x_j = c_i \ , \qquad i = 1, \ldots, n \tag{3.1}$$

sont des variables aléatoires, de moyenne nulle ; désignons par s_{ik} la covariance $E\,x_i x_k$; nous supposons que l'on résout le système par la méthode indiquée au paragraphe 1, en introduisant comme plus haut des grandeurs aléatoires μ, ν et π à chaque calcul ; les considérations du paragraphe 1 restent valables, si l'on suppose, ce que nous ferons, que les μ, ν, π sont stochastiquement indépendants des x_i ; il reste dès lors simplement (en donnant la même signification aux M_k) :

[3]) L'étude d'autres systèmes a donné des résultats tout à fait analogues.

$$E\varrho_i\varrho_k = \begin{cases} M_4 \sum_{j=1}^{i} b_{ij}b_{kj} , & i<k , \\ M_4 \sum_{j=1}^{i} b_{ij}^2 + M_3 b_{ii}^2 + M_1 \sum_{j=2}^{i} s_{jj} + M_2 b_{ii}^2 \sum_{j=i+1}^{n} s_{jj} , & i = k < n , \\ M_4 \sum_{j=1}^{n-1} b_{nj}^2 + M_3 b_{nn}^2 + M_1 \sum_{j=2}^{n} s_{jj} , & i = k = n ; \end{cases} \tag{3.2}$$

on peut en tirer ensuite les $E\eta_i\eta_k$; ces covariances se calculent donc comme si les x_i étaient des grandeurs certaines égales aux s_{ii}. Il en résulte en particulier que ces covariances ne dépendent pas de la covariance des x_i, mais seulement de leurs variances.

Ces considérations trouvent une application dans l'étude des solutions approchées de certains problèmes. On a montré ailleurs (voir [1], [2], [3]) que l'on peut se placer alors à un point de vue stochastique, en considérant les données du problème comme aléatoires, la solution l'étant donc aussi, de même que les erreurs de méthode si l'on adopte une méthode approchée ; si l'on se propose alors l'étude non seulement des erreurs de méthode, mais aussi des erreurs de chute, et si le calcul numérique comporte essentiellement la résolution d'un système algébrique linéaire (comme c'est par exemple le cas lorsqu'on substitue des différences finies à des dérivées dans une équation différentielle linéaire), les considérations précédentes s'appliquent parfaitement.

BIBLIOGRAPHIE

[1] *Ch. Blanc*, Etude stochastique de l'erreur dans un calcul approché, Comment. Math. Helv. 26, 1952, 225–241.

[2] *Ch. Blanc*, Sur les formules d'intégration approchée d'équations différentielles, Arch. Math., 5, 1954, 301–308.

[3] *Ch. Blanc et W. Liniger*, Stochastische Fehlerauswertung bei numerischen Methoden, Z. Angew. Math. Mech., 35, 1955, 121–130.

[4] *H. H. Goldstine et J. v. Neumann*, Numerical Inverting of Matrices of high Order. II. Proc. Am. Math. Soc. 2, 1951, 188–202.

[5] *W. E. Milne*, « Numerical Calculus », Princeton 1949, p. 17 et suiv.

[6] *J. v. Neumann et H. H. Goldstine*, Numerical Inverting of Matrices of high Order, Bull. Am. Math. Soc., 53, 1947, 1027.

(Reçu le 20 juillet 1955.)