On the Separation of Two Matrices

by

J. M. Varah

Technical Report 77-20

December 1977

Department of Computer Science
The University of British Columbia
Vancouver, British Columbia, V6T 1W5

Abstract

The sensitivity of the solution X to the matrix equation $AX - XB = C$ is primarily dependent on the quantity sep(A,B) introduced by Stewart in connection with the resolution of invariant subspaces. In this paper, we discuss some properties of sep(A,B), give some examples to show how very small it can be for seemingly harmless problems, and discuss the feasibility of the iteration $AX^{(k+1)} = X^{(k)}B + C$ for solving the matrix equation.

1. Introduction

We begin with the matrix equation

$$AX - XB = C \tag{1.1}$$

for $A(n\times n)$ and $B(m\times m)$ square matrices, so that X and C are $n\times m$. This equation arises in many applications; for example in the solution of linear elliptic boundary value problems when the unknowns are set up as a matrix X (see Bickley and McNamee [1960], Wan [1973]). Much is known about the problem: there is a unique solution whenever A and B have no eigenvalues in common; see Lancaster [1970] for a discussion of properties and iterative methods for obtaining X, and Bartels and Stewart [1972] for a direct method of solution.

However we are interested in the sensitivity of the solution X to perturbations in A, B, and C. For this, it is illuminating to recast the problem in the form of finding invariant subspaces $\left(\begin{array}{c|c} I & X \\ \hline 0 & I \end{array}\right)$ for the block matrix $\left(\begin{array}{c|c} A & -C \\ \hline 0 & B \end{array}\right)$. Then the results of Stewart [1973] apply: his Theorem 4.1 shows that the sensitivity of X is inversely proportional to the separation between A and B,

$$\operatorname{sep}(A,B) = \min_{||P||=1} ||AP - PB|| \tag{1.2}$$

It is this quantity we wish to discuss here, in particular with the Frobenius norm $||Z||_F^2 = \operatorname{tr}(Z^*Z)$.

Of course, (1.1) can also be recast as a linear system

$$\begin{pmatrix} A - b_{11}I & -b_{21}I & \cdots & -b_{m1}I \\ -b_{12}I & A - b_{22}I & & \vdots \\ \vdots & & \ddots & \vdots \\ -b_{1m}I & \cdots & \cdots & A - b_{mm}I \end{pmatrix} \begin{pmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_m \end{pmatrix} = \begin{pmatrix} \underline{c}_1 \\ \vdots \\ \underline{c}_m \end{pmatrix} \tag{1.3}$$

where $\underline{x}_i$ and $\underline{c}_i$ are the columns of X and C. The matrix is of course the Kronecker sum of A and -B,

$$T = I \otimes A - B^T \otimes I.$$

Seen in this light, the sensitivity of X should be proportional to the condition number $\kappa(T)$; however since

$$\sigma_{\min}(T) = \min_{||x||_2 = 1} ||T\underline{x}||_2 = \min_{||P||_F = 1} ||AP-PB||_F = \text{sep}_F(A,B),$$

the two are equivalent if we scale A and B so $\sigma_{\max}(T) = 1$. (Here $\sigma_{\min}(T)$, $\sigma_{\max}(T)$ denote the smallest and largest singular values of T.)

In the next section, we discuss some properties of sep(A,B) and show with some examples how incredibly small this quantity can be for non-normal matrices. In Section 3, we relate it to the perturbation required to give equal eigenvalues in A and B. Then in Section 4, we discuss an iterative method for solving (1.1) which is useful for some applications.

