Iterative Numerical Methods for Real Eigenvalues and Eigenvectors of Matrices

John Coffey, Cheshire, UK.

August 2016

Key words: matrix, eigenvalue, eigenvector, iteration, power method, inverse power method, shifting, Rayleigh quotient, LU decomposition, matrix deflation, rank order reduction, QR, Schur decomposition, geometric series, Jacobi method, Hessenberg matrix, Householder reflector, Francis' implicitly shifted QR algorithms.

1 Introduction

This article gives a brief, informal account of some aspects of iterative numerical methods for finding the real eigenvalues and eigenvectors of square matrices with real elements. There is a large, well documented literature on this subject and many computer algorithms and sophisticated programs to implement them. The whole subject is interesting because of the innovative methods that have been devised to persuade matrices to reveal their eigen pairs (values and vectors). This article touches on only a few aspects which I have looked at for personal interest, my aim being to remind myself of the properties of matrices and thence to gain some understanding of the numerical techniques which have been developed.

I came to this subject through the modelling of vibrating structures using finite element methods. In these the structure is represented by elastic elements defined over a mesh of nodes, and the mass and stiffness are represented by symmetric matrices $\mathbf{M}$ and $\mathbf{K}$ respectively. In a previous article on *www.mathstudio.co.uk* entitled 'Periodic Forced Vibrations, Normal Modes and Damping, with Measurements on a 'Cello' I explain how the equations of motion can be written in the form

$$\mathbf{K}\,\mathbf{x} = -\omega^2\mathbf{M}\mathbf{x} \qquad \text{or} \qquad \mathbf{M}^{-1}\,\mathbf{K}\,\mathbf{x} \;=\; \omega^2\mathbf{x}\,. \tag{1}$$

This is a linear eigenvalue problem with the standard form $\mathbf{E}\mathbf{p} = \lambda\mathbf{p}$ where the natural frequencies of vibration are given by the square roots of the eigenvalues $\lambda = \omega^2$. The corresponding eigenvectors $\mathbf{p}$ give the relative amplitudes of motion at the mesh nodes. When these calculations are carried out by finite element programs, the eigenvalue/eigenvector pairs are determined to chosen precision by iterative numerical methods. This contrasts with the approach in analytical mathematics where the steps are as follows:

1. Let the $n \times n$ square matrix be $\mathbf{E}$ and let $\mathbf{I}$ be the unit matrix of the same dimension. Form $\mathbf{E} - \lambda\mathbf{I}$ where λ is a scalar to be determined.

2. Evaluate the determinant of $\mathbf{E} - \lambda\mathbf{I}$. This will be a polynomial in λ of degree n. Solve for the n zeros, which may be real and distinct, real with multiplicity or in complex conjugate pairs. These are the eigenvalues λ_1, λ_2, ... λ_n.

3. To find the corresponding eigenvectors, substitute λ_j into $\mathbf{E} - \lambda\mathbf{I} = \mathbf{0}$. The result will be a singular matrix which represents the coefficients of a set of simultaneous equations in the components of the eigenvector $\mathbf{p}_j$. Take one of these components to have a given value (usually 1) and solve for the other $n-1$ components. This can be done by inverting the $n-1$ by $n-1$ matrix obtained by deleting the row and column indexed by the chosen component.

This procedure is of wide applicability and would in principle give the exact eigenvalues and eigenvectors provided arbitrary precision arithmetic were used. It applies equally to real matrices with real eigenvalues, ones with some complex conjugate eigenvalue pairs, and to matrices with complex elements. However, in practice it is only applicable to relatively small matrices – say up to 10×10. The effort in evaluating the determinant, solving numerically for all n roots, and then solving the set of $n-1$ simultaneous equations is prohibitive for large matrices and rounding errors become troublesome. Bear in mind that some matrices in finite element calculations may have thousands of elements, making computer storage an issue even today. To avoid the obstacles to computation in the direct approach, several iterative numerical schemes have been developed over many years. There are three main computational challenges:

- to find all the eigenvalues and eigenvectors, or at least all of interest. In vibration problems often the eigenvalues with smallest magnitude are most important because they correspond to the lowest frequencies,
- for any identified eigenvalue, to converge rapidly and accurately,
- for any identified eigenvector, to converge rapidly and accurately. Some methods converge on both an eigenvector its eigenvalue simultaneously.

I deal entirely with real square matrices, and most examples will have only real eigenvalues. I describe some of the basic methods, commenting on convergence rates and applicability, but there is no deep analysis of stability or the computational effort required. Though these matter are central to numerical analysis, the reader must look to the literature for such details. Four books which give thorough accounts of this large subject are

- 'The Algebraic Eigenvalue Problem' by J. H. Wilkinson, Oxford Univ, Press, 1965
- 'Matrix Computations' by Gene H. Golub and Charles F. Van Loan, Third Edn. 1996, John Hopkins Univ Press. Available electronically on the internet.
- 'The Matrix Eigenvalue Problem: GR and Krylov Subspace Methods' by David S. Watkins, pub. SIAM, 2007
- 'Fundamentals of Matrix Computations' by David S. Watkins, Third Edn. 2010. Pub. John Wiley.

There are also many original papers and lecture notes on the internet, including the review of the QR algorithm 50 years on by Gene Golub and Frank Uhlig[1].

In this article the scene is set in §2 by listing some useful properties of matrices and their eigenvalues and eigenvectors and giving numerical illustrations in §3. The similarity transformation is perhaps the most important concept because it changes the matrix to a more tractable form without changing its eigenvalues. The next two sections are all related to the Power Method by

[1] IMA Journal of Numerical Analysis Vol 29, 467-485, 2009.

which the dominant eigenvalue (the one with largest absolute value) and its eigenvector are found simultaneously. §4 describes the basic, direct power method with an example. Shifting the diagonal of the matrix by subtracting a constant can lead to faster convergence since the rate depends on the ratio of eigenvalues. §5 explains how fast convergence can be obtained with the inverse power method, and how the direct power and inverse power methods can be use together to find one eigen pair at a time. The inverse method in principle involves the inverse of the given matrix, but the problems in actually finding the inverse are in practice circumvented by the equivalent process of solving a system of simultaneous equations, and this in turn is made straightforward by factorising the given matrix into a product LU of a lower (L) and an upper (U) triangular matrix.

After the dominant and possibly one or two other eigen pairs have been found, 'matrix deflation' may be used to allow more to be found. Deflation means finding a smaller matrix which has the same eigenvalues and vectors as the given matrix expect for one or two known eigen pairs which have been removed. It is a way of chopping off the eigen pairs which have already been determined from the matrix, allowing further calculation of the dominant eigenpairs in the reduced matrix which has smaller eigenvalues. Some deflation and matrix order reduction algorithms are described in §6.

§7 deals with an algorithm first described by Carl Jacobi in 1841. Though this is effective only for symmetric matrices, it has the attractive property of converging on all eigenvalues simultaneously, in contrast with the power method. It is probably the earliest method with this property. Jacobi's method works through a sequence of nested similarity transformations at each of which one pair of the elements in the lower left and upper right of the matrix is mapped to 0 by, in effect, a rotation of axes. In this way the matrix is gradually transformed to a diagonal matrix where the eigenvalues can be read down the diagonal.

§8 returns to the idea of factorising a matrix into a product of two with special properties. Historically, the LU decomposition described in §5 was extended by Heinz Rutishauser in the late 1950s into an iterative algorithm called LR. This had serious stability problems for some types of matrix, but the idea was sound and was part of the inspiration of John Francis in Britain and Vera Kublanovskaya in Russia to develop the more stable QR method in about 1960. The basic or 'explicit' QR method is described with examples in §8. At each stage of iteration the starting matrix **A** is factored into the product **QR** where **Q** is an orthogonal matrix, found by the Gram-Schmidt procedure, and **R** is upper triangular, the so-called Schur equivalent form of **A**. These factors are then multiplied in reverse order **RQ** which happens to have smaller elements below the main diagonal. After several iterations **RQ** is itself almost triangular, at which stage the all the eigenvalues can be read down the diagonal.

John Francis published two seminal papers in 1960 and 1961 respectively. In the second he developed modifications of the QR method which were so profound that they constitute a distinct and powerful algorithm which has since been known as the 'implicitly shifted QR algorithm'. In recent years David Watkins of Washington State University, who has written extensively on the subject, has urged that it be renamed 'Francis's algorithm', pointing out that it is better understood in its own right rather than as a version of the basic QR algorithm. Francis' algorithm requires first that the given matrix be put into so-called Hessenberg form in which all elements below the sub-diagonal (lower off-diagonal) are zero. Methods for transforming to Hessenberg form using matrices equivalent to reflections are described in §9. Francis' single and double shift algorithms are outlined in §10.

Paradoxically, though my interest was stimulated by eigenvalue solutions of finite element models, I do not say much in this article about the numerical methods most suited to the large, sparse matrices which typically arise with finite elements. Very large matrices are best not stored in a computer as $n \times n$ arrays in which almost all elements are zero, but rather as a list of the non-zero elements together with their two positional indices. Several of the more powerful numerical methods for general, dense matrices, such as the QR family of algorithms, involve matrix-by-matrix multiplication and in this a matrix which starts as sparse generally becomes densely filled-in within one or two iterations. If the computer storage cannot cope with these large, dense product matrices, the method cannot be applied. For this reason the powerful 'implicitly single and double shifted QR' methods due to John Francis are not suitable for very large matrices. Instead, an algorithm must be used which avoids matrix-by-matrix multiplication, and has nothing more dense than matrix-by-vector multiplication. The most popular is called the 'implicitly restarted Arnoldi' algorithm. I mention it fleetingly, along with Krylov subspaces, in §8.4.

Appendix 1, §11, is an analysis of the sum of two or more geometric series, a feature of the Power Method. I show how the coefficients of each series can be determined by iterative solution of a set of simultaneous non-linear equations, and the results used to give an accurate estimation of the eigenvalue and eigenvector. Appendix 2, §12, is an example of the range of Power Methods being applied to solve a straightforward 5×5 real matrix. Appendix 3, §13 tackles a problem matrix which has some eigenvalues close together. Such matrices can defeat the Power Method but the QR-Schur decomposition solves it, and Francis' single shift algorithm makes light work of it.

2 Some properties of eigenvalues and eigenvectors

This section lists some facts in no particular order and §3 gives numerical examples. The $n \times n$ square matrix $\mathbf{E}$ is assumed to have real elements, but is otherwise arbitrary unless stated. Some of these properties will also apply to complex elements, but they are not of interest to my type of vibration modelling, so I ignore them. The eigenvalues are λ_j and eigenvectors are $\mathbf{p}_j,\ j = 1, n$. $\mathbf{v}$ is a general n-vector.

1. Eigenvalues and eigenvectors arise as solutions of the equation $\mathbf{Ep} = \lambda\mathbf{p}$ for all square matrices, invertible and non-invertible. Symmetric matrices have $\mathbf{E} = \mathbf{E}^T$, where ^{T} denotes transpose. A symmetric matrix with real coefficients always has real eigenvalues and its eigenvectors are mutually orthogonal. A 'positive definite' matrix is a symmetric matrix such that for every vector $\mathbf{v}$, $\mathbf{v}^T\mathbf{Ev} > 0$; all its eigenvalues are real and positive. The form $\mathbf{v}^T\mathbf{Ev}$ appears in the physics of vibration as $\frac{1}{2}\mathbf{v}^T\mathbf{Mv}$ for the kinetic energy and $\frac{1}{2}\mathbf{v}^T\mathbf{Kv}$ for potential strain energy. $\mathbf{M}$ and $\mathbf{K}$ must be positive definite as energy cannot be negative.

2. The eigenvectors $\mathbf{p}_j$ are only determined up to the ratio of their components. They are usually normalised by multiplying by a scale factor chosen to give one component the value 1, or to make the modulus 1 so that each $\mathbf{p}_j$ is a unit vector.

3. The trace of a square matrix (the algebraic sum of its diagonal elements) equals the algebraic sum of the eigenvalues. Hence in an $n \times n$ matrix if $n-1$ eigenvalues have been found, the last is given by subtracting from the trace.

