Aplikace matematiky

Petr Liebl; Markéta Nováková

A method for dealing with ill-conditioned symmetric linear systems

Aplikace matematiky, Vol. 15 (1970), No. 6, 407--412

Persistent URL: http://dml.cz/dmlcz/103314

A METHOD FOR DEALING WITH ILL-CONDITIONED SYMMETRIC LINEAR SYSTEMS

PETR LIEBL and MARKÉTA NOVÁKOVÁ

(Received October 21, 1969)

1. Let us first introduce the notation used here. Let $\boldsymbol{A}$ be a regular real hermitian matrix of order n with elements a_{ik}. Let $\boldsymbol{A}^{-1} = \boldsymbol{B}$ have elements b_{ik}. Denote the eigenvalues of $\boldsymbol{A}$, arranged by ascending value, by $\lambda_1, \lambda_2, \ldots, \lambda_n$, so that

$$0 < |\lambda_1| \leqq |\lambda_2| \leqq \ldots \leqq |\lambda_n| . \tag{1}$$

Let the corresponding eigenvectors, written for convenience as row-vectors, be $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$, with coordinates $\boldsymbol{v}_i = (v_{i1}, v_{i2}, \ldots, v_{in})$, so that

$$\boldsymbol{v}_i \boldsymbol{A} = \lambda_i \boldsymbol{v}_i . \tag{2}$$

Assume the eigenvectors normalized, so that

$$\sum_{k=1}^{n} v_{ik}^2 = 1 , \quad i = 1, 2, \ldots, n , \tag{3}$$

and orthogonal to each other, which is always possible. Introduce the matrices

$$\Lambda = \begin{pmatrix} \lambda_1, & 0, & & 0 \\ 0, & \lambda_2, & & 0 \\ & & & \\ 0, & 0, & & \lambda_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{V} = \begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \\ \boldsymbol{v}_n \end{pmatrix} = \begin{pmatrix} v_{11}, v_{12}, \ldots, v_{1n} \\ v_{21}, v_{22}, \ldots, v_{2n} \\ \ldots\ldots\ldots\ldots \\ v_{n1}, v_{n2}, \ldots, v_{nn} \end{pmatrix} .$$

Then it is well known that $\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}^{-1}$ and hence

$$\sum_{i=1}^{n} v_{ik}^2 = 1 , \quad k = 1, 2, \ldots, n . \tag{4}$$

Further it holds that $\boldsymbol{A} = \boldsymbol{V}^{\mathrm{T}} \Lambda \boldsymbol{V}$ and $\boldsymbol{B} = \boldsymbol{V}^{\mathrm{T}} \Lambda^{-1} \boldsymbol{V}$. These equalities can be written in elements as follows

$$a_{ik} = \sum_j v_{ji} \lambda_j v_{jk} \tag{5}$$

(6) $$b_{ik} = \sum_j v_{ji}\lambda_j^{-1}v_{jk} .$$

The summations, here as elsewhere, will be assumed to extend from 1 to n, unless explicitly stated otherwise. By a matrix norm, a mapping of the set of matrices into the interval $\langle 0, \infty)$ with certain properties is meant. Of the properties of matrix norms, we shall need the fact that any norm of the square matrix $\boldsymbol{M}$ is not less than the value of any of the eigenvalues of $\boldsymbol{M}$. In this paper, we shall use throughout the "row sum" norm

(7) $$\|\boldsymbol{M}\| = \max_i \sum_k |m_{ik}|$$

where m_{ik} is the element in row i and column k of $\boldsymbol{M}$. There is common agreement that the "ill-conditionedness" of a system of linear algebraic equations can be described, to a certain degree, by the value of a "condition number" of the coefficient matrix. A matrix norm given, the corresponding condition number of a regular matrix $\boldsymbol{M}$ is defined as the product of the norm of $\boldsymbol{M}$ itself and the norm of $\boldsymbol{M}^{-1}$. The condition number of a singular matrix is usually put equal to ∞. From properties of norms it follows that a condition number of any matrix is never less than 1. Using the norm defined by (7), introduce the condition number for any regular matrix $\boldsymbol{M}$