2. Properties of Sep(A,B)

For A and B normal, Stewart [1973] shows that $\text{sep}_F(A,B) = \min_{i,j} |\lambda_i(A)-\lambda_j(B)|$, the minimum distance between the eigenvalues of A and B. However, for A or B non-normal, the separation can be much smaller than this. When B is one-dimensional (B = the scalar b),

$$\mathrm{sep}_F(A,B) = \min_{\|\underline{x}\|_2=1} \|(A-bI)\underline{x}\|_2 = \sigma_{\min}(A-bI),$$

which was used in Varah [1971] to measure the sensitivity of the eigenvector associated with b in the augmented matrix $\left(\begin{array}{c|c} A & -c \\ \hline 0 & b \end{array}\right)$. At this point it is interesting to relate this to the quantity s_b commonly used to measure the sensitivity of the eigenvalue b (see Wilkinson [1965, page 68]). The augmented matrix has $v_b = \begin{pmatrix} x \\ \hline 1 \end{pmatrix}$ and $u_b^T = (0|1)$ as the right and left eigenvectors corresponding to the eigenvalue b, where $\underline{x}$ is the solution to $A\underline{x} - b\underline{x} = \underline{c}$. Thus

$$s_b^2 = \cos^2(u_b, v_b) = \frac{1}{1+\|x\|_2^2},$$

whereas

$$\mathrm{sep}_F(A,B) = \sigma_{\min}(A-bI) = \|(A-bI)^{-1}\|_2^{-1}.$$

Hence s_b depends on $\underline{c}$ but $\mathrm{sep}_F(A,B)$ does not. However they are certainly related: in some sense $\mathrm{sep}_F(A,B)$ gives the smallest possible s_b over all vectors $\underline{c}$ of norm one. We have

$$\frac{1}{s_b^2} - 1 = \|x\|_2^2 \le \|(A-bI)^{-1}\|_2^2 \|c\|_2^2 = \frac{\|c\|_2^2}{[\mathrm{sep}_F(A,B)]^2}$$

and this is an equality for certain vectors $\underline{c}$.

For general non-normal matrices A and B, we feel it is extremely important to realize that sep(A,B) can be very small even though the eigenvalues of A and B are well separated.

Example 1:

$$A = \begin{pmatrix} 1 & -1 & & \\ & 1 & -1 & \\ & & \ddots & \ddots \\ & & & -1 \\ & & & 1 \end{pmatrix}_{n\times n} , \quad B = \begin{pmatrix} 1-\alpha & 1 & & \\ & 1-\alpha & 1 & \\ & & \ddots & \ddots \\ & & & 1 \\ & & & 1-\alpha \end{pmatrix}_{m\times m} .$$

We claim $\mathrm{sep}_F(A,B) = O(\alpha^{m+n-1})$ as $\alpha \to 0$. To see this, first form the matrix T of (1.3):

$$T = \begin{pmatrix} J_n(\alpha) & & & \\ -I & J_n(\alpha) & & \\ & \ddots & \ddots & \\ & & -I & J_n(\alpha) \end{pmatrix}, \quad J_n(\alpha) = \begin{pmatrix} \alpha & -1 & & \\ & \alpha & -1 & \\ & & \ddots & \ddots \\ & & & -1 \\ & & & \alpha \end{pmatrix},$$

where I and $J_n(\alpha)$ are m×m. Since $\mathrm{sep}_F(A,B) = \sigma_{min}(T)$, we need to exhibit a vector $\underline{x} = (\underline{x}_1 , \ldots , \underline{x}_m)^T$ with $\dfrac{||T\underline{x}||_2}{||\underline{x}||_2} = O(\alpha^{n+m-1})$. Take $\underline{x}_1 = (\alpha^{m-1}, \alpha^m, \ldots, \alpha^{m+n-2})^T$; it is easy to see $||\, J_n(\alpha)\underline{x}_1|| = O(\alpha^{n+m-1})$. Now solve for $\underline{x}_2, \underline{x}_3, \ldots, \underline{x}_m$ using the block lower triangular nature of T: i.e., solve $J_n(\alpha)\underline{x}_k = \underline{x}_{k-1}$, $k = 2, \ldots, m$. We obtain

$$\underline{x}_k = (p_{1k}{}^{m-k}, p_{2k}{}^{m-k+1}, \ldots, p_{nk}{}^{m-k+n-1})^T$$

where $P = (p_{ij})$ is the Pascal triangle matrix

$$P = \begin{pmatrix} 1 & n & \cdot & \cdot & & & \\ 1 & \cdot & \cdot & \cdot & & & \\ 1 & 3 & 6 & 10 & \cdots & & \\ 1 & 2 & 3 & 4 & \cdots & & \\ 1 & 1 & 1 & 1 & \cdots & \cdots & 1 \end{pmatrix} .$$

Thus $||\underline{x}||_2 = O(\alpha^\circ)$ and $||T\underline{x}||_2 = O(\alpha^{n+m-1})$. So $\text{sep}_F(A,B)$ can be very small even for moderate sized α: we computed $\text{sep}_F(A,B)$ for several values of m, n, and α, and show some results in Table 1.

