elements of $A^T A$, is always adequate. While as Taussky shows [2], the system $A^T AX = A^T b$ is more ill-conditioned than $AX = b$, this ill condition reflects itself in a larger computing time but does not affect accuracy.

The system has been applied with success at the Canadian Armament Research and Development Establishment on the ALWAC III machine. The circumstance was such that the routine for proper vectors was programmed prior to that for matrix inversion. It was originally meant as a stop-gap measure but has turned out to be a practical method of much merit.

N. S. MENDELSOHN

Canadian Armament Res. and Dev. Est.
and
The University of Manitoba
Winnipeg, Canada


The Computation of Complex Proper Values and Vectors of a
Real Matrix with Application to Polynomials

Introduction. The power method for computing the largest real proper value and corresponding proper vector of a real matrix $A$ can with very minor adjustments be applied to the complex proper values and vectors. A good exposition of the power method may be found in von Mises and Pollaczek-Geiringer [4]. When applied to the companion matrix of a real polynomial the method yields a real root or pair of complex roots of the polynomial. It turns out that in this application the method is equivalent to Bernoulli’s method, as was shown by Aitken [1]. Having obtained the root or roots of maximum modulus other roots are usually obtainable by the elementary devices indicated below. The method admits of easy programming on a digital computer. The program has been carried out on the ALWAC III computer at the Canadian Armament Research and Development Establishment and has turned out to be very practical indeed.

Method and Theory. Let $A$ be an $(r, r)$ real matrix with proper values $\lambda_1, \lambda_2, \ldots, \lambda_r$ and corresponding column proper vectors $x_1, x_2, \ldots, x_r$. It is assumed that $|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_r|$. If any one of the $\lambda_i$ is not real its conjugate appears among the remaining $\lambda_i$ and the corresponding proper vectors are conjugate. If $\eta$ is an arbitrary non-zero column vector with $r$ components, $\eta$ can be expressed in the form $\eta = c_1 x_1 + c_2 x_2 + \cdots + c_r x_r$. It follows that $\eta_n = A^n \eta = c_1 \lambda_1^n x_1 + c_2 \lambda_2^n x_2 + \cdots + c_r \lambda_r^n x_r$.

If $|\lambda_1| > |\lambda_2|$, then $\lambda_1$ is real and

$$\eta_n = \lambda_1^n \left[ c_1 x_1 + c_2 \left( \frac{\lambda_2}{\lambda_1} \right)^n x_2 + \cdots + c_r \left( \frac{\lambda_r}{\lambda_1} \right)^n x_r \right].$$

If $c_1 \neq 0$, the dominant term in $\eta_n$ is $\lambda_1^n c_1 x_1$. Hence, for sufficiently large $n$, $\eta_{n+1} \approx \lambda_1 \eta_n$. It is ordinarily convenient at each stage to normalize $\eta_n$ so that the
last component is 1. Accordingly putting \( \eta_n = \text{col} (y_1^{(n)}, y_2^{(n)}, \ldots, y_{r-1}^{(n)}, 1) \), the relationship between \( \eta_n \) and \( \eta_{n+1} \) is given by:

\[
\begin{align*}
\mu_{n+1} &= a_{r1}y_1^{(n)} + a_{r2}y_2^{(n)} + \cdots + a_{r,r-1}y_{r-1}^{(n)} + a_{rr} \\
y_1^{(n+1)} &= \frac{1}{\mu_{n+1}} \{ a_{11}y_1^{(n)} + a_{12}y_2^{(n)} + \cdots + a_{1,r-1}y_{r-1}^{(n)} + a_{1r} \} \\
y_2^{(n+1)} &= \frac{1}{\mu_{n+1}} \{ a_{21}y_1^{(n)} + a_{22}y_2^{(n)} + \cdots + a_{2,r-1}y_{r-1}^{(n)} + a_{2r} \} \\
&\quad \vdots \\
y_{r-1}^{(n+1)} &= \frac{1}{\mu_{n+1}} \{ a_{r-1,1}y_1^{(n)} + a_{r-1,2}y_2^{(n)} + \cdots + a_{r-1,r-1}y_{r-2}^{(n)} + a_{r-1,r} \}.
\end{align*}
\]

Here \( a_{ij} \) is the \((i, j)\) element of the matrix \( A \), \( \mu_{n+1} \) is a factor of proportionality which makes the last component of \( \eta_{n+1} = 1 \), and the iteration is started using a non-exceptional real vector \( \eta = \eta_0 \) with last component 1. The vector \( \eta_n \) converges to the proper vector \( x_1 \) and \( \mu_{n+1} \) converges to \( \lambda_1 \).

