On Pre-Conditioning Matrices

E. E. Osborne

Space Technology Laboratories

Los Angeles

Summary

That which follows is concerned with improving the condition of a matrix for the purpose of obtaining its eigenvalues and eigenvectors.

Some of the difficulties encountered in the eigenvalue problem are essentially due to the fact that the eigenvalues of the given matrix are small compared to the norm of the matrix. Thus it appears that there is something to be gained by applying similarity transformations to the matrix so as to reduce its norm.

The notation employed in Householder's [1] paper is used here. Thus, let A be an n^{th} order matrix with complex elements and $\lambda_i(A)$ $(i=1,2,\ldots,n)$ its eigenvalues. Use is made of the Euclidean vector norm

$$||x|| = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2} \tag{1}$$

Two matrix norms are considered. One is the Euclidean norm

$$||A||_E = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2} . \tag{2}$$

The other norm is the spectral norm

$$||A||_s = \left(\rho(AA^*)\right)^{1/2} , \tag{3}$$

where $\rho(A)$ is the spectral radius

$$\rho(A) = \max_i |\lambda_i(A)| . \tag{4}$$

Because of the relation

$$1/\sqrt{n}\,||A||_E \leqslant ||A||_s \leqslant ||A||_E , \tag{5}$$

a substantial reduction of either norm will involve a reduction of the other.

The following theorem is implied by Theorem 4.4 of Householder's paper [1].

Theorem 1. Given $\epsilon > 0$, there exists a unitary matrix U and a non-singular diagonal matrix D such that if P = UD, then $||P^{-1}AP||_s = \rho(A) + \epsilon$. In general one cannot take $\epsilon = 0$.

Theorem 1 does not indicate how the matrix P = UD is to be found. For this reason the use of elementary matrices is investigated. Consider matrices T_e obtained from the identity by replacing one of its zero elements by a scalar.

Theorem 2. There exists a sequence of matrices T_e, (e=1,2,...,) such that if $A_e = T_e^{-1}A_{e-1}T_e$, $A_o = A$, them $\lim_{e\to\infty} ||A_e||_s = \rho(A)$.

It turns out that the associated sequence $||A_e||_s$, (e=1,2,...) is not necessarily monotone decreasing, however. Also, the use of matrices T_e will reduce the norm by additive cancellation and therefore is likely to cause loss of accuracy.

The use of diagonal matrices on the other hand does not lead to much loss of accuracy. The process of applying similarity transformations induced by diagonal matrices is referred to as "scaling" in the remainder of this paper. Scaling is clearly desirable in case the computations are based on fixed point arithmetic. In case floating point arithmetic is used, scaling appears to affect results in two ways. The first is in the formation of inner products. One consequence of this is in the application of certain convergence criteria such as

$$||Ax - \lambda x||/||x|| < \epsilon \qquad (6)$$

which is used to see if λ is an eigenvalue of A having x as its associated eigenvector. By introducing bad scaling, the author has succeeded in causing the failure of computer programs using this test.

The second way in which scaling can affect results is in making decisions based on the relative size of numbers. Such decisions enter into the selection of pivotal elements. It is surprising that in many cases bad scaling does not seem to impair the results. However, examples can be given in which bad scaling causes the wrong choice of pivotal elements to be made, leading to worthless results. Indeed, such an example has been constructed and has caused the failure of a computer program.

An iterative process for improving the scaling of a matrix will now be given. It is assumed that the matrix A is irreducible by permutation matrices. Let it be supposed that the k^{th} step of the process involves the i_kth row and column of A, where i_k-1 is the least positive residue of k-1 modulo n. Form

$$D_{k+1} = \bar{D}_k D_k, \quad D_1 = I \ , \tag{7}$$

where

$$D_k = \begin{cases} \text{diag } (1,1,\ldots,1,d_{i_k},1,\ldots,1), & i_k < n \\ \text{diag } (d_{i_k}^{-1}, d_{i_k}^{-1},\ldots,d_{i_k}^{-1}, 1) \ , & i_k = n \ , \end{cases} \tag{8}$$

d_{i_k} to be determined. Then,

$$A_{k+1} = \bar{D}_k A_k \bar{D}_k^{-1} \ , \quad A_1 = A \ . \tag{9}$$

Let $A_k = (a_{rs}^{(k)})$ and compute

$$\begin{aligned} R_{i_k}^{(k)} &= \left(\sum_{\substack{j=1 \\ j \neq i_k}}^{n} \left| a_{i_k j}^{(k)} \right|^2 \right)^{1/2} \\ S_{i_k} &= \left(\sum_{\substack{j=1 \\ j \neq i_k}}^{n} \left| a_{j i_k}^{(k)} \right|^2 \right)^{1/2} \ . \end{aligned} \tag{10}$$

The positive real number d_{i_k} is chosen so as to minimize

$$d_{i_k}^2 \, R_{i_k}^{(k)^2} + S_{i_k}^{(k)^2} / \, d_{i_k}^2 \ . \tag{11}$$

Thus,

$$d_{i_k} = \left(S_{i_k}^{(k)} / \, R_{i_k}^{(k)} \right)^{1/2} \ . \tag{12}$$

The above steps are repeated until $||A_k||_E$ changes by sufficiently small amounts for n consecutive steps.

The following theorem can be stated for the process above.

Theorem 3. The iterative process converges. That is,

(i) $\lim_{k \to \infty} A_k = A_f$ exists

(ii) $A_f = D_f A D_f^{-1}$

(iii) $D_f = \lim_{k \to \infty} D_k$

(iv) $||A_f||_E = \min_{D_\alpha} ||D_\alpha A D_\alpha^{-1}||_E$, D_α non-singular diagonal.

A Fortran program based on the above method was found to give rapid convergence.

In case a fixed point matrix is to be scaled, the process will yield the necessary diagonal matrix if one scales a floating point representation of the matrix.

A faster method can be obtained from the present one by forming a new matrix whose elements are the exponents in the floating point representations of the corresponding elements of the original matrix. The arithmetic steps would change accordingly and in place of (10) one could use

$$R_{i_k}^{(k)} = \max_{j \neq i_k} \left| b_{i_k j}^{(k)} \right| , \quad S_{i_k}^{(k)} = \max_{j \neq i_k} \left| b_{j i_k}^{(k)} \right| .$$

Reference

[1] Householder, A. S., "The Approximate Solution of Matrix Problems", J. Assoc. Comp. Mach. 5 (1958), 205 - 43.