Table 1

n	m	α	sep
4	4	1/2	$3.4\ 10^{-4}$
6	3	1/4	$7.0\ 10^{-7}$
6	4	1/8	1.3×10^{-10}
6	6	1/16	2.2×10^{-16}

Example 2: Some matrices of order 12.

We first considered the Frank matrices F_n, defined by

$$\begin{aligned}(F_n)_{ij} &= n + 1 - \max(i,j), \text{ if } j \geq i - 1 \\ &= 0, \text{ otherwise.}\end{aligned}$$

These are well-known to have ill-conditioned eigenvalues, and have been used often as test matrices (see for example Golub and Wilkinson [1976]). We first used the QR method to put F_{12} into upper triangular form, with the computed eigenvalues (all real and positive) arranged in decreasing order. Then we took A as the first k rows and columns, and B as the last (12-k) rows and columns, and computed $\text{sep}_F(A,B)$. These are given in Table 2.

Table 2

k	sep	k	sep	k	sep
1	9.2	5	0.24	9	5.7×10^{-7}
2	4.2	6	6.6×10^{-3}	10	2.3×10^{-7}
3	3.4	7	1.0×10^{-4}	11	3.2×10^{-7}
4	1.7	8	4.4×10^{-6}		

Thus, as is well known, the invariant subspace corresponding to the smallest few eigenvalues is not well determined (see Wilkinson [1963, page 153] for the corresponding condition numbers s_i of the eigenvalues).

What is surprising (to this author at least) is that although this behaviour appears pathological, it is not; this amount of ill-condition is to be expected in non-normal matrices of this order. We generated upper triangualr matrices of order 12 with elements chosen randomly from (0,1), and took A and B as above. The results are in Table 3.

Table 3

k	sep	k	sep	k	sep
1	$3.0\ 10^{-5}$	5	$2.3\ 10^{-8}$	9	$3.1\ 10^{-8}$
2	$5.2\ 10^{-5}$	6	$2.0\ 10^{-8}$	10	$1.4\ 10^{-6}$
3	$8.9\ 10^{-5}$	7	$4.9\ 10^{-8}$	11	$5.4\ 10^{-7}$
4	$1.0\ 10^{-6}$	8	$2.9\ 10^{-8}$		

When we took the diagonal elements from (0,1) and the upper triangular elements from (-2,-1), the results were even more remarkable: the separations were all less than 10^{-8}, and some were less than 10^{-12}. Similar results were obtained when the diagonal elements were fixed at $i/12$, $i=1,\ldots,12$.

The conclusion is clear: the invariant subspaces of non-normal matrices are incredibly ill-conditioned in general, even for moderate-sized matrices; they can only be resolved accurately using extended precision arithmetic. We feel strongly that these separations should be calculated whenever one is attempting to resolve invariant subspaces of non-normal matrices.

3. Spectrum Overlap

Give matrices A and B, it is also of interest to measure the amount required to perturb A and/or B so they have a common eigenvalue; this is discussed in Golub and Wilkinson [1976]. Towards this end (and for other reasons) it is useful to make the following definition:

Definition 3.1: The ε-spectrum of A is the region

$$S_\varepsilon(A) = \{\lambda\varepsilon\mathbb{C} \mid \sigma_{\min}(A-\lambda I)\le\varepsilon\}$$

For A normal, this consists of circles of radius ε around each of the eigenvalues of A. For A non-normal, this is a more complicated region of the complex plane. For ε small, this region gives the values λ where $(A-\lambda I)$ is nearly singular; indeed if $\lambda\varepsilon S_\varepsilon(A)$, there is a matrix E with $||E||_2 \le \varepsilon$ so that $(A-\lambda I+E)$ is singular. (Take $E = -\sigma_{\min} uv^T$, where u and v are the proper left and right singular vectors.)