In the case where \(|\lambda_1| = |\lambda_2| > |\lambda_3|, \lambda_1 \neq \lambda_2\) the iteration (I) no longer converges. This case occurs when either \( \lambda_1 = -\lambda_2 \) or when \( \lambda_1 \) and \( \lambda_2 \) form a conjugate complex pair. We will assume the latter occurs. Put \( \lambda_2 = \tilde{\lambda}_1 \) and \( x_2 = \tilde{x}_1 \). Although the iteration (I) diverges it is still useful, as is shown by the following theorem.

**Theorem.** Let \( A \) be an \((r, r)\) matrix with real entries and proper values \( \lambda_1, \lambda_2, \lambda_3, \lambda_4, \ldots, \lambda_r \) and corresponding proper vectors \( x_1, \tilde{x}_1, x_2, \tilde{x}_2, \ldots, x_r \) such that \( \lambda_1 \) is not real and \(|\lambda_1| = |\tilde{\lambda}_1| > |\lambda_3| \geq |\lambda_4| \geq \cdots \geq |\lambda_r|\). Let \( \mu_{n+1} \) be given by the iteration (I), where (I) is started by means of a real, non-zero vector \( \eta_0 \) which has 1 as its last component. Then the sequence

\[
\nu_n = \frac{1}{2} \mu_{n+1} \left( \frac{\mu_n - \mu_{n+2}}{\mu_n - \mu_{n+1}} \right)
\]

converges to a limit \( \alpha \). Furthermore, the sequence

\[
\sigma_n = \mu_n \mu_{n+1} \left( \frac{\mu_{n+1} - \mu_{n+2}}{\mu_n - \mu_{n+1}} \right)
\]

converges to a limit \( \beta^2 \) and the proper values \( \lambda_1, \tilde{\lambda}_1 \) are given by \( \alpha \pm i\beta \) where \( \beta = \sqrt{\beta^2 - \alpha^2} \).

**Proof.** Let \( \eta_0 = c_1x_1 + c_2x_2 + \cdots + c_rx_r \). Then

\[
\eta_n = k_n \{ c_1\lambda_1^n x_1 + c_2\lambda_2^n x_2 + \cdots + c_r\lambda_r^n x_r \}
\]

where \( k_n \) is a constant which makes the last component of \( \eta_n = 1 \). For sufficiently large \( n \) all but the first two terms of the sum may be neglected. Let \( N \) be a specific value of \( n \) sufficiently large so that \( \eta_N = ax_1 + \bar{a}\tilde{x}_1 \) the neglected terms being negligible. Since the vector \( x_1 \) is only determined up to a constant factor we may
replace \(x_1\) by \(ax_1\). With this convention \(\eta_N = x_1 + \bar{x}_1\). Let \(\xi_r\) be the last component of \(x_1\). Since \(\eta_N\) has last component 1, \(\xi_r + \bar{\xi}_r = 1\). By the iteration (I) it follows that \(\eta_{N+1} = A\eta_N/\mu_{N+1} = (\lambda_1 x_1 + \bar{\lambda}_1 \bar{x}_1)/\mu_{N+1}\). Also, since \(\eta_{N+1}\) has last component 1, \(\mu_{N+1} = \lambda_1 \xi_r + \bar{\lambda}_1 \bar{\xi}_r = (\lambda_1 \xi_r + \bar{\lambda}_1 \bar{\xi}_r)/(\xi_r + \bar{\xi}_r)\). Put \(\lambda_1 = \alpha + i\beta = re^{i\theta}, \xi_r = e^{i\theta}\).

Then \(\mu_{N+1} = \rho \cos (\theta + \gamma)\). In the same way it follows that \(\mu_{N+2} = \rho \cos (2\theta + \gamma)\). On eliminating \(\gamma\) the relation \(\rho^2 - 2\mu_{N+1}\rho \cos \theta + \mu_{N+1}\mu_{N+2} = 0\) is obtained. Replacing \(N\) by \(N + 1\) yields \(\rho^2 - 2\mu_{N+2}\rho \cos \theta + \mu_{N+2}\mu_{N+3} = 0\). Solving these latter two equations, one obtains

\[
\alpha = \rho \cos \theta = \frac{1}{2} \frac{\mu_{N+2}(\mu_{N+1} - \mu_{N+3})}{\mu_{N+1} - \mu_{N+2}} = \nu_{N+1}; \quad \rho^2 = \frac{\mu_{N+1}\mu_{N+2}(\mu_{N+2} - \mu_{N+3})}{\mu_{N+1} - \mu_{N+2}} = \sigma_{N+1}.
\]

To solve the problem of determining \(\lambda_1\) on a digital computer, the numbers \(\mu_n\) are first generated from the iteration (I) and then the \(\nu_n\) and \(\sigma_n\) are computed from the \(\mu_n\). When successive values of \(\nu_n\) and \(\sigma_n\) differ by less than the allowable preassigned error the iterations are stopped and \(\alpha\) and \(\beta\) are given the values \(\alpha = \nu_{n+1}, \beta = \sqrt{\sigma_{n+1} - \nu_{n+1}^2}\). This algorithm for obtaining \(\alpha, \beta, \) and \(\rho\) is also discussed in Rutishauser [3] and Aitken [1].