4. Given an eigenvector $\mathbf{p}$, the corresponding eigenvalue is readily found in one of two ways. i) Normalise the vector so that one component is 1 then multiply by $\mathbf{E}$; that component will be replaced by the eigenvalue. ii) Since $\mathbf{Ep} = \lambda\mathbf{p}$, $\mathbf{p}^T\mathbf{Ep} = \lambda\mathbf{p}^T\mathbf{p}$. The dot product $\mathbf{p}^T\mathbf{p} = |\mathbf{p}|^2$ so $\lambda = \mathbf{p}^T\mathbf{Ep}/|\mathbf{p}|^2$. This quantity is called the Rayleigh quotient after Lord Rayleigh who

introduced it in volume 1 of his book 'The Theory of Sound', his §90, page 113 *et seq.*. The Rayleigh quotient of a matrix achieves its least value when λ is the smallest absolute eigenvalue, corresponding for a vibrating system with the lowest frequency.

5. The eigenvectors are linearly independent, though not orthogonal unless the matrix is symmetric. Two or more eigenvectors can share the same eigenvalue; the eigenvalue is then said to be degenerate. For a non-degenerate matrix the eigenvectors form a complete basis set of dimension n, meaning that they span the space of n dimensions. Then any other vector $\mathbf{v}$ in n dimensions can be written as a linear combination

$$\mathbf{v} \;=\; c_1\mathbf{p}_1 + c_2\mathbf{p}_2 + ...c_n\mathbf{p}_n. \tag{2}$$

6. An important consequence of item 5 is that powers of $\mathbf{E}$ applied to an arbitrary vector $\mathbf{v}$ converge to the eigenvector with the largest absolute eigenvalue. To see this, suppose that with some relabelling $\lambda_1 > \lambda_2 > ... > \lambda_n$. Apply $\mathbf{E}$ repeatedly to Eq 2.

$$\begin{aligned}
\mathbf{E}^k\mathbf{v} &= c_1\lambda_1^k\mathbf{p}_1 + c_2\lambda_2^k\mathbf{p}_2 + ...c_n\lambda_1^k\mathbf{p}_n, \\
&= c_1\lambda_1^k\left(\mathbf{p}_1 + \frac{c_2\lambda_2^k}{c_1\lambda_1^k}\mathbf{p}_2 + ...\frac{c_n\lambda_n^k}{c_1\lambda_1^k}\right). \qquad (3) \\
&\rightarrow \quad c_1\lambda_1^k\mathbf{p}_1 \quad \text{as } k\rightarrow\infty\,.
\end{aligned}$$

   This is the basis of the Power Method for finding the eigenvalue of largest absolute value and simultaneously the corresponding eigenvector. The method is described in §3. Clearly convergence depends on both the choice of initial vector (through c_2/c_1) and on the ratio of absolute eigenvalues $|\lambda_2|/|\lambda_1|$.

7. If elementary row operations of replacing a row by itself plus a multiple of another row are used to convert a matrix to triangular form, the signs of the diagonal elements (called the 'pivots') give the signs of the eigenvalues, though not their values. The product of all the pivots happens to be the determinant of the original matrix. However, the elementary row and column operations in general change the trace, the characteristic equation and hence the eigenvalues and eigenvectors.

8. If a constant β is subtracted from all the diagonal elements of $\mathbf{E}$, the eigenvalues of $\mathbf{E}-\beta\mathbf{I}$ are $\lambda_j - \beta$, $1 \le j \le n$. This is called 'shifting'. The proof is quite simple; $(\mathbf{E}-\beta\mathbf{I})\mathbf{p} = \lambda\mathbf{p} - \beta\mathbf{I}\mathbf{p} = (\lambda-\beta)\mathbf{p}$. Shifting by a carefully chosen constant can be used to enhance the convergence of the Power Method by changing the ratio λ_2/λ_1.

9. The eigenvalues of the transpose $\mathbf{E}^T$ are the same as those of $\mathbf{E}$, but the eigenvectors are different unless the matrix is symmetric. To prove this suppose that

$$\mathbf{E}\mathbf{p} \;=\; \lambda\mathbf{p} \qquad \text{and} \qquad \mathbf{E}^T\mathbf{q} \;=\; \mu\mathbf{q}\,.$$

   Using the order-reversing property of transposes,

$$\mathbf{q}^T\mathbf{E} \;=\; \mu\mathbf{q}^T \quad \text{so} \quad \mathbf{q}^T\mathbf{E}\mathbf{p} \;=\; \mu\mathbf{q}^T\mathbf{p}\,.$$

   But $\mathbf{q}^T\mathbf{E}\mathbf{p}$ is also obtained by multiplying $\mathbf{E}\mathbf{p} = \lambda\mathbf{p}$ on the left by $\mathbf{q}^T$, and it gives $\mathbf{q}^T\lambda\mathbf{p}$. This equals $\mu\mathbf{q}^T\mathbf{p}$ only if $\mu = \lambda$.
   The relation $\mathbf{q}^T\mathbf{E} = \lambda\mathbf{q}^T$ explains why $\mathbf{q}^T$ is referred to as the 'left eigenvector' of $\mathbf{E}$. $\mathbf{p}$ would then be called the right eigenvector.

10. The eigenvectors of the inverse of a matrix $\mathbf{E}^{-1}$ are the same as the eigenvectors of $\mathbf{E}$ and its eigenvalues are the reciprocals $1/\lambda_j$. The proof is: $\mathbf{E}^{-1}\,\mathbf{E}\,\mathbf{p}_j = \mathbf{E}^{-1}\lambda_j\,\mathbf{p}_j$ so $\mathbf{p}_j = \lambda_j\mathbf{E}^{-1}\,\mathbf{p}_j$ or $\mathbf{E}^{-1}\,\mathbf{p}_j = 1/\lambda_j\,\mathbf{p}_j$. There is a close relation between the eigenvectors of the inverse $\mathbf{E}^{-1}$ and the eigenvectors of the transpose $\mathbf{E}^T$, described and used in §?? on 'matrix deflation' which is the name given to eliminating a selected eigenvalue-vector pair from a matrix.

11. Suppose we have three $n \times n$ matrices, $\mathbf{E}$, $\mathbf{F}$ and $\mathbf{G}$, where $\mathbf{G}$ is invertible. If they are related by $\mathbf{E} = \mathbf{G}^{-1}\mathbf{F}\,\mathbf{G}$, then $\mathbf{E}$ and $\mathbf{F}$ are said to be 'similar' and linked by the similarity transformation (also called conjugate transformation) of pre-and post-multiplication by $\mathbf{G}$. $\mathbf{E}$ and $\mathbf{F}$ have the same determinant, characteristic equation, trace, and the same eigenvalues, though different eigenvectors. To see this, observe that if $\mathbf{Ep} = \lambda\mathbf{p}$, then $\mathbf{FGp} = \lambda\,\mathbf{Gp}$. So $\mathbf{F}$ also has eigenvalue λ, but eigenvector $\mathbf{Gp}$. Similar matrices represent the same linear transformation of a space from different sets of basis vectors.

12. A special case of the previous item is that a square matrix can be diagonalised – that is, represented by a diagonal matrix with the same eigenvalues and eigenvectors. The operation is effected by the matrix $\mathbf{P}$ whose columns are the column eigenvectors $\mathbf{p}_j$, $i \le j \le n$. Thus

$$\mathbf{D} \;=\; \mathbf{P}^{-1}\mathbf{EP} \tag{4}$$

    is diagonal, and its elements are its eigenvalues which are also the eigenvalues of $\mathbf{E}$. The eigenvectors are the orthogonal set $(1,0,0,...0)$, $(0,1,0,...0)$, ..., $(0,0,0,....1)$. We have here a method for producing an infinite family of matrices with the same prescribed eigenvalues: start with a diagonal matrix built of the given eigenvalues and transform it with any chosen invertible matrix in a similarity transformation.

13. A triangular matrix (one with 0s below or above the main diagonal) also has its eigenvalues arrayed down the diagonal. Only the diagonal elements contribute to the characteristic polynomial. If the eigenvalues are all real, the characteristic polynomial is a product of linear factors $(\lambda_1 - d_{11})(\lambda_2 - d_{22})\ldots(\lambda_N - d_{NN})$ where d_{jj} are the diagonal elements.

14. Several mathematicians have proved bounds on eigenvalues in terms of the value of the matrix elements. A simple test by Alfred Brauer, 1946, is based on the absolute values of the elements. Let a_{ij} be the elements, $R_i = \sum_j |a_{ij}|$ be the sum of absolute values along the i^{th} row and $C_j = \sum_i |a_{ij}|$ the sum down the j^{th} column. Let $\mathcal{R}$ be the largest R_i and $\mathcal{C}$ the largest C_j. Then for all eigenvalues $|\lambda| \le min(\mathcal{R}, \mathcal{C})$.

15. Other bounds were derived by Gershgorin (1931) by analogy with the diagonalised matrix $\mathbf{D}$ in item 11. He judged that if the off-diagonal elements are relatively small compared with the diagonal ones, the eigenvalues cannot be too far away from the values down the diagonal. This is quantified in Gershgorin's two Circle Theorems. Let the $n \times n$ matrix have elements a_{ij}, and along each row i add up the absolute values $|a_{ij}|$, $i \neq j$ of the off-diagonal elements. Call this sum r_i. Next, in the complex plane mark a point on the real axis at the point corresponding to the diagonal element a_{ii}. Using this as centre, draw a circle with radius r_i. Repeat this for all rows to give n such circles. These will overlap if some diagonal elements are close to each other. Theorem I states that every eigenvalue lies within at least one of these circles. Theorem II states that if n_1 circles overlap each other and another n_2 overlap each other but are disjoint from the first set, then exactly n_1 eigenvalues lie in the first set and n_2 in the second. The circles allow for the possibility of the eigenvalues being complex, though I shall not be concerned with such matrices. The theorem also holds if the sums of absolute values are taken down the columns instead of across the rows.

16. A third set of bounds on eigenvalues has been given by Wolkowicz and Styan ('Bounds on eigenvalues using traces' in Linear Algebra and its Applications, Vol 29, p 471-506, 1980). Suppose the n real eigenvalues of an $n \times n$ matrix are ordered $\lambda_1 \geq \lambda_2 \geq ....\lambda_k \geq .... \geq \lambda_n$. By analogy with statistics, Wolkowicz and Styan define a mean m and standard deviation s for the trace of a matrix by

$$m \;=\; \frac{TrE}{n} \;=\; \frac{1}{n}\sum_j^n \lambda_j\,, \qquad\qquad s^2 = \frac{1}{n}^2\left[nTr(E^2) - (TrE)^2\right] \;=\; \frac{1}{n}Tr(E^2) - m^2\,.$$

$$\text{Then} \quad m - s\sqrt{n-1} \;\leq\; \lambda_{min} \;\leq\; m - \frac{s}{\sqrt{n-1}}\,,$$

$$m + \frac{s}{\sqrt{n-1}} \;\leq\; \lambda_{max} \;\leq\; m + s\sqrt{n-1}\,,$$

$$m - s\sqrt{\frac{k-1}{n-k+1}} \;\leq \lambda_k \;\leq\; m + s\sqrt{\frac{n-k}{k}}\,.$$

As a special case, when $n = 3$

$$m - s\sqrt{2} \leq \lambda_3 \;\leq\; m - \frac{s}{\sqrt{2}} \;\leq \lambda_2 \;\leq\; m + \frac{s}{\sqrt{2}} \;\leq \lambda_1 \;\leq\; m + s\sqrt{2}\,. \tag{5}$$

17. There is a special class of matrix which arises is vibration analysis; a non-symmetric matrix made by the product of two symmetric matrices. The eigenvectors of these have orthogonality properties which are crucial for vibration modal analysis[2]. This is seen in Eq 1 where $\mathbf{M}$, $\mathbf{M}^{-1}$ and $\mathbf{K}$ are symmetric. If the eigenvectors of $\mathbf{E} = \mathbf{M}^{-1}\mathbf{K}$ are $\mathbf{p}_j$, the orthogonality condition is

$$\mathbf{p}_j^T\mathbf{M}\mathbf{p}_i \;=\; 0 \quad \text{if } i \neq j\,. \tag{6}$$

If $i = j$, the product is non-zero and used to scale and normalise the components of the eigenvectors to make $\mathbf{p}_j^T\mathbf{M}\mathbf{p}_j = \mathbf{I}$, an operation called 'mass normalisation'.