Returning to spectrum overlap, suppose the -spectra of A and B overlap at λ; then λ is an eigenvalue of $A+E_1$ and $B+E_2$, with $||E_1||_2 \le \varepsilon$, $||E_2||_2 \le \varepsilon$. This motivates

Definition 3.2: The spectrum separation of A and B,

$$\mathrm{sep}_\lambda(A,B) = \min_{\varepsilon_1,\varepsilon_2}\{\varepsilon_1+\varepsilon_2 \mid S_{\varepsilon_1}(A)\cap S_{\varepsilon_2}(B)\ne\phi\}$$

This is related to $\mathrm{sep}_F(A,B)$ as follows:

Theorem 3.1: $\mathrm{sep}_F(A,B) \le \mathrm{sep}_\lambda(A,B)$.

Proof: Let ε_1 and ε_2 give the minimum values in Definition 3.2. Thus there is some $\lambda\varepsilon S_{\varepsilon_1}(A)\cap S_{\varepsilon_2}(B)$. Now $\lambda\varepsilon S_{\varepsilon_1}(A)$ means there is a vector v, $||v||_2 = 1$, with $||(A-\lambda I)v||_2 \le \varepsilon_1$; similarly there is a vector u, $||u||=1$, with $||u^*(B-\lambda I)||_2 \le \varepsilon_2$.

Now take $P = vu^*$; $||P||_F^2 = tr(P^*P) = ||u||_2^2||v||_2^2 = 1$, and

$$\begin{aligned} AP - PB &= Avu^* - vu^*B \\ &= (A-\lambda I)vu^* - vu^*(B-\lambda I) \\ &= w_1u^* - vw_2^* \quad \text{(say)}. \end{aligned}$$

Thus

$$\begin{aligned} ||AP - PB||_F &\le ||w_1u^*||_F + ||vw_2^*||_F \\ &= ||w_1||_2||u||_2 + ||v||_2||w_2||_2 \\ &= \varepsilon_1 + \varepsilon_2. \end{aligned}$$

Hence from (1.2), $sep_F(A,B) \le \varepsilon_1 + \varepsilon_2$. QED.

However, this is about as much as can be said in general relating the two separations: $sep_F(A,B)$ may be very much smaller than $sep_\lambda(A,B)$, if the corresponding matrix P is not of the form uv^T, so there are no corresponding nearly null vectors.

Another way to see this is through the matrix T of (1.3). If $sep_\lambda(A,B) = \varepsilon_1 + \varepsilon_2$, then there are perturbation matrices E_1, E_2 (with $||E_1||=\varepsilon_1$, $||E_2||_2 = \varepsilon_2$) so that $A+E_1$ and $B+E_2$ have a common eigenvalue. Thus the Kronecker sum

$$I \otimes (A+E_1) - (B+E_2) \otimes I = T + (I \otimes E_1 - E_2 \otimes I)$$

is singular; however this is a very <u>special</u> perturbation of T; there could easily be a more general perturbation (T+E) which is singular, with $||E||_2 \ll \varepsilon_1 + \varepsilon_2$ and this would imply $sep_F(A,B) \ll \varepsilon_1 + \varepsilon_2$.

This points out another characterization of $sep_\lambda(A,B)$.

Theorem 3.2:

$$\text{Let } \eta = \min_{\substack{||P||_F=1 \\ P=uv^*}} ||AP-PB||_F = \min_{||\alpha||_2 ||v||_2=1} ||T(\bar{v} \otimes u)||_2 .$$

Then $\eta \le \mathrm{sep}_\lambda(A,B) \le \eta\sqrt{2}$.