(8) $$C(\boldsymbol{M}) = \|\boldsymbol{M}\| \, \|\boldsymbol{M}^{-1}\| .$$

2. Consider the system of n linear algebraic equations

(9) $$\boldsymbol{A}\boldsymbol{x}^{\mathrm{T}} = \boldsymbol{y}^{\mathrm{T}}$$

where $\boldsymbol{A}$ is the matrix introduced above, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ is a given row-vector, and $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is the row-vector of the unknowns. Assume now the, say, q-th eigenvector $\boldsymbol{v}_q$ and the corresponding eigenvalue λ_q to be known. We may write, using (2) and (9)

$$\boldsymbol{v}_q\boldsymbol{x}^{\mathrm{T}} = \lambda_q^{-1}\boldsymbol{v}_q\boldsymbol{A}\boldsymbol{x}^{\mathrm{T}} = \lambda_q^{-1}(\boldsymbol{v}_q\boldsymbol{y}^{\mathrm{T}})$$

where the leftmost member is a linear combination of the unknowns with known coefficients and the rightmost member a known number. So, we have derived a new equation for the unknowns x_k

(10) $$Kv_{q1}x_1 + Kv_{q2}x_2 + \ldots + Kv_{qn}x_n = \lambda_q^{-1}K\sum_k v_{qk}y_k$$

where $K \neq 0$ is an arbitrary parameter. The idea now is to replace one, the p-th, say, of the original equations in (9) by the equation (10), obtaining the system

(9a) $$\boldsymbol{A}^{(p,q)}\boldsymbol{x}^{\mathrm{T}} = \boldsymbol{y}^{(p,q)\mathrm{T}}$$

where

$$\boldsymbol{A}^{(p,q)} = \begin{bmatrix} a_{11}, & a_{12}, & \ldots, a_{1n} \\ a_{21}, & a_{22}, & \ldots, a_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{p-1,1}, & a_{p-1,2}, & \ldots, a_{p-1,n} \\ Kv_{q1}, & Kv_{q2}, & \ldots, Kv_{qn} \\ a_{p+1,1}, & a_{p+1,2}, & \ldots, a_{p+1,n} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}, & a_{n2}, & \ldots, a_{nn} \end{bmatrix}$$

and

$$\boldsymbol{y}^{(p,q)} = (y_1, y_2, \ldots, y_{p-1}, \lambda_q^{-1} K \sum_k v_{qk} y_k, y_{p+1}, \ldots, y_n) .$$

Obviously, the solution $\boldsymbol{x}$ of (9) is a solution of (9a) as well. When the system (9) is ill-conditioned this is often caused by λ_1 being in value much smaller than the other ones. In this situation it is however rather easy to find λ_1 and $\boldsymbol{v}_1$. It is expected, from heuristic considerations, that by replacing one equation in (9) by (10) for $q = 1$, the "conditionedness" of the system is substantially improved. A theorem showing that this is indeed the case is proved in sec. 4.

3. Denote by $\boldsymbol{A}^{(p,1)}$ the matrix obtained from matrix $\boldsymbol{A}$ by replacing its p-th row by the vector $K\boldsymbol{v}_1$.

$\boldsymbol{A}$ being given, $\boldsymbol{A}^{(p,1)}$ depends on p and K. It is easily verified that $\boldsymbol{A}^{(p,1)} = \boldsymbol{T}\boldsymbol{A}$ where

$$\boldsymbol{T} = \begin{bmatrix} 1, & 0, & \ldots, & 0 \\ 0, & 1, & \ldots, & 0 \\ \dfrac{K}{\lambda_1} v_{11}, & \dfrac{K}{\lambda_1} v_{12}, & \ldots, & \dfrac{K}{\lambda_1} v_{1n} \\ 0, & 0, & \ldots, & 1 \end{bmatrix} \leftarrow p$$

If now

$$v_{1p} \neq 0 , \tag{12}$$

$\boldsymbol{T}^{-1}$ exists and is found to be

$$\boldsymbol{T}^{-1} = \begin{bmatrix} 1, & 0, & \ldots, & 0 \\ 0, & 1, & \ldots, & 0 \\ \ldots & & \ldots & \\ -\dfrac{v_{11}}{v_{1p}}, & -\dfrac{v_{12}}{v_{1p}}, \ldots, & \dfrac{\lambda_1}{Kv_{1p}}, \ldots, & -\dfrac{v_{1n}}{v_{1p}} \\ \ldots & & \ldots & \\ 0, & 0, & \ldots, & 1 \end{bmatrix} \leftarrow p \tag{13}$$

and

$$\mathbf{A}^{(p,1)-1} = \mathbf{A}^{-1}\mathbf{T}^{-1} = \mathbf{B}\mathbf{T}^{-1}. \tag{14}$$