18. A matrix problem is said to be 'ill-conditioned' if numerical attempts to determine its solutions, such as finding the eigenvalues and eigenvectors, are over-sensitive to the precision of or errors in the input numbers. The 'condition number' ≥ 1 is defined as the ratio of error in output to error in the input and is given numerically as the product of the direct and inverse norms, $\|\,\mathbf{E}^{-1}\,\|\;.\;\|\,\mathbf{E}\,\|$. Here $\|\;...\;\|$ denotes the Euclidean norm (also called the spectral or L_2 norm) which is the square root of the sum of the squares of all the matrix elements. A large condition number indicates troublesome sensitivity to input. Ill-conditioned matrices can arise in finite-element models where a few matrix elements are orders of magnitude different from the rest. In practice finding $\mathbf{E}^{-1}$ is a challenge so the condition number may have to be estimated from intermediate numbers produced during a calculation.

19. The Cayley-Hamilton theorem: Each matrix satisfies its own characteristic equation. So if the matrix is $\mathbf{E}$ and its characteristic equation is $a_n\lambda^n + a_{n-1}\lambda^{n-1} + ... + a_1\lambda + a_0 = 0$, then $a_n\mathbf{E}^n + a_{n-1}\mathbf{E}^{n-1} + ... + a_1\mathbf{E} + a_0\mathbf{I} = \mathbf{0}$, the zero matrix. This provides a method for calculating higher powers of $\mathbf{E}$, since $\mathbf{E}^n = -(a_{n-1}\mathbf{E}^{n-1} + ... + a_1\mathbf{E} + a_0\mathbf{I})/a_n$. Moreover, the inverse can be found as a polynomial in $\mathbf{E}$ since $a_n\mathbf{E}^{n-1} + a_{n-1}\mathbf{E}^{n-2} + ... + a_1\mathbf{I} = -a_0\mathbf{E}^{-1}$.

[2] Proof of orthogonality is given in §2 of my article on modal vibration analysis on *www.mathstudio.co.uk*, entitled 'Periodic Forced Vibrations, Normal Modes and Damping, with Measurements on a 'Cello'.

20. The eigenvalues and eigenvectors of a positive definite matrix have a geometrical representation in terms of a 'representation ellipsoid'. This is most readily envisaged with a diagonalised 3-dimensional matrix, **D**. (Recall that all eigenvalues are real and positive and lie on the diagonal.) Consider the product $\mathbf{v}^T\mathbf{D}\mathbf{v}$ with $\mathbf{v} = (x, y, z)$. This is the dot or inner product of the vectors **v** and **Dv** and evaluates to $\lambda_1 x^2 + \lambda_2 y^2 + \lambda_3 z^2$. The locus of points which satisfy $\lambda_1 x^2 + \lambda_2 y^2 + \lambda_3 z^2 = 1$ is an ellipsoid with semi-axes of lengths $1/\sqrt{\lambda_1}$, $1/\sqrt{\lambda_2}$, $1/\sqrt{\lambda_3}$ and with principal axes along the x, y and z axes. If **E** is similar to **D**, it will also represent a quartic surface, but skewed and distorted. However three position vectors will be unchanged in direction by the transformation – these are the eigenvectors. The ellipsoid will be stretched in these directions by amounts given by the respective eigenvalues. Note that each of the lengths $1/\sqrt{\lambda_j}$ is a local maximum, minimum or stationary value with respect to small changes in the vector direction. This has prompted an important concept that the eigenvalues and vectors in a more general sense maximise or minimise some function of the matrix. The area of mathematical which investigates such problems is the 'calculus of variations'.

3 Illustrative examples of matrix properties

Some numerical examples will help the reader appreciate the above interesting facts. I will take the eigenvalue with largest absolute value to be λ_1, the next to be λ_2.

3.1 A general 3×3 matrix

Our first example matrix is

$$\mathbf{E} = \begin{pmatrix} -4 & -2 & 3 \\ 1 & 3 & 4 \\ -1 & 1 & 5 \end{pmatrix}. \tag{7}$$

Its trace is 4. To find bounds on the eigenvalues first apply Brauer's criterion, §2.1, item 12. The largest absolute row sum is 9 and largest column sum is 12, so $|\lambda_j| \le 9$. The matrix is not diagonally dominant, but let us see what Gershgorin theorems state as bounds of the eigenvalues. Applied to the rows, the circle centres are at −4, 3 and 5 and the respective radii are 5, 5 and 2. These overlap and span from −9 to +8. Applied to the columns the circles span from −6 to 12 so together the eigenvalues must lie within (−6, 8), a modest narrowing of bounds from ±9. The circle for column 1 spans $(-6, 2)$ and the other two overlap and together span (−2, 12). They only touch at −2 so it is likely that there is only one eigenvalue in (−6, −2) and two in $(-2, 8)$.

Solving for the eigenvalues in the classical analytic way, the characteristic equation is

$$(-4-\lambda)([3-\lambda][5-\lambda]-4) \;+\; 2(5-\lambda+4) \;+\; 3(1+3-\lambda) \;=\; -\lambda^3+4\lambda^2+16\lambda-14.$$

To check the Cayley-Hamilton theorem use

$$\mathbf{E}^2 = \begin{pmatrix} 11 & 5 & -5 \\ -5 & 11 & 35 \\ 0 & 10 & 26 \end{pmatrix}, \qquad \mathbf{E}^3 = \begin{pmatrix} -34 & -12 & 28 \\ -4 & 78 & 204 \\ -16 & 56 & 170 \end{pmatrix}$$

and the zero matrix does result. Since we have here $\mathbf{E}^2$, we can evaluate the Wolkowicz-Styan bounds using Eq 5 of item 14. $m = 4/3$ and $s^2 = 48/3 - 16/9$ so $s = 8\sqrt{2}/3 = 3{\cdot}77$. The eigenvalues are bounded

$$-4 \;<\lambda\; < \; -1\tfrac{1}{3} \; < \; \lambda \; < \; 4 \; < \; \lambda \; < \; 6\tfrac{2}{3}\,.$$

The roots of the characteristic equation are found by Newton's method to be

$$\lambda_1 \;=\; 6\cdot 212664048, \qquad \lambda_2 \;=\; -2\cdot 971119456\,, \qquad \lambda_3 \;=\; 0\cdot 7584554087\,.$$

I give them to high precision for later comparison with the results of approximate methods. It is heartening that all three bound estimates are consistent with these values. The Wolkowicz-Styan ones are the tightest.

To find eigenvector $\mathbf{p}_1$ we form the matrix equation $\mathbf{E} - 6\cdot 21266\,\mathbf{I} = \mathbf{0}$ and take p_{13} to be 1. There are then two independent simultaneous equations for the other two components:

$$-10\cdot 12166\,p_{11} \;-\; 2p_{12} \;-\; 3 \;=\; 0$$

$$p_{11} \;-\; 3\cdot 21266\,p_12 \;+\; 4 \;=\; 0$$

with solution $p_{11} = 0\cdot 0470554$, $p_{12} = 1\cdot 2597195$. The other eigenvectors are found in the same way, giving

$$\mathbf{p}_1 \;=\; \begin{pmatrix} 0\cdot 0470554 \\ 1\cdot 2597195 \\ 1 \end{pmatrix}, \qquad \mathbf{p}_2 \;=\; \begin{pmatrix} 6\cdot 2538745 \\ -1\cdot 7172449 \\ 1 \end{pmatrix}, \qquad \mathbf{p}_3 \;=\; \begin{pmatrix} 1\cdot 6990701 \\ -2\cdot 5424745 \\ 1 \end{pmatrix}. \tag{8}$$

As a demonstration that they span 3-space but are not orthogonal, the angles between pairs of eigenvectors are as follows: $\mathbf{p}_1$ and $\mathbf{p}_2$: $94\cdot 7^\circ$, $\mathbf{p}_2$ and $\mathbf{p}_3$: $40\cdot 8^\circ$, $\mathbf{p}_3$ and $\mathbf{p}_1$: $114\cdot 2^\circ$. Here is an example of a fairly arbitrary vector expressed as a sum of these eigenvectors:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \;=\; 0\cdot 89458\,\mathbf{p}_1 \;+\; 0\cdot 17098\,\mathbf{p}_2 \;-\; 0\cdot 06556\,\mathbf{p}_3\,. \tag{9}$$

Using elementary row and/or column operations the inverse of $\mathbf{E}$ is found to be

$$\mathbf{E}^{-1} \;=\; \frac{1}{14}\begin{pmatrix} -11 & -13 & 17 \\ 9 & 17 & -19 \\ -4 & -6 & 10 \end{pmatrix}. \tag{10}$$

This can also be found from the characteristic equation using the Cayley-Hamilton theorem in the form $14\mathbf{E}^{-1} = -\mathbf{E}^2 + 4\mathbf{E} + 16\mathbf{I}$. Its trace is $16/14$. Brauer's criterion gives that $|\lambda_j| < 45/14 = 3\cdot 21$. In fact $1/|\lambda_3| = 1\cdot 318$, comfortably under this upper bound. The three Gershgorin circles derived from the matrix rows overlap and cover the wide interval $(-41/14,\ 45/14)$, that is $(-2\cdot 93,\ 3\cdot 21)$. Applied to the columns, the circles span $(-26/14,\ 46/14)$ so combining all Gershgorin circles with Brauer's estimate reduces the interval to $(-26/14,\ 45/14) = (-1\cdot 86,\ 2\cdot 71)$.

The characteristic equation of $\mathbf{E}^{-1}$ is $-14h^3 + 16h^2 + 4h - 1 = 0$ with roots $h = 1\cdot 318469$, $0\cdot 160962$, $-0\cdot 336573$, equal to $1/\lambda_3$, $1/\lambda_1$, $1/\lambda_2$ respectively. To find the eigenvectors $\mathbf{q}_3$ of the largest eigenvalue, the simultaneous equations to solve are

$$-29\cdot 45857\,q_{31} \;-\; 13\,q_{32} \;+\; 17 \;=\; 0$$

$$9\,q_{31} \;-\; 1\cdot 45857\,q_{32} \;-\; 19 \;=\; 0$$

and the solution is $q_{31} = 1\cdot 69907$, $q_{32} = -2\cdot 54247$, precisely the same as for p_{31}, p_{32}. This illustrates the intriguing fact that a matrix and its inverse share the same eigenvectors.

The matrix $\mathbf{P}$ and its inverse which will diagonalise $\mathbf{E}$ are

$$\mathbf{P} = \begin{pmatrix} 0\cdot047055 & 1\cdot699070 & 6\cdot253875 \\ 1\cdot259719 & -2\cdot542475 & -1\cdot717245 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \mathbf{P}^{-1} = \begin{pmatrix} -0\cdot0441735 & 0\cdot2438131 & 0\cdot6949425 \\ 0\cdot2035268 & 0\cdot0884303 & -0\cdot1209744 \\ -0\cdot1593532 & -0\cdot3322434 & 0\cdot4260319 \end{pmatrix}.$$

Within machine accuracy

$$\mathbf{P}^{-1}\mathbf{E}\,\mathbf{P} = \begin{pmatrix} 6\cdot2126640 & 0 & 0 \\ 0 & -2\cdot9711195 & \\ 0 & 0 & 0\cdot7584554 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}.$$

The transposed matrix $\mathbf{E}^T$ has the same eigenvalues but these eigenvectors, which are proportional to the respective rows of $\mathbf{P}^{-1}$:

$$\mathbf{q}_1 = \begin{pmatrix} -0\cdot063564 \\ 0\cdot350839 \\ 1 \end{pmatrix}, \qquad \mathbf{q}_2 = \begin{pmatrix} -1\cdot682395 \\ -0\cdot730984 \\ 1 \end{pmatrix}, \qquad \mathbf{q}_3 = \begin{pmatrix} -0\cdot374042 \\ -0\cdot779856 \\ 1 \end{pmatrix}.$$

Direct calculation shows that the dot (inner) product $\mathbf{q}_i^T\mathbf{p}_j = 0$ if $i \neq j$. Therefore $\mathbf{q}_i$ and $\mathbf{p}_j$ are orthogonal. A useful normalisation of the non-zero products sets the $|\mathbf{p}_j| = 1$ and $\mathbf{q}_j^T\mathbf{p}_j = 1$. I make use of this scheme in §5 on matrix deflation, but there introduce notation which has the eigenvectors of $\mathbf{E}$ denoted $\mathbf{x}$ and normalised $|\mathbf{x}_j| = 1$, and the eigenvectors of $\mathbf{E}^T$ denoted $\mathbf{y}$ with $\mathbf{y}_j^T\mathbf{x}_j = 1$. This leaves $\mathbf{p}$ and $\mathbf{q}$ meaning the same eigenvectors normalised such that the last vector component is 1.

