Balancing a Matrix for Calculation of Eigenvalues and Eigenvectors

by B. N. Parlett and C. Reinsch

1. Theoretical Background

Introduction

This algorithm is based on the work of Osborne [1]. He pointed out that existing eigenvalue programs usually produce results with errors at least of order $\varepsilon \|A\|_E$, where $\varepsilon$ is the machine precision and $\|A\|_E$ is the Euclidean (Frobenius) norm of the given matrix $A$. Hence he recommends that one precede the calling of such a routine by certain diagonal similarity transformations of $A$ designed to reduce its norm. To this end Osborne discussed the case $p = 2$ of an algorithm which produces a sequence of matrices $A_k$, ($k = 1, 2, \ldots$) diagonally similar to $A$ such that for an irreducible $A$

(i) $A_f = \lim_{k \to \infty} A_k$ exists and is diagonally similar to $A$,

(ii) $\|A_f\|_p = \inf (\|D^{-1}AD\|_p)$, where $D$ ranges over the class of all non-singular diagonal matrices,

(iii) $A_f$ is balanced in $\|\cdot\|_p$, i.e. $\|a_i\|_p = \|a^T_i\|_p$, $i = 1, \ldots, n$ where $a_i$ and $a^T_i$ denote respectively the $i$-th column and $i$-th row of $A_f$,

(iv) $A$ and $D^{-1}AD$ produce the same $A_f$.

The process may be organized to compute either $A_f$ explicitly or only the corresponding diagonal transformation matrix. In the latter case, rounding errors can not cause a loss of accuracy in the eigensystem; however, existing eigenvalue programs must be modified so that they act on the matrix $D^{-1}AD$, implicitly given by $A$ and $D$. This realization is preferable from a systematic point of view. In practice, however, it appears simpler to scale the given matrix explicitly without changing existing programs. No rounding errors need occur in the execution if the process is implemented on a digital computer in floating point arithmetic, and if the elements of the diagonal matrices are restricted to be exact powers of the radix base employed, usually 2, 10, or 16.

---


** The following norms will be used. Let $1 \leq p < \infty$, $x = (x_1, \ldots, x_n)^T$ and $B = (b_{ij})$ of order $n$, then $\|x\|_p = (\sum |x_i|^p)^{1/p}$ and $\|B\|_p = (\sum \sum |b_{ij}|^p)^{1/p}$, where $\sum$ denotes summation over the index values 1, $\ldots$, $n$. $\|B\|_E = \|B\|_F$.

*** For a definition of irreducible see [1].
The Algorithm

Let $A_0$ denote the off-diagonal part of $A$. For any non-singular diagonal matrix $D$, 
\[
D^{-1} A D = \text{diag}(A) + D^{-1} A_0 D
\]
and only $A_0$ is affected. Assume that no row or column of $A_0$ identically vanishes.

From $A_0$ a sequence $\{A_k\}$ is formed. The term $A_k$ differs from $A_{k-1}$ only in one row and the corresponding column. Let $k = 1, 2, \ldots$ and let $i$ be the index of row and column modified in the step from $A_{k-1}$ to $A_k$. Then, if $n$ is the order of $A_0$, $i$ is given by

\[
\begin{align*}
    i - 1 & \equiv k - 1 \pmod{n}.
\end{align*}
\]

Thus, the rows are modified cyclically in their natural order. The $k$-th step is as follows.

(a) Let $R_k$ and $C_k$ denote the $\| \cdot \|_p$ norms of row $i$ and column $i$ of $A_{k-1}$. According to the above assumption $R_k C_k \neq 0$. Hence, if $\beta$ denotes the radix base, there is a unique signed integer $\sigma = \sigma_k$ such that

\[
\beta^{\sigma - 1} < R_k / C_k \leq \beta^{\sigma + 1}.
\]

Define $f = f_k$ by

\[
f = \beta^\sigma.
\]

(b) For given constant $\gamma \leq 1$ take

\[
D_k = \begin{cases} 
    I + (f - 1) e_i e_i^T & \text{if } (C_k f + R_k / f)^p < \gamma (C_k^p + R_k^p) \\
    I & \text{otherwise.}
\end{cases}
\]

Here $I = (e_1, \ldots, e_n)$ is the identity matrix.

(c) Form

\[
D_k = D_k D_{k-1}; \quad D_0 = I, \\
A_k = D_k^{-1} A_{k-1} D_k.
\]

We note that with this algorithm if $\gamma = 1$ then in every step, the factor $f$ is that integer power of the machine base $\beta$ which gives maximum reduction of the contribution of the $i$-th row and column to $\|A_k\|_p$. If $\gamma$ is slightly smaller than 1, a step is skipped if it would produce an insubstantial reduction of $\|A_{k-1}\|_p$. Iteration is terminated if for a complete cycle $D_k = I$.

For $\gamma = 1$ and $p = 2$ the algorithm differs from Osborne’s proposal only by the constraint on $f$ given in (a). Nevertheless it shows different properties. Corresponding to the above listed points we have (under the assumption that $A_0$ has no null rows or columns):

(i) The process may also be applied to a reducible $A$ and in all cases terminates after a finite number of iterations with a matrix $A_f$,

(ii) $\|A_f\|_p \geq \inf \|D_\beta^{-1} A D_\beta\|_p$, where $D_\beta$ ranges over the class of all diagonal matrices with entries restricted to integer powers of the machine base $\beta$, and inequality may well hold,