Denoting the elements of $\mathbf{A}^{(p,1)-1}$ by $\tilde{b}_{ik}$, we have from (13) and (14)

$$\tilde{b}_{ik} = b_{ik} - v_{1k}v_{1p}^{-1}b_{ip} \quad \text{for} \quad k \neq p$$

$$\tilde{b}_{ip} = \lambda_1 K^{-1} v_{1p}^{-1} b_{ip}$$

and using further (6)

$$\tilde{b}_{ik} = \sum_j v_{ji}\lambda_j^{-1}(v_{jk} - v_{1k}v_{1p}^{-1}v_{jp}) \quad \text{for} \quad k \neq p \tag{15}$$

$$\tilde{b}_{ip} = \lambda_1 K^{-1} v_{1p}^{-1} \sum_j v_{ji}\lambda_j^{-1} v_{jp}. \tag{16}$$

Note that in (15), the term for $j = 1$ vanishes. We may therefore in (15) assume the summation to run from 2 to n and write

$$\tilde{b}_{ik} = \sum_{j\neq 1} \lambda_j^{-1} v_{ji}v_{jk} - v_{1k}v_{1p}^{-1} \sum_{j\neq 1} \lambda_j^{-1} v_{ji}v_{jp}. \tag{17}$$

4. Our further aim is to find upper bounds for the values of the elements of $(\mathbf{A}^{(p,1)})^{-1}$ and hence an upper bound for $\|(\mathbf{A}^{(p,1)})^{-1}\|$. First we observe, using (3), that for any r, s

$$\sum_j |v_{jr}|\,|v_{js}| \leqq \tfrac{1}{2}\sum_j (v_{jr}^2 + v_{js}^2) = 1. \tag{18}$$

Now we start from (17), use (1) and (18) and obtain

$$\begin{aligned} |\tilde{b}_{ik}| &\leqq \sum_{j\neq 1} |\lambda_j^{-1}|\,|v_{ji}|\,|v_{jk}| + |v_{1k}|\,|v_{1p}^{-1}| \sum_{j\neq 1} |\lambda_j^{-1}|\,|v_{ji}|\,|v_{jp}| \leqq \\ &\leqq |\lambda_2^{-1}| \left(\sum_j |v_{ji}|\,|v_{jk}| + |v_{1k}|\,|v_{1p}^{-1}| \sum_j |v_{ji}|\,|v_{jp}|\right) \leqq \\ &\leqq |\lambda_2^{-1}| \left(1 + |v_{1k}|\,|v_{1p}^{-1}|\right) \quad \text{for} \quad k \neq p. \end{aligned} \tag{19}$$

From (16), using (1) and (18) we have

$$|\tilde{b}_{ip}| \leqq |\lambda_1|\,|K^{-1}|\,|v_{1p}^{-1}|\,|\lambda_1^{-1}| \sum_j |v_{ji}|\,|v_{jp}| \leqq |K^{-1}|\,|v_{1p}^{-1}|. \tag{20}$$

Putting together (19) and (20) we have

$$\sum_k |\tilde{b}_{ik}| \leqq (n-1)\,|\lambda_2^{-1}| + |v_{1p}^{-1}|\,|\lambda_2^{-1}| \sum_{k\neq p} |v_{1k}| + |v_{1p}^{-1}|\,|K^{-1}|. \tag{21}$$

Note that the right hand side of the inequality is independent of i and hence is an upper bound for $\|(\mathbf{A}^{(p,1)})^{-1}\|$ itself. Now everything is ready to prove the

Theorem. *Putting*

$$K = \left(\sum_k |v_{1k}|\right)^{-1} \|\boldsymbol{A}\| \tag{22}$$

and choosing p in such a way that

$$|v_{1p}| \geqq |v_{1k}| \quad \textit{for all } k\,, \tag{23}$$

it holds that

$$C(\boldsymbol{A}^{(p,1)}) < 3n|\lambda_1|\,|\lambda_2^{-1}|\,C(\boldsymbol{A})\,. \tag{24}$$