Proof: The first inequality follows directly from the proof of Theorem 3.1. To see the other, take vectors u,v with $||u||_2 = ||v||_2 = 1$, and form $P=uv^*$; then

$$\begin{aligned} AP - PB &= Auv^* - uv^*B \\ &= (A-\lambda I)uv^* - uv^*(B-\lambda I) \end{aligned}$$

for any λ. Using $x = (A-\lambda I)u$ and $y^* = v^*(B-\lambda I)$, we have

$$\delta^2(u,v) = ||AP-PB||_F^2 = ||x||_2^2 ||v||_2^2 + ||y||_2^2 ||u||_2^2 - 2\mathrm{Re}[(u^*x)(v^*y)].$$

Now take $\lambda = u^*Au/u^*u$ so $u^*x = 0$. Then

$$\delta^2(u,v) = ||(A-\lambda I)u||_2^2 + ||v^*(B-\lambda I)||_2^2 = \varepsilon_1^2 + \varepsilon_2^2 .$$

Thus we have exhibited λ so that $\sigma_1(A-\lambda I) \le \varepsilon_1$ and $\sigma_1(B-\lambda I) < \varepsilon_2$. Hence

$$\mathrm{sep}_\lambda(A,B) \le \varepsilon_1 + \varepsilon_2 \le \sqrt{2}\,\delta(u,v).$$

Since this holds for all possible u,v, it holds for $\delta(u,v) = \eta$. QED.

4. An Iteration for X

Given the problem (1.1), a rather obvious iteration is

$$AX^{(k+1)} = X^{(k)}B + C \tag{4.1}$$

studied by Lancaster [1970] and others. Once can also include a shift (μI) in A and B. In the light of our discussion earlier, it is of interest to express this iteration in terms of an iteration for the linear system

$$T\underline{x} \equiv (I \otimes A - B^T \otimes I)\underline{x} = \underline{c}. \tag{4.2}$$

Indeed, it is clear that (4.1) is equivalent to solving (4.2) by the linear iteration

$$M\underline{x}^{(k+1)} = N\underline{x}^{(k)} + \underline{c}$$

using the splitting

$$T = M-N = I \otimes A - B^T \otimes I \;.$$

Thus the convergence rate is determined by the spectral radius

$$\begin{aligned}\rho(M^{-1}N) &= \rho((I \times A)^{-1}(B^T \times I) \\ &= \rho(B^T \times A^{-1})\end{aligned}$$

and this last matrix has eigenvalues $\{b_i/a_j,\ i=1,\ldots,m,\ j=1,\ldots,n\}$, where $\{b_i\}$ and $\{a_j\}$ are the eigenvalues of B and A. So the iteration converges if

$$\max|b_i| < \min|a_j|.$$

If we include a general shift μ, this condition means there is a circle with centre μ which includes all the $\{b_i\}$, but excludes all of the $\{a_j\}$. An equivalent condition is given in Lancaster [1970].

This, of course, is a much stronger condition than sep(A,B) > 0; however sep(A,B) can be very small and the iteration can still converge. In this case, the convergence rate is not affected, only the limiting accuracy of the $X^{(k)}$.

This iteration may in fact be useful in some cases of practical interest; in particular in separating blocks occurring in singular perturbation problems in ordinary differential equations, where A has eigenvalues of order ε^{-1} and B eigenvalues of order 1.

References

R. Bartels and G. W. Stewart (1972), Algorithm 432. CACM 15, pp 820-826.

W. G. Bickley and J. McNamee (1960), Matrix and other direct methods for the solution of linear difference equations. Phil. Trans. Roy. Soc. (London). Series A, 252, pp. 69-131.

G. Golub and J. Wilkinson (1976), Ill-conditioned eigensystems and the computation of the Jordan canomical form. SIAM Review.

P. Lancaster (1970), Explicit solution of linear matrix equations. SIAM Review 12, pp 544-566.

G. W. Stewart (1973), Error and perturbation bounds associated with certain eigenvalue problems. SIAM Review 15, pp 727-764.

J. M. Varah (1971), Invariant subspace perturbations for a non-normal matrix. IFIP 71 Proceedings (North-Holland), pp 1251-1253.

F. Wan (1973), An in-core finite difference method for separable boundary value problems on a rectangle. Studies in Appl. Math. LII, pp 103-113.

J. Wilkinson (1963), Rounding Errors in Algebraic Processes. Prentice-Hall, New York.

J. Wilkinson (1965), The Algebraic Eigenvalue Problem. Clarendon Press, Oxford.