A word of warning may be appropriate here. While theoretically the \(\nu_n\) and \(\sigma_n\) converge, it may happen in practice that rounding off errors in computed values of these will leave the magnitude of the difference of successive values greater than the preassigned error. Hence, if the computation continues for too long a time it may be necessary to print out a few consecutive values of \(\nu_n\) and \(\sigma_n\) and examine the differences.

In the case of complex proper values, there may be no occasion to obtain the corresponding proper vectors, but the computation presents no difficulty. Having computed \(\lambda_1\) the vector \(x_1\) may be obtained from the equation 

\(x_1 = k(\mu_{N+1}\eta_{N+1} - \bar{\lambda}_1 \eta_n)\) where \(k\) is a constant of proportionality.

Application to Polynomials. Let \(f(x) = x^r + c_1 x^{r-1} + \cdots + c_r = 0\) have roots \(\lambda_1, \lambda_2, \cdots, \lambda_r\) such that \(|\lambda_1| \geq |\lambda_2| \cdots \geq |\lambda_r|\). The companion matrix

\[
A = \begin{bmatrix}
0 & 1 & 0 & 0 & \cdots & 0 \\
0 & 0 & 1 & 0 & \cdots & 0 \\
& \ddots & & \ddots & \ddots & \ddots \\
& & 0 & 0 & 0 & \cdots & 0 \\
& & & -c_r & -c_{r-1} & \cdots & -c_2 - c_1
\end{bmatrix}
\]

has proper values \(\lambda_1, \lambda_2, \cdots, \lambda_r\). The computation method here described will yield either a single real root or a pair of complex roots. Further roots of \(f(x) = 0\) may be obtained by various devices. For instance the matrix \(A + kI\) has proper values \(\lambda_1 + k, \lambda_2 + k, \cdots, \lambda_r + k\). It may well happen that \(|\lambda_1 + k|\) is not the largest of the numbers \(|\lambda_1 + k|, |\lambda_2 + k|, \cdots, |\lambda_r + k|\). In this case, another root or pair of roots of \(f(x) = 0\) is obtained. If \(\lambda_1\) is real the choice \(k = -\lambda_1\) is
certain to yield further roots. If \( \lambda_1 = \bar{\lambda}_1 \) are a pair of conjugate complex roots, taking \( k = -\) real part of \( \lambda_1 \) will minimize \( |\lambda_1 + k| \) and will possibly yield further roots. By forming the polynomial \( g(x) = x^r f\left(\frac{1}{x}\right) \) one obtains the roots of minimum modulus of \( f(x) \). These devices have the advantage that the precision obtained in any new root is independent of the errors in previously obtained roots. It is only after these devices have become exhausted, that it may become necessary to remove from \( f(x) \) factors corresponding to computed roots.

**Concluding Remarks.** The method described here fails only in the case where \( f(x) \) has more than one pair of roots of the same maximum modulus (e.g., the equation \( x^r - 1 = 0 \)). It could be refined without difficulty to take in these more general cases, as indeed was done by Rutishauser [3] and Aitken [1]. The results are considerably more complicated in form and it is doubtful if they have any practical value.

N. S. MENDELSON

Canadian Armament Res. and Dev. Est.

and

The University of Manitoba

Winnipeg, Canada


1. A Method of Inverting Large Matrices of Special Form

**1. Introduction.** A method is suggested to invert large matrices, \( M_n \), of the form

\[
M_n = \begin{pmatrix}
\alpha_0 & -P_1 & 0 & 0 & \cdots & 0 \\
-P_1 & \alpha_1 & -P_2 & 0 & \cdots & 0 \\
0 & -P_2 & \alpha_2 & -P_3 & & \\
0 & 0 & -P_3 & \alpha_3 & \cdots & \\
0 & 0 & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & -P_n \\
\end{pmatrix}
\]

(1)

in which the elements are themselves special square matrices. The diagonal elements, \( \alpha_n \), are symmetric matrices. The minus signs are not required; they were a convenience in the particular problems we studied. This type of matrix arises, e.g., in the least-square fitting of survey data or in the difference equation approximation, to some common partial differential equations.

The method suggested takes advantage of the special properties of the large matrix and involves only algebraic operations and inversions of matrices the size of the elements. It is thus particularly applicable for use by “medium” sized