Proof. Note first that with (23), (12) is satisfied and our considerations are valid. Obviously, from (3) it follows that

$$\sum_{k \neq p} |v_{1k}| < \sum_k |v_{1k}| \leqq \sqrt{n}\,.$$

Also, from (3) and (23) it follows that $|v_{1p}^{-1}| \leqq \sqrt{n}$. According to our remark on norms, $\|\boldsymbol{A}\| \geqq |\lambda_2|$. Using these three facts and (22) and (21), we have

$$\|(\boldsymbol{A}^{(p,1)-1}\| \leqq (3n - 1)\,|\lambda_2^{-1}| < 3n|\lambda_2^{-1}|\,.$$

Now, from (22) it follows that

$$\|\boldsymbol{A}^{(p,1)}\| = \|\boldsymbol{A}\|\,. \tag{25}$$

Finally,

$$C(\boldsymbol{A}^{(p,1)}) = \|\boldsymbol{A}^{(p,1)}\|\,\|(\boldsymbol{A}^{(p,1)})^{-1}\| < \|\boldsymbol{A}\|\,.\,3n|\lambda_2^{-1}| = 3n|\lambda_2^{-1}|\,\|\boldsymbol{A}^{-1}\|^{-1}\,C(\boldsymbol{A})\,.$$

As the eigenvalues of $\boldsymbol{A}^{-1}$ are λ_i^{-1}, $\|\boldsymbol{A}^{-1}\| \geqq |\lambda_1^{-1}|$, and (24) follows.

5. Based on the theorem and on what was said before, an algorithm may be suggested. Assume a system (9) to be given, with $|\lambda_1|$ much smaller than $|\lambda_2|$. First, by backward iteration or any other method, a good approximation to λ_1 and the corresponding eigenvector $\boldsymbol{v}_1$ is found. Then, p is chosen with respect to (23), and K is found according to (22). Finally, in (9), the p-th equation is replaced by the equation (10). The resulting system (9a) has the same solution, $\boldsymbol{x}$, as (9), while the condition number of $\boldsymbol{A}^{(p,1)}$ is at least $|\lambda_1^{-1}\lambda_2|\,(3n)^{-1}$ times smaller than the condition number of $\boldsymbol{A}$. It is apparent that the smaller the ratio $|\lambda_1|\,|\lambda_2|^{-1}$, the easier the evaluation of λ_1 and $\boldsymbol{v}_1$ and the more substantial the improvement in the condition number. It should be noted, however, that one aspect of the illconditioned-ness of the system, the sensitivity of $\boldsymbol{x}$ to small changes in $\boldsymbol{y}$, cannot be changed by any adjustments of the system. In the described method, it is hidden in the sensitivity of the right hand side of (10) to changes in $\boldsymbol{y}$. This suggests that that one number should be evaluated with special attention to accuracy. Also, it should be kept in mind that the accuracy of the result substantially depends on the accuracy in determining λ_1 and $\boldsymbol{v}_1$.

References

[1] *J. W. Head, G. M. Oulton:* The Solution of "Ill-conditioned" Linear Simultaneous Equations. Aircraft Engineering, Vol. XXX, 1958, Math., 309—312.
[2] *Ф. Р. Гантмахер:* Теория матриц. Изд. Наука, Москва 1966.
[3] *J. H. Wilkinson:* Sounding Errors in Algebraic Processes. London, H. M. Stationery Office, 1963.

Souhrn

METODA PRO ZACHÁZENÍ SE ŠPATNĚ PODMÍNĚNÝMI SYMETRICKÝMI SOUSTAVAMI LINEÁRNÍCH ALGEBRAICKÝCH ROVNIC

PETR LIEBL, MARKÉTA NOVÁKOVÁ

V [1] je naznačena metoda, kterou lze „zlepšit" podmíněnost soustavy lineárních rovnic se symetrickou maticí. V matici koeficientů se nahradí jeden řádek vlastním vektorem, příslušejícím k vlastnímu číslu v absolutní hodnotě nejmenšímu.

V této práci je tato metoda popsána, proveden výběr parametrů ve formulích metody a dokázána věta o dosaženém zlepšení.

Authors' addresses: Petr Liebl, Matematický ústav ČSAV v Praze, Žitná 28, Praha 1; *Markéta Nováková,* CSc., Matematicko-fyzikální fakulta KU, Malostranské nám. 25, Praha 1.