3.2 A symmetric matrix

As an example, take

$$\mathbf{B} = \begin{pmatrix} 1 & 1 & 3 & -1 \\ 1 & 2 & 5 & 1 \\ 3 & 5 & -2 & 3 \\ -1 & 1 & 3 & -2 \end{pmatrix}$$

and let us see what we can learn about its eigenvalues without solving the characteristic equation. The trace is -1. Brauer's row-and-column sum test gives all $|\lambda| < 13$. In Wolkowicz and Styan's statistics criterion the trace of B^2 is 105, so $m = -1/4$, $s = 5\cdot12$. The intervals in which the four eigenvalues are predicted to lie overlap:

$$-9\cdot1 < \lambda < -3\cdot2\,, \qquad -5\cdot4 < \lambda < 2\cdot7\,, \qquad -3\cdot2 < \lambda < 4\cdot9\,, \qquad 2\cdot7 < \lambda < 8\cdot6\,.$$

Using elementary row addition and multiplication operations I have transformed $\mathbf{B}$ to a row-equivalent triangular matrix

$$\mathbf{T} = \begin{pmatrix} 1 & 1 & 3 & -1 \\ 0 & 1 & 2 & 2 \\ 0 & 0 & -15 & 2 \\ 0 & 0 & 0 & -101 \end{pmatrix}.$$

Note that no target row is multiplied by a negative number in these row operations as that would change the sign of the pivot. The pivots are the diagonal elements and their signs show that two

eigenvalues are positive, two negative. We can immediately revise the bounds from Wolkowicz and Styan to

$$-9\cdot 1 < \lambda < -3\cdot 2\,, \qquad -5\cdot 4 < \lambda < 0\,, \qquad 0 < \lambda < 4\cdot 9\,, \qquad 2\cdot 7 < \lambda < 8\cdot 6\,.$$

By subtracting a constant β from the diagonal elements, the value dividing positive from negative eigenvalues is moved by β. This allows us to test various ranges to see if an eigenvalue lies within a given range. For instance, take $\beta = 4$. The shifted matrix and its row-equivalent triangular matrix are

$$\begin{pmatrix} -3 & 1 & 3 & -1 \\ 1 & -2 & 5 & 1 \\ 3 & 5 & -6 & 3 \\ -1 & 1 & 3 & -6 \end{pmatrix} \rightarrow \begin{pmatrix} -3 & 1 & 3 & -1 \\ 0 & -30 & 108 & 12 \\ 0 & 0 & 93 & 22 \\ 0 & 0 & 0 & -599 \end{pmatrix}.$$

This has three negative pivots so only one of $\lambda_j - 4$, $j = 1,\ 4$ is positive: that is, only one eigenvalue is > 4. Since there are two < 0, there must be exactly one in the interval $(0,\ 4)$. Similarly shifting by adding 4 gives

$$\begin{pmatrix} 5 & 1 & 3 & -1 \\ 1 & 6 & 5 & 1 \\ 3 & 5 & 2 & 3 \\ -1 & 1 & 3 & 2 \end{pmatrix} \rightarrow \begin{pmatrix} 5 & 1 & 3 & -1 \\ 0 & 29 & 22 & 6 \\ 0 & 0 & -7 & 6 \\ 0 & 0 & 0 & 27 \end{pmatrix}$$

which has only one negative pivot. So one of $\lambda_j + 4$ is < 0, or one < -4. Taking all this evidence together we know that the eigenvalues are ordered

$$-9\cdot 1 < \lambda < -4\,, \qquad -4 < \lambda < 0\,, \qquad 0 < \lambda < 4\,, \qquad 4 < \lambda < 8\cdot 6\,.$$

This case study illustrates how we can gain a rough idea of the positions of the eigenvalues on the real number line. This may guide the choice of algorithms and indeed the whole strategy for solving the problem. However, we still have no idea as to the eigenvectors.

Solving in the classic way, the characteristic equation is $\lambda^4 + \lambda^3 - 52\lambda^2 - 47\lambda + 101 = 0$ with solutions

$$\lambda_1 = -7\cdot 1056967\,, \qquad \lambda_2 = 7\cdot 0423731\,, \qquad \lambda_3 = -1\cdot 9642282\,, \qquad \lambda_4 = 1\cdot 0275518$$

where they are indexed in order of absolute value. These values lie within the predicted intervals. The corresponding eigenvectors are

$$\mathbf{p}_1 = \begin{pmatrix} 0\cdot 6705359 \\ 0\cdot 7691055 \\ -1\cdot 7347554 \\ 1 \end{pmatrix}, \qquad \mathbf{p}_2 = \begin{pmatrix} 1\cdot 5949580 \\ 3\cdot 0293222 \\ 2\cdot 5360030 \\ 1 \end{pmatrix}, \qquad \mathbf{p}_3 = \begin{pmatrix} 0\cdot 2432323 \\ -0\cdot 7435009 \\ 0\cdot 3408350 \\ 1 \end{pmatrix}, \qquad \mathbf{p}_4 = \begin{pmatrix} -2\cdot 0849973 \\ 0\cdot 6999624 \\ 0\cdot 0808640 \\ 1 \end{pmatrix}.$$

The pair-wise dot products of these are all zero to within machine accuracy, proving that these are a mutually orthogonal set. This is always the case with symmetric matrices.

3.3 Products of symmetric matrices

Matrices of the form $\mathbf{AB}$, where $\mathbf{A}$ and $\mathbf{B}$ are symmetric, occur in finite element calculations as $\mathbf{M^{-1}K}$ where $\mathbf{M}$ and $\mathbf{K}$ represent the mass and stiffness distributions of the structure. Here is a simple contrived example to illustrate that the eigenvectors are orthogonal with respect to weight

functions $\mathbf{A}^{-1}$ and $\mathbf{B}$. Two arbitrary symmetric matrices will in general give complex eigenvalues, so I have chose $\mathbf{A}$ and $\mathbf{B}$ so that $\mathbf{E} = \mathbf{AB}$ has only real ones. Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 4 & 0 \\ 1 & 2 & 5 & 0 \\ 4 & 5 & -2 & 1 \\ 0 & 0 & 1 & 6 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 1 & 3 & -1 \\ 1 & 2 & 5 & 1 \\ 3 & 5 & -2 & 3 \\ -1 & 1 & 3 & -2 \end{pmatrix}.$$

We also need to know that

$$\mathbf{A}^{-1} = \frac{1}{115}\begin{pmatrix} 176 & -133 & 18 & -3 \\ -133 & 109 & 6 & -1 \\ 18 & 6 & -6 & 1 \\ -3 & -1 & 1 & 19 \end{pmatrix}, \qquad \mathbf{E} = \mathbf{AB} = \begin{pmatrix} 14 & 23 & 0 & 12 \\ 18 & 30 & 3 & 16 \\ 2 & 5 & 44 & -7 \\ -3 & 11 & 16 & -9 \end{pmatrix}.$$

The characteristic equation for $\mathbf{E}$ is $\lambda^4 - 79\lambda^3 + 1107\lambda^2 + 17062\lambda - 11615 = 0$ with roots $50 \cdot 43142$, $37 \cdot 34133$, $-9 \cdot 42702$, $0 \cdot 65427$ and respective eigenvectors $\mathbf{p_j}$

$$\begin{pmatrix} 2 \cdot 27800 \\ 3 \cdot 08656 \\ 2 \cdot 01958 \\ 1 \end{pmatrix}, \qquad \begin{pmatrix} -3 \cdot 2814 \\ -3 \cdot 85184 \\ 4 \cdot 92921 \\ 1 \end{pmatrix}, \qquad \begin{pmatrix} -0 \cdot 183353 \\ -0 \cdot 334982 \\ 0 \cdot 169233 \\ 1 \end{pmatrix}, \qquad \begin{pmatrix} -1 \cdot 30900 \\ 0 \cdot 237810 \\ 0 \cdot 194459 \\ 1 \end{pmatrix}.$$

Incidentally, there is no simple relationship between the eigenvalues and eigenvectors of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{E}$. The point to be demonstrated is that $\mathbf{p_i}^T \mathbf{A}^{-1} \mathbf{p_j} = 0$ and $\mathbf{p_i}^i \mathbf{B} \mathbf{P_j} = 0$ unless $i = j$. I have confirmed this for the above matrices, obtaining products of typically 10^{-14}, which is essentially 0 to machine accuracy.

Here is a proof of this intriguing fact. The eigenvalue equation is $\mathbf{A}^{-1}\,\mathbf{B}\,\mathbf{p}_i = \lambda_i \mathbf{p}_i$. Use the property of transposes that $\mathbf{CD}^T = \mathbf{D}^T\mathbf{C}^T$ to show that

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

Since $\mathbf{A}$, and hence $\mathbf{A}^{-1}$, and $\mathbf{B}$ are symmetric,

$$\mathbf{p}_i^T\,\mathbf{B}\,\mathbf{A}^{-1} = \lambda_i \mathbf{p}_i^T\,.$$

Now multiply on the right by $\mathbf{A}\,\mathbf{p}_j$ to get

$$\mathbf{p}_i^T\,\mathbf{B}\,\mathbf{p}_j = \lambda_i \mathbf{p}_i^T\,\mathbf{A}\,\mathbf{p}_j\,. \tag{11a}$$

Swap the labels and take the transform

$$(\mathbf{p}_j^T\,\mathbf{B}\,\mathbf{p}_i)^T = \lambda_j (\mathbf{p}_j^T\,\mathbf{A}\,\mathbf{p}_i)^T.$$

$$\mathbf{p}_i^T\,\mathbf{B}\,\mathbf{p}_j = \lambda_j \mathbf{p}_i^T\,\mathbf{A}\,\mathbf{p}_j\,. \tag{11b}$$

The left sides of Eq 6a and b are the same and their right sides differ only through the subscripts on λ. But the eigenvalues λ_i and λ_j are not equal, so their multiplying factors must be zero.

3.4 A degenerate matrix

Here is an example of a matrix with three equal eigenvalues: $\lambda_1 = \lambda_2 = \lambda_3 = 3$, $\lambda_4 = 1$.

$$\mathbf{A} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ -2 & 2 & 0 & 1 \\ 0 & 0 & 3 & 0 \\ 2 & 1 & 0 & 2 \end{pmatrix}.$$

The trace is 10, the determinant 27 and the characteristic equation is $(\lambda-3)^3(\lambda-1) = 0$. Substituting $\lambda = 1$ into $(\mathbf{A} - \lambda\mathbf{I})\mathbf{p}$ gives the eigenvector $\mathbf{p}_4 = (0, -1, 0, 1)$. Substituting $\lambda = 3$ gives only one relation amongst the vector components $2p_{j1} + p_{j2} - p_{j4} = 0$, $j = 1, 3$. Geometrically, this is the equation of a plane and so must be spanned by two independent vectors. There is wide scope to choose two basis vectors for this space, but $(0, 1, 1, 1)$ and $(\frac{1}{4}, \frac{1}{2}, 0, 1)$ will serve – they both have the last component equal to 1. Alternatively the Gram-Schmidt procedure can be used to produce two orthonormal base vectors

$$\mathbf{b_1} = \begin{pmatrix} 0 \cdot 574285 \\ -0 \cdot 607517 \\ 0 \cdot 091549 \\ 0 \cdot 541053 \end{pmatrix}, \quad \mathbf{b_2} = \begin{pmatrix} -0 \cdot 034508 \\ 0 \cdot 360020 \\ 0 \cdot 885726 \\ 0 \cdot 291005 \end{pmatrix}.$$

We see that each of the three equal eigenvalues does not have associated with it a unique eigenvector, but instead the three share a space, here equivalent to a plane. Degenerate matrices are likely to arise in engineering and physics where the structure being described has rotational or mirror symmetry.

4 The direct power method

The Power Method was known in Victorian times as an iterative procedure converging on the eigenvalue with the largest absolute value and simultaneously upon its eigenvector. Indeed, the algorithm focuses on the eigenvector and produces the eigenvalue as a by-product. It exists to two forms, direct and inverse, the latter being described in the next section, §4. As above, the largest absolute eigenvalue is called λ_1, the second λ_2, *etc.*

4.1 Matrix multiplication in the basic power method

The basic power method operates as follows. Using the example in §3.1, apply $\mathbf{E}$ of Eq 7 repeatedly to the vector $\mathbf{v} = (1, 1, 1)$. This is a fairly arbitrary starting place, though the linear expansion in terms of the eigenvectors at Eq 8 does show that $\mathbf{p}_1$ makes the largest contribution. Some values are

$$\mathbf{E}^3\mathbf{v} = \begin{pmatrix} -18 \\ 278 \\ 210 \end{pmatrix}, \quad \mathbf{E}^4\mathbf{v} = \begin{pmatrix} 146 \\ 1656 \\ 1346 \end{pmatrix}, \quad \mathbf{E}^5\mathbf{v} = \begin{pmatrix} 142 \\ 10498 \\ 8240 \end{pmatrix}, \quad \mathbf{E}^6\mathbf{v} = \begin{pmatrix} 3156 \\ 64596 \\ 51556 \end{pmatrix}.$$

This looks opaque until the matrices are normalised by scaling so that, say, the bottom component is 1. Some values are

$$\mathbf{E}^3\mathbf{v} \to \begin{pmatrix} -0 \cdot 08571 \\ 1 \cdot 32381 \\ 1 \end{pmatrix}, \quad \mathbf{E}^4\mathbf{v} \to \begin{pmatrix} 0 \cdot 10847 \\ 1 \cdot 23031 \\ 1 \end{pmatrix}, \quad \mathbf{E}^5\mathbf{v} \to \begin{pmatrix} 0 \cdot 01723 \\ 1 \cdot 27403 \\ 1 \end{pmatrix}, \quad \mathbf{E}^6\mathbf{v} \to \begin{pmatrix} 0 \cdot 06121 \\ 1 \cdot 25293 \\ 1 \end{pmatrix}.$$

By iteration 20 the agreement with $\mathbf{p_1}$ is correct to 6 decimal places. If a further multiplication is make on the normalised vector we get

$$\mathbf{v}_{20} = \begin{pmatrix} 0 \cdot 0470544 \\ 1 \cdot 2597199 \\ 1 \end{pmatrix}, \qquad \mathbf{E}\mathbf{v}_{20} \approx \lambda_1\mathbf{v}_{20} = \begin{pmatrix} 0 \cdot 2923380 \\ 7 \cdot 8262135 \\ 6 \cdot 2126633 \end{pmatrix}.$$

The bottom component is the largest eigenvalue, λ_1, correct almost to 6 decimal places. In practice the normalisation would be performed at each iteration and the cycle stopped when the difference between successive iterations is less than, say, 10^{-8}.

The subtleties are in encouraging the method to converge fairly rapidly whatever the values of the eigenvectors. I now give some algebraic analysis of convergence. Using superscripts to index the iterations, suppose that the starting guess is the vector in Eq 2,

$$\mathbf{v}^{(0)} \;=\; c_1\mathbf{p}_1 \,+\, c_2\mathbf{p}_2 \,+\, \ldots. c_n\mathbf{p}_n \qquad \textit{Copy of } (2)$$

where the last component of each $\mathbf{p}_j$ is 1 and the eigenvalues are ordered $|\lambda_1| > |\lambda_2| > \ldots. > |\lambda_n|$. Multiply by $\mathbf{E}$ and renormalise:

$$\mathbf{v}^{(1)} \;=\; \frac{c_1\lambda_1\mathbf{p}_1 \,+\, c_2\lambda_2\mathbf{p}_2 \,+\, \ldots. c_n\lambda_n\mathbf{p}_n}{c_1\lambda_1 + c_2\lambda_2 + \ldots.. + c_n\lambda_n}\,.$$

The weightings of the eigenvectors are now

$$c'_j \;=\; \frac{c_j\lambda_j}{\sum_{k=1} c_k\lambda_k}\,, \qquad 1 \le j \le n.$$

After a second iteration

$$c''_j \;=\; \frac{c'_j\lambda_j}{\sum_{k=1} c'_k\lambda_k} \;=\; \frac{c_j\lambda_j^2}{\sum c_k\lambda_k \,\sum c'_k\lambda_k} \;=\; \frac{c_j\lambda_j^2}{\sum c_k\lambda_k^2}\,.$$

After m iterations

$$c_j^{(m)} \;=\; \frac{c_j\lambda_j^m}{\sum_k c_k\lambda_k^m}\,. \qquad (12)$$

In particular the coefficients of $\mathbf{p}_1$ and $\mathbf{p}_2$ are

$$c_1^{(m)} \;=\; \frac{\frac{c_1}{c_2}\left(\frac{\lambda_1}{\lambda_2}\right)^m}{\frac{c_1}{c_2}\left(\frac{\lambda_1}{\lambda_2}\right)^m + 1 + \ldots..}\,, \qquad c_2^{(m)} \;=\; \frac{\frac{c_2}{c_1}\left(\frac{\lambda_2}{\lambda_1}\right)^m}{1 \,+\, \frac{c_2}{c_1}\left(\frac{\lambda_2}{\lambda_1}\right)^m + \ldots..}\,.$$

Even though c_1 may be much smaller that c_2 (a possibility examined below), it is clear that $(\lambda_1/\lambda_2)^m$ will grow indefinitely with m while $(\lambda_2/\lambda_1)^m$ will shrink. Therefore $c_1^{(m)} \to 1$ and $c_j^{(m)} \to 0,\ j \ge 2$, and $\mathbf{v}^{(m)} \to \mathbf{p}_1$. Subject to control of rounding errors, the iterations will almost certainly converge on the eigenvector for the largest eigenvalue, provided $|\lambda_2| \ne \lambda_1$.

What starting vector would give the worse convergence? Theoretically if $c_1 = 0$, convergence to $\mathbf{p}_1$ should be impossible. Because the eigenvectors are not orthogonal, it is not sufficient to invent a starting vector $\mathbf{v}^{(0)}$ which is just a linear combination of $\mathbf{p}_2$ and $\mathbf{p}_2$; it is also necessary for it be to normal to $\mathbf{p}_1$. The vector with these properties is $(9\cdot 4104456,\ -1\cdot 1453442,\ 1)$. After only two multiplications by $\mathbf{E}$ it is clear that the pull of λ_1 has been suppressed and that the $\mathbf{v}^{(j)}$ are tending towards the second largest eigenvector, $\mathbf{p}_2$. Closest approach to $\mathbf{p}_2$ is attained at 12 iterations where the error in λ_2 is about 3×10^{-7}. Perhaps surprisingly, with more iterations it diverges from $\mathbf{p}_2$ in the direction of $\mathbf{p}_1$. After 50 iterations the method has λ_1 and its eigenvector correct to 3 decimal places. The power method, therefore, has an inevitability pull towards the eigenvector whose eigenvalue has the largest magnitude.

Even with a favourable choice of starting vector the rate of convergence will depend strongly on λ_1/λ_2 and to a lesser extent on c_1/c_2. To emphasise these, let us evaluate the difference $\delta_{m+1} =$

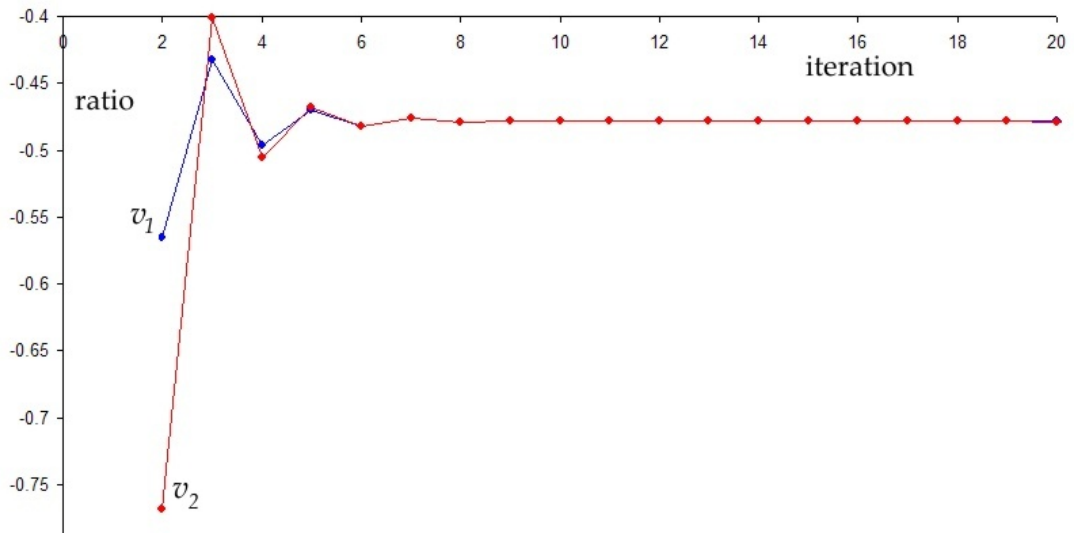

Figure 1: Ratio of differences of two components of $\mathbf{v}^{(m)} - \mathbf{v}^{(m-1)}$ over 20 iterations.

$\mathbf{v}^{(m+1)} - \mathbf{v}^{(m)}$ between two consecutive estimates of the eigenvector. For $\lambda_1 > \lambda_2$ the denominator of Eq 7 can be expanded as a Taylor series:

$$\frac{1}{c_1\lambda_1^m + \sum_{k\geq 2} c_k\lambda_k^m} \approx \frac{1}{c_1\lambda_1^m}\left[1 - \Sigma_m + \Sigma_m^2 - ...\right], \qquad \Sigma_m = \sum_{k\geq 2}\frac{c_k}{c_1}\left(\frac{\lambda_k}{\lambda_1}\right)^m .$$

$$\delta_{m+1} \equiv \mathbf{v}^{(m+1)} - \mathbf{v}^{(m)} =$$

$$(c_1\lambda_1^{m+1}\mathbf{p}_1 + \sum_{k\geq 2} c_k\lambda_k^{m+1}\mathbf{p}_k)\frac{1}{c_1\lambda_1^{m+1}}[1 - \Sigma_{m+1} + ...] - (c_1\lambda_1^m\mathbf{p}_1 + \sum_{k\geq 2} c_k\lambda_k^m\mathbf{p}_k)\frac{1}{c_1\lambda_1^m}[1 - \Sigma_m + ...].$$

$$= [\Sigma_m - \Sigma_{m+1}]\mathbf{p}_1 + \sum_{k\geq 2}\frac{c_k}{c_1}\left(\frac{\lambda_k}{\lambda_1}\right)^m\left\{[1 - \Sigma_{m+1}]\frac{\lambda_k}{\lambda_1} - [1 - \Sigma_m]\right\}\mathbf{p}_k$$

$$= [\Sigma_m - \Sigma_{m+1}]\mathbf{p}_1 + \sum_{k\geq 2}\frac{c_k}{c_1}\left(\frac{\lambda_k}{\lambda_1}\right)^m\left\{\left(1 - \frac{\lambda_k}{\lambda_1}\right)(\Sigma_m - 1) + \frac{\lambda_k}{\lambda_1}(\Sigma_m - \Sigma_{m+1})\right\}\mathbf{p}_k$$

$$\text{where} \quad \Sigma_m - \Sigma_{m+1} = \sum_{k\geq 2}\frac{c_k}{c_1}\left(\frac{\lambda_k}{\lambda_1}\right)^m\left(1 - \frac{\lambda_k}{\lambda_1}\right).$$

The factor $(\lambda_k/\lambda_1)^m$ features strongly here, but the expression does not become simple unless $|c_3|$ is sufficiently small compared with $|c_2|$ to be neglected. In other words, the three largest eigenvalues must be sufficiently widely separated. When this condition holds,

$$\mathbf{v}^{(m+1)} - \mathbf{v}^{(m)} \approx \frac{c_2}{c_1}\left(\frac{\lambda_2}{\lambda_1}\right)^m\left(1 - \frac{\lambda_2}{\lambda_1}\right)(\mathbf{p}_1 - \mathbf{p}_2) + ... \tag{13}$$

where the .... indicate terms in $\mathbf{p}_1 - \mathbf{p}_3$ and in $(\lambda_2/\lambda_1)^{2m}$. Eq 8 expresses a geometric series with common ratio (λ_2/λ_1). It will alternate if λ_1 and λ_2 have opposite signs. If $|\lambda_2| \approx |\lambda_1|$, convergence will be very slow and the method may drown in rounding errors. If $\lambda_2 \approx \lambda_1$, little can be done, but if they differ in sign it is highly advantageous to shift the diagonal elements by a constant β, as at item 5 of §2.1 and discussed in §3.2 below.

As numerical evidence for a geometric series, Figure 1 plots in blue the ratio

$$\frac{v_1^{(m)} - v_1^{(m-1)}}{v_1^{(m-1)} - v_1^{(m-2)}}$$

of the first component $v_1^{(m)}$ of $\mathbf{v}^{(m)}$ obtained with the matrix $\mathbf{E}$ of §3.1. The red points are ratios for the second component $v_2^{(m)}$. (The third component is normalised to 1 for all $\mathbf{v}^{(m)}$.) Even at iteration 5 the ratios for the two components are close, being respectively $-0\cdot470$ and $-0\cdot468$. At iteration 10 they agree on $-0\cdot478430$ to 6 decimal places. At this stage the λ_1 eigenvalue has been determined as $6\cdot215$ so the common ratio tells us that λ_2 is close to $-2\cdot973$. Moreover, if all we were interested in were the eigenvalues and not the other two eigenvectors for this 3×3 matrix $\mathbf{E}$, λ_3 is Trace $-\lambda_1 - \lambda_2 = 0\cdot758$. All three eigenvalues have been found from one short sequence of iterations on one starting vector.

4.1.1 Rayleigh quotients

I should point out that there is an alternative way of calculating the iterated eigenvalue using the Rayleigh quotient, introduced at item 3 in §2. Writing μ_k for the estimated eigenvalue at the k^{th} iteration (so as not to confuse it with the λ obtained by direct multiplication)

$$\mathbf{E}\mathbf{v}^{(k)} = \mu_k\mathbf{v}^{(k)} \quad \text{so} \quad \mathbf{v}^{(k)T}\mathbf{E}\mathbf{v}^{(k)} = \mu_k\mathbf{v}^{(k)T}\mathbf{v}^{(k)} \quad \text{and} \quad \mu_k = \frac{\mathbf{v}^{(k)T}\mathbf{E}\mathbf{v}^{(k)}}{\mathbf{v}^{(k)T}\mathbf{v}^{(k)}}\,.$$

The expression on the right is the Rayleigh quotient. It involves extra calculation beyond determining $\mathbf{E}\mathbf{v}^{(k)}$ and so is a half-way-house between iterations k and $k+1$ of the direct power method. To illustrate this Figure 2 plots the natural logarithm of the absolute error in the estimated eigenvalue of our example matrix $\mathbf{E}$ of Eq 7. The number of the iteration forms the horizontal axis and I have plotted the points for the direct power method at integer values and the Rayleigh quotient values at the next $\frac{1}{2}$. For this case Rayleigh quotient is slightly more accurate than the next direct power estimate, though it converges at the same rate. It is a matter of judgement whether it is worth the extra effort in calculation compared with making another couple of multiplications by $\mathbf{E}$.

4.2 Predicted values using summed geometric series

There is considerable scope to improve the estimation of the eigenvector components and leading eigenvalue using the fact that the differences between successive iterations form geometric series. I present here a summary and refer the reader to Appendix 1 for supporting analysis. The essential concept is that if the first term $\mathcal{C}$ and common ratio r of a geometric series can be identified, then the sum to infinity, S, is given by the formula we learned at school

$$S = \frac{\mathcal{C}}{1-r}, \qquad |r| < 1\,.$$

This allows the value currently estimated by the basic power method – by multiplication by $\mathbf{E}$ – to be projected as if through an infinity of further iterations to the true value. Indeed if the differences δ_m did form exactly a single geometric series, the precise values of $\mathbf{p}$ and λ would be known after only three consecutive iterates. I have not seen this aspect of the Power Method described in textbooks or the literature, but it seems quite obvious so I suppose it to be well known. The analysis here is my own.

Consider the first order approximation where a single geometric series is used to project a value for the eigenvalue. When three consecutive evaluations, $\lambda^{(1)}$, $\lambda^{(2)}$, $\lambda^{(3)}$, have been made by

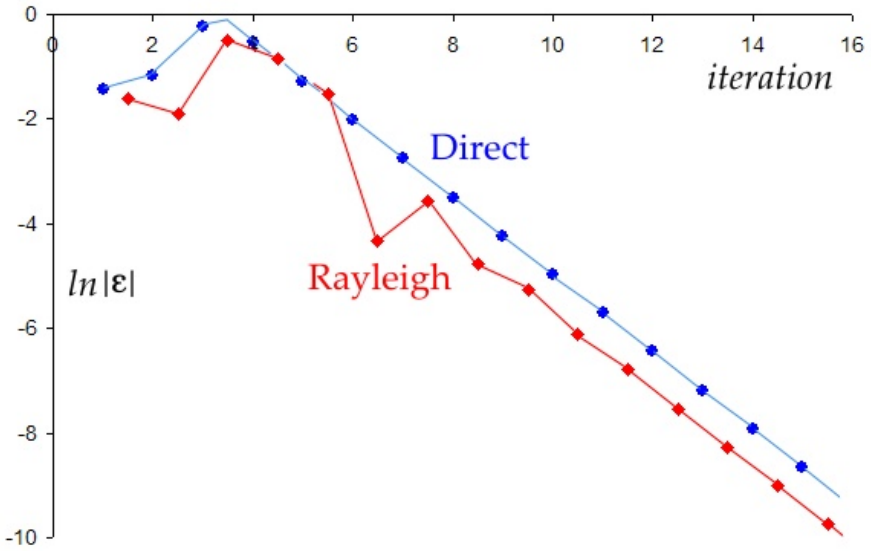

Figure 2: Error in estimated value of $\lambda_1 = 6 \cdot 2126641$. $\ln|\varepsilon|$ versus iteration number. Blue: direct power method. Red: Rayleigh quotient. Starting vector was $(1, 1, 1)$. The errors alternate in sign.

k	$\lambda^{(k)}$	δ_k	r_k	$\Sigma \times 100$	λ_∞	ε_A	ε_B	ratio
1	5							
2	$7 \cdot 2$	$2 \cdot 2$						
3	$5 \cdot 833333$	$-1 \cdot 366667$	$-0 \cdot 62121$	$-84 \cdot 29907$	$6 \cdot 35700934$	$0 \cdot 1444$	$-0 \cdot 3793$	$0 \cdot 381$
4	$6 \cdot 409524$	$0 \cdot 576190$	$-0 \cdot 42160$	$40 \cdot 531046$	$6 \cdot 23864379$	$2 \cdot 60_{E-2}$	$0 \cdot 197$	$0 \cdot 132$
5	$6 \cdot 121843$	$-0 \cdot 287681$	$-0 \cdot 49928$	$-19 \cdot 187944$	$6 \cdot 21764437$	$4 \cdot 98_{E-3}$	$-9 \cdot 08_{E-2}$	$0 \cdot 055$
6	$6 \cdot 256796$	$0 \cdot 134954$	$-0 \cdot 46911$	$9 \cdot 186092$	$6 \cdot 21370342$	$1 \cdot 04_{E-3}$	$4 \cdot 41_{E-2}$	$0 \cdot 024$
7	$6 \cdot 191714$	$-0 \cdot 065082$	$-0 \cdot 48226$	$-4 \cdot 390756$	$6 \cdot 21288856$	$2 \cdot 25_{E-4}$	$-2 \cdot 10_{E-2}$	$0 \cdot 011$
8	$6 \cdot 222718$	$0 \cdot 031004$	$-0 \cdot 47638$	$2 \cdot 099999$	$6 \cdot 21271385$	$4 \cdot 98_{E-5}$	$1 \cdot 01_{E-2}$	$5 \cdot 0_{E-3}$
9	$6 \cdot 207864$	$-0 \cdot 014854$	$-0 \cdot 47910$	$-1 \cdot 004263$	$6 \cdot 21267524$	$1 \cdot 12_{E-5}$	$-4 \cdot 80_{E-3}$	$2 \cdot 3_{E-3}$
10	$6 \cdot 214961$	$0 \cdot 007098$	$-0 \cdot 47783$	$0 \cdot 480277$	$6 \cdot 21266658$	$2 \cdot 54_{E-6}$	$2 \cdot 30_{E-3}$	$1 \cdot 1_{E-3}$
11	$6 \cdot 211566$	$-0 \cdot 003396$	$-0 \cdot 47843$	$-0 \cdot 229686$	$6 \cdot 21266462$	$5 \cdot 77_{E-7}$	$-1 \cdot 10_{E-3}$	$5 \cdot 3_{E-4}$
12	$6 \cdot 213189$	$0 \cdot 001624$	$-0 \cdot 47814$	$0 \cdot 109844$	$6 \cdot 21266418$	$1 \cdot 31_{E-7}$	$5 \cdot 25_{E-4}$	$2 \cdot 5_{E-4}$
13	$6 \cdot 212413$	$-0 \cdot 000777$	$-0 \cdot 47828$	$-0 \cdot 052531$	$6 \cdot 21266408$	$2 \cdot 93_{E-8}$	$-2 \cdot 51_{E-4}$	$1 \cdot 2_{E-4}$
14	$6 \cdot 212784$	$0 \cdot 000371$	$-0 \cdot 47821$	$0 \cdot 025122$	$6 \cdot 21266405$	$4 \cdot 85_{E-9}$	$1 \cdot 20_{E-4}$	$4 \cdot 0_{E-5}$

Table 1: Comparison of two convergence scheme for λ_1 of $\mathbf{E}$: A) forward projection by sum of geometric series, B) multiplication by $\mathbf{E}$. Correct value is $6 \cdot 212664048$.

the direct power method, they furnish two differences δ and one ratio of differences, and so project to a value of $\lambda^{(\infty)}$, the supposed ultimate value:

$$\delta_2 = \lambda^{(2)} - \lambda^{(1)}, \qquad \delta_3 = \lambda^{(3)} - \lambda^{(2)}, \qquad r = \frac{\delta_3}{\delta_2}.$$

$$\Sigma = \sum_{k=2}^{\infty} \delta_k = \frac{\delta_2}{1-r}, \qquad \lambda^{(\infty)} = \lambda^{(1)} + \Sigma = \frac{\lambda^{(2)} - r\lambda^{(1)}}{1-r} \approx \frac{\lambda^{(3)} - r\lambda^{(2)}}{1-r}. \tag{14}$$

This final formula on the right should give an even better estimate than $[\lambda^{(2)} - r\lambda^{(1)}]/(1-r)$.

An example is shown in detail in Table 1. It lists the results of two methods for estimating the eigenvalue $\lambda_1 = 6{\cdot}212664048$ for matrix **E**, Eq 7. The first starting vector was (1, 1, 1). However the parameters of the geometric series are calculated in a rolling way from the most recent three multiplications by **E**. Thus $\lambda^{(\infty)}$ at iteration 8, say, uses $\lambda^{(6)}$, $\lambda^{(7)}$ and $\lambda^{(8)}$. The columns in Table 1 are:

1. the iteration number of the basic power method,
2. the estimate of eigenvalue by the direct power method, by multiplying $\lambda^{(k-1)}$ by **E**,
3. the difference $\delta_k = \lambda^{(k)} - \lambda^{(k-1)}$,
4. the ratio $r = \delta_k/\delta_{k-1}$,
5. the sum to infinity of the geometric series formula in Eq 14, multiplied by 100 for convenience of reading,
6. the projected value, $\lambda^{(\infty)}$, of the eigenvalue,
7. the error $\varepsilon_A = \lambda^{(\infty)} - 6 \cdot 212664048$ in the geometric series scheme,
8. the error $\varepsilon_B = \lambda^{(k)} - 6 \cdot 212664048$ in the current estimate of the power method for comparison,
9. the ratio of errors $\varepsilon_A/\varepsilon_B$ in the two methods.

The sequence of $\lambda^{(\infty)}$ clearly converges much more rapidly with iteration than the sequence from the basic power method. By iteration 7 it is 100 times more accurate, by iteration 10, 1000 times. Note also that the errors ε_A in the geometric series projection all have the same sign. The process can also be applied to the components of the eigenvector and gives the same rapid convergence.

Prediction using geometric series can be made more sophisticated. Appendix 1 shows that the difference between iterations $\delta_m = \mathbf{v}_{m+1} - \mathbf{v}_m$ is more accurately given by the five terms

$$\begin{aligned} & Cr^m(1-r)(\mathbf{p_1}-\mathbf{p_2}) \;+\; Ds^m(1-s)(\mathbf{p_1}-\mathbf{p_3}) \\ +\; C^2r^{2m}(1-r^2)\mathbf{p_2} \;+\; & CDr^ms^m(1-rs)(\mathbf{p_2}+\mathbf{p_3}) \;+\; D^2s^{2m}(1-s^2)\mathbf{p_3}\,. \qquad (15) \\ C \;=\; \frac{c_2}{c_1}, \qquad & D \;=\; \frac{c_3}{c_1}, \qquad r \;=\frac{\lambda_2}{\lambda_1}, \qquad s \;=\frac{\lambda_3}{\lambda_1} \end{aligned}$$

where the c_1, c_2, c_3 are the contributions of the three eigenvectors to the starting vector[3] as at Eq 2. The geometric series given at Eq 14 above come from the first term in Eq 15, with common ratio r. Clearly there are four other series here with ratios s, r^2, rs and s^2, and further, less significant series. It is possible in principle to determine the first term and common ratio of each of these contributing series by solving a set of non-linear simultaneous equations. This can be achieved using a multi-variable version of Newton's iterative method, given with a sufficiently close initial guess of the ratios λ_2/λ_1 and λ_3/λ_1 obtained either from the series itself or from the bounds on eigenvalues obtained in the preliminary survey. In Appendix 1 I shows how a double and a triple geometric series can be fitted to a sequence of five or six (respectively) values of eigenvector or eigenvalue iterate. Using the notation

$$C(1-r)(p_1-p_2) = \mathcal{C}, \qquad D(1-s)(p_1-p_3) = \mathcal{D}, \qquad C^2(1-r^2)p_2 = \mathcal{E}\,. \qquad (\textit{copy of A1.4})$$

[3] Starting vector here means the first in the current sequence of values used in calculating the geometric series. This increments at each iteration of the power method to give a running sequence of values being used.

iteration	power method	single series	double series	triple series
3	$5\cdot833333$			
4	$6\cdot409524$			
5	$6\cdot121842$	$6\cdot217644372$		
6	$6\cdot256796$	$6\cdot213703418$		
7	$6\cdot191714$	$6\cdot212888564$	$6\cdot212664546$	
8	$6\cdot222718$	$6\cdot212713857$	$6\cdot212667297$	$6\cdot212668260$
9	$6\cdot207864$	$6\cdot212675248$	$6\cdot212664177$	$6\cdot212663888$
10	$6\cdot214961$	$6\cdot212666586$	$6\cdot212664093$	$6\cdot212664075$
11	$6\cdot211566$	$6\cdot212664625$	$6\cdot212664050$	$6\cdot212664046$
12	$6\cdot213189$	$6\cdot212664179$	$6\cdot212664048$	
3	$-0\cdot3793$			
4	$0\cdot1969$			
5	$-0\cdot0908$	$4\cdot98_{E-3}$		
6	$4\cdot41_{E-2}$	$1\cdot04_{E-3}$		
7	$-2\cdot10_{E-2}$	$2\cdot25_{E-4}$	$4\cdot98_{E-7}$	
8	$1\cdot01_{E-2}$	$4\cdot98_{E-5}$	$3\cdot25_{E-6}$	$4\cdot21_{E-6}$
9	$-4\cdot80_{E-3}$	$1\cdot12_{E-5}$	$1\cdot29_{E-7}$	$-1\cdot59_{E-7}$
10	$2\cdot30_{E-3}$	$2\cdot54_{E-6}$	$4\cdot57_{E-8}$	$2\cdot71_{E-8}$
11	$-1\cdot10_{E-3}$	$5\cdot78_{E-7}$	$2\cdot69_{E-9}$	$-1\cdot73_{E-9}$
12	$5\cdot25_{E-4}$	$1\cdot32_{E-7}$	$6\cdot50_{E-10}$	

Table 2: Example of convergence to the eigenvalue $\lambda_1 = 6\cdot212664048$ by four means: 1) simple multiplication by matrix **E**, Eq 7, 2) fitting a single geometric series, 3) fitting two geometric series, 3) three geometric series. Upper panel lists estimated values of λ_1, lower panel lists errors.

the double series is $\mathcal{C} + \mathcal{C}r + \mathcal{C}r^2 + \mathcal{C}r^3 + \ \cdots\ + \ \mathcal{D} + \mathcal{D}s + \mathcal{D}s^2 +$ ..... The triple series adds to this the series $\mathcal{E} + \mathcal{E}r^2 + \mathcal{E}r^4 + \ \cdots$.

Table 2 lists the projected values of eigenvalue λ_1 of matrix **E** using a) simple multiplication by matrix **E**, b) projection with the single series as in Table 1, c) projection with two geometric series with ratios $r = \lambda_2/\lambda_1$ and $s = \lambda_3/\lambda_1$, and d) projection with three series with ratios r, s and r^2. The upper panel lists the projected values of λ_1 and the lower panel lists the errors. I wrote a computer program to obtain these values. The parameters of the double series are calculated in a rolling way from the five most recent multiplications by **E**, and the triple series uses the most recent six. Appendix 1 gives a reasonable way of obtaining initial values for Newton's method so that the iterations converge. (To be clear, iterations of Newton's method are nested within the iterations of Power Method.) In this example the calculation was stopped at iteration 12 when the double series projection changed by less than 10^{-8} from one Power Method iteration to the next. The ratio $r = \lambda_2/\lambda_1 = -0\cdot478235769$ obtained for the double series gives $\lambda_2 = -2\cdot9711182$, very close to the true value of $-2\cdot9711194$, so essentially the second eigenvalue has been determined. The agreement from $s = 0\cdot2273$ with λ_3 is less good: $1\cdot41$ compared with 0.758. The eigenvector $\mathbf{p}_1$ was simultaneously calculated as $(0\cdot047055409236, 1\cdot25971945692, 1)$, correct to 11 decimal places.

The sequence of three-series projected values shows little advantage in rate of convergence over the double series, but the parameters can yield projected values of the second eigenvector $\mathbf{p}_2$. To see this note that the constant ($m = 0$) terms of the series with ratios r and r^2 are $\mathcal{C} = C(1-r)(p_1-p_2)$

and $\mathcal{E}C^2(1-r^2)p_2$ respectively. Given numerical values for these and of the component p_1 just calculated, we have two simultaneous quadratic equations in C and p_2. For example, in the calculation which gave Table 2 at iteration 11 the parameters for the first component of the dominant eigenvector were

$$\mathcal{C} = -0\cdot 02098\,, \qquad \mathcal{E} = 0\cdot 00002338\,, \qquad r = -0.478236\,.$$

The solution is $C = c_2/c_1 = 0\cdot 002121$, $p_2 = 6\cdot 739$. If you refer back the exact eigenvector as Eq 8, the correct value is about $6\cdot 254$. The parameters fitted to the triple geometric series for iterations 8 to 11 are listed in Table 3. p_{21} refers to the first component of the eigenvector $\mathbf{p}_2$, and p_{22} to the second, the third being normalised to 1. A few things to note in this table are

- the values of $r = \lambda_2/\lambda_1$ are very close and consistent and agree to about 6 decimal places with the r values in the double series.
- the values of $s = \lambda_3/\lambda_1$ are more scattered. Averages are shown in the right panel. These averages give $\lambda_3 = 0\cdot 87$, to compare with the true value of $0\cdot 758$. Although the double series has a much more consistent value of s at $0\cdot 227$, the projected $\lambda_3 = 1\cdot 41$ is about twice the correct value.

The above solution of two simultaneous equations can be done for each iteration, so I list the solutions, $C = c_2/c_1$ and p_{21}, in Table 4. The first column states the component of $\mathbf{p}_2$ and the second names the variables in the simultaneous equations. With the first eigenvector component solutions at all four iterations are consistent. Their average is $6\cdot 2726$ which compares well with the correct value of $6\cdot 264$ (refer to Eq 8, §2.2.1). It is disappointing to find that the solutions for the second vector component are complex. I have no explanation for this. However in any iteration the value of C obtained for component p_{21} should also hold for p_{21}. Using this C with the values of $\mathbf{C}$ in the p_{22} panel of Table 3 give the values in the bottom row of Table 4. These average at $-1\cdot 726$, compared with the true value of $-1\cdot 717$.

To summarise, by fitting a double and a triple geometric series to four or then five consecutive values of the differences δ maximum information has been wrung from the power series iterations. The basic power method would have resulted in an estimate of $\mathbf{p}_1$ and λ_1 accurate after 12 iterations to 5×10^{-4}. After the same 12 iterations the double series has found $\mathbf{p}_1$ and λ_1 to 6×10^{-10}, and also λ_2 to 6 decimal places. The triple geometric series has added a useful approximation to the second eigenvector $\mathbf{p}_2$ as $(6\cdot 273,\ -1\cdot 726,\ 1)$. This is an excellent starting point for a power method search for precise $\mathbf{p}_2 = (6\cdot 254,\ -1\cdot 7171,\ 1)$. The weakness in fitting the double and triple geometric series is in obtaining convergence of Newton's method.

It is interesting to see what the power method with geometric series project makes of the degenerate matrix of §3.4. I ran the computer program I had written to implement the direct power method, giving three arbitrary starting vectors from which the program chooses the best after three iterations. At iteration 9 it stopped, having converged through the triple-series projection to $\lambda_1 = 3$ with error (that is, change from the last iteration) less that $2\cdot 5\times 10^{-9}$ and eigenvector

$$\begin{pmatrix} 1\cdot 0614213 \\ -1\cdot 1228426 \\ 0\cdot 1692047 \\ 1 \end{pmatrix}.$$

This does satisfy the relation amongst the vector component given in §3.4 that $2p_1 + p_2 = p_4$ for any p_3. The ratio r in the geometric series pointed to the second largest eigenvalue being $0\cdot 999985$,

	iteration	8	9	10	11	average
p_{21}	$\mathcal{C}$	0·191822	-0·09173	0·043869	-0·02098	
	$\mathcal{D}$	0·000231	$-1·627_{E-5}$	$3·085_{E-6}$	$1·936_{E-6}$	
	$\mathcal{E}$	0·002132	0·000512	0·00011	$2·338_{E-5}$	
	r	-0·47823	-0·47823	-0·47824	-0·47824	
	s	0·045245	0·284264	0·063567	0·218592	0·152917
p_{22}	$\mathcal{C}$	-0·09201	0·043997	-0·02104	0·010062	
	$\mathcal{D}$	-0·00045	$-5·849_{E-5}$	$-6·562_{E-6}$	$-8·435_{E-7}$	
	$\mathcal{E}$	-0·00105	-0·00022	$-5·292_{E-5}$	$-1·194_{E-5}$	
	r	-0·47822	-0·47823	-0·47824	-0·47824	
	s	0·096026	0·146348	0·105815	0·137759	0·121487
λ	$\mathcal{C}$	0·593755	-0·28383	0·13573	-0·06491	
	$\mathcal{D}$	0·017619	-0·00067	-0·00011	$-9·514_{E-6}$	
	$\mathcal{E}$	-0·03518	-0·00319	-0·00067	-0·00016	
	r	-0·47816	-0·47823	-0·47823	-0·47824	
	s	0·242613	0·077125	0·176494	0·090935	0·146792

Table 3: Parameters of three geometric series fitted to differences δ of two components v_1, v_2 of eigenvector and to eigenvalue over iterations 8 to 11.

		8	9	10	11	average
p_{21}	C	-0·02114	0·010615	−0·00477	0·002121	
	p_2	6·186396	5·892949	6·272302	6·738926	6·272643
p_{22}	C	complex				
	p_2	-1·68495	-1·54413	-1·72608	-1·94989	-1·72626

Table 4: Projected components 1 and 2 of the second eigenvector $\mathbf{p}_2$ obtained from triple geometric series from iterations 8 to 11 of the basic power method. p_{22} calculated using C from p_{12}.

remarkably close to 1. Running the program again with different starting vectors gave convergence at iteration 9 to eigenvalue 3 but a different eigenvector

$$\begin{pmatrix} 0·9452984 \\ -0·8905967 \\ 0·4520796 \\ 1 \end{pmatrix}$$

which satisfies the same relation, $2p_1 + p_2 = p_4$. The fact that different starting vectors give the same eigenvalue but different eigenvectors is clear evidence of a degenerate matrix. Both these vectors should be expressible as linear combinations of the orthonormal base vectors in §3.4. Indeed they are $1·848248\,\mathbf{b_1}$ and $1·666364\,\mathbf{b_1} + 0·338170\,\mathbf{b_2}$ respectively. We might expect a similar behaviour with a non-degenerate matrix which has two or more eigenvalues close together. I may term such a 'near-degenerate matrix'. Appendix 2 gives an example.

This use of geometric series to accelerate convergence seem very obvious, yet I could not find an account of it in the books I consulted or on the internet. I therefore wrote to Prof. David Watkins at Washington State University, and he kindly replied that the method is equivalent to Aiken's 'delta-squared' method. This was first described by the New Zealander Alexander Aiken in

1927 as a general method for accelerating the convergence of any series that is geometric or almost geometric. David Watkins explains

> The extrapolation technique is equivalent to Aitken acceleration (Aitken's delta-squared process), found in Wilkinson's book (page 578) and covered in many numerical analysis texts. It is not usually presented in terms of geometric series, but it is nevertheless the same. I didn't make use of Aitken acceleration in either of my books.
>
> Aitken acceleration is useful whenever a sequence converges linearly, or geometrically as you call it. The QR algorithm with standard shifting strategies normally converges quadratically, so Aitken is not of use there. I think that is the reason Aitken has not become an important tool in the world of eigenvalue computations. It is only good for accelerating linearly convergent processes, and it is not competitive against quadratically convergent processes.

4.3 Shifting to improve convergence

Shifting the diagonal elements of the matrix is a way to manipulate the ratio of largest to next largest eigenvalues and so promote convergence. As shown at point 7 in §2.1, the offset β changes the ratio of eigenvalues to

$$\frac{\lambda_1 + \beta}{\lambda_2 + \beta} \tag{16}$$

and this is largest when $\beta = -\lambda_2$. Using our example 3×3 matrix $\mathbf{E}$, suppose that with some happy guess of starting vector it is becoming clear after a few iterations that λ_1 is near $6 \cdot 2$. Now λ_2 is unknown, but we do know that the average value of λ_2 and λ_3 is (Trace$-\lambda_1$)/2. The convergence rate towards λ_1 can be increased with $\beta = (4 - 6 \cdot 2)/2 = -1 \cdot 1$, the mean of the other eigenvalues. To confirm this numerically note the following alternative calculations. Taking $(1,\ 1,\ 1)$ as the starting vector $\mathbf{v}^{(0)}$ and no shift of diagonal elements, the first five estimates of largest eigenvalue are 5, $7 \cdot 2$, $5 \cdot 83$, $6 \cdot 4$ and $6 \cdot 12$. A measure of change between consecutive estimates $\mathbf{v}^{(m)}$ and $\mathbf{v}^{(m+1)}$ next is required, so I use δ defined as $\sqrt{(\sum[v_j^{(m+1)} - v_j^{(m)}]^2)}$ where $v_j^{(m)}$ is the j^{th} component of vector $\mathbf{v}^{(m)}$. δ is the length of the vector joining the points defined by $\mathbf{v}^{(m)}$ and $\mathbf{v}^{(m+1)}$ when regarded as position vectors. If the iterations continue with no shift, δ reduces to $7 \cdot 6 \times 10^{-7}$ at 21 iterations of the direct power method. If instead $\mathbf{E}$ is replaced by $\mathbf{E} - 1 \cdot 1\,\mathbf{I}$ after the first five iterations, δ has reduced to the same level after 9 further iterations, making 14 in all. This is not a great saving in calculational effort in this case, but still cuts it to 2/3. The eigenvalue arrived at is $7 \cdot 312665$ which needs adjusting downwards by $1 \cdot 1$. The other two eigenvalues, could they be found, are $-1 \cdot 87$ and $+1 \cdot 86$, straddling zero evenly, the best that could be achieved by choice of β.

Having found λ_1 a second eigenvalue-vector pair can be found by shifting by $\beta = \lambda_1$ to make λ_1 play no role in the iteration sequence. If λ_1 is positive and λ_2 is negative, as here, the shifted sequence will converge to λ_2. If λ_2 is also positive, shifting by λ_1 will give convergence on λ_3. Multiplying $(1,\ 1,\ 1)$ a few times by $\mathbf{E} - 6 \cdot 2\,\mathbf{I}$, the first four eigenvalue estimates are $-1 \cdot 2$, $-10 \cdot 4$, $-9 \cdot 75$, $-9 \cdot 5$ and, if allowed to continue with $\beta = 6 \cdot 2$, δ falls to 9×10^{-7} by iteration 27. It is possible to optimise β as above based on the estimate $\lambda_2 \approx -9 \cdot 2 + 6 \cdot 2 = -3 \cdot 0$. With β set to (Trace$-(-3 \cdot 0)$)/2 $= 3 \cdot 5$ after the first four iterations, in a further 15 iterations (19 in all) δ has reduced to 7×10^{-7}, another modest but worthwhile improvement in convergence rate. The other two eigenvalues have been shifted to $2 \cdot 71$ and $-2 \cdot 74$ giving the smallest absolute value of the next largest eigenvalue.

This suggests a two stage approach to finding an eigenvalue λ. First a few iterations on the starting vector are used to find a rough value, then a shift of matrix diagonal is introduced to enhance the rate of convergence. Since a poor choice of starting vector will give poor convergence at the start of the process, it is probably worth trying two or three vectors and picking the one which has the smallest change δ at each step, since small changes in the iterates signal closeness of the true value. Three starting choices might be $(1, 1, 1)$, $(1, -2, 1)$ and $(-1, 0, 1)$, which are mutually orthogonal. Further increase is the rate of convergence will probably be possible using the summed geometric series described in §3.2. We see that in the direct power method there is considerable scope for rapidly finding at least some of the eigenvector-value pairs.

In Appendix 1, §9, I give details of an attempt to find the eigen pairs of a challenging 6×6 matrix which has some eigenvalues close together.

5 Inverse power method with LU decomposition

Can the Power Method be adapted further to find the third and further eigenvalue-eigenvector pairs? The Inverse Power Method is a scheme which applies the iteration process not to $\mathbf{E}$ but to its inverse to take advantage of the fact that the eigenvalues of $\mathbf{E}^{-1}$are reciprocals $1/\lambda$:

$$\mathbf{E}\mathbf{p} = \lambda\mathbf{p} \quad \text{so} \quad \mathbf{p} = \lambda\mathbf{E}^{-1}\mathbf{p}\,, \quad \mathbf{E}^{-1}\mathbf{p} = \frac{1}{\lambda}\mathbf{p}\,.$$

Suppose one suspected an eigenvalue of $\mathbf{E}$ near β. Shifting the diagonal by β will make that eigenvalue shift to near zero and then its reciprocal will be very large. Convergence through multiplication by the inverse shifted matrix should therefore be very rapid. Finally the resulting eigenvalue must be transformed into an eigenvalue of $\mathbf{E}$ by reverse shift of its reciprocal.

Though this sounds a good scheme, it would be of little practical use if we actually had to invert $\mathbf{E}$. With large matrices this would be a heavy challenge subject to rounding errors. Instead the equation

$$\mathbf{E}^{-1}\,\mathbf{v}^{(m)} \;=\; \frac{1}{\lambda_{m+1}}\mathbf{v}^{(m+1)} \quad \text{is written} \qquad \mathbf{E}\left(\frac{\mathbf{v}^{(m+1)}}{\lambda_{m+1}}\right) \;=\; \mathbf{v}^{(m)} \tag{17}$$

and the simultaneous equations implied by the second form are solved for the components of $\mathbf{v}^{(m+1)}/\lambda_{m+1}$. Solution is made easier by factorising $\mathbf{E}$ using 'LU decomposition'.

5.1 LU decomposition

LU decomposition is well described in texts on linear algebra so I mention it only briefly. The idea is to factor the given matrix into the product of two triangular matrices, $\mathbf{L}$ and $\mathbf{U}$ where $\mathbf{L}$ has 1s on the diagonal and 0s above the diagonal, and $\mathbf{U}$ has 0s below the diagonal. The motivation is that the given equation $\mathbf{A}\mathbf{u} = \mathbf{v}$, say, can be replaced by two equations each of which is far easier to solve:

$$\mathbf{A}\mathbf{u} = \mathbf{v} \quad \rightarrow \quad \mathbf{L}\mathbf{U}\mathbf{u} = \mathbf{v} \quad \rightarrow \quad \mathbf{L}\mathbf{w} = \mathbf{v},\ \mathbf{U}\mathbf{u} = \mathbf{w}\,. \tag{18}$$

With $\mathbf{L}$ and $\mathbf{U}$ being strictly triangular, solving each of the sets of simultaneous equations on the right is simply by a sequence of substitutions from row to row, solving one equation in one variable at each row. $\mathbf{U}$ is produced by applying elementary row operations *except swapping rows*[4] to $\mathbf{A}$. Each such

[4] Not all matrices can be LU decomposed without some swapping of rows. The technique can be extended to these by multiplying $\mathbf{L}$ on the left by an elementary permutation matrix.

row operation is effected by multiplying by an invertible matrix $\mathbf{e}_j$. Suppose that $\mathbf{U} = \mathbf{e}_n \ldots \mathbf{e}_2\, \mathbf{e}_1\, \mathbf{A}$. Then $\mathbf{L} = \mathbf{e}_1^{-1}\, \mathbf{e}_2^{-1} \ldots \mathbf{e}_n^{-1}$.

I will now find an LU decomposition of our example matrix $\mathbf{E}$ (without a shift) and use it with the inverse power method to find the eigenvector with smallest magnitude. The answer should be $0 \cdot 7584554$. The matrix is

$$\mathbf{E} \;=\; \begin{pmatrix} -4 & -2 & 3 \\ 1 & 3 & 4 \\ -1 & 1 & 5 \end{pmatrix}.$$

There is not a unique decomposition – they differ depending on the order and nature of the elementary row operations. The scheme I have used has 7 steps:

$$R_1 \to \tfrac{-1}{4} R_1, \quad R_2 \to R_2 - R_1, \quad , R_2 \to \tfrac{2}{5} R_2, \quad R_3 \to R_3 + R_1,$$

$$R_3 \to \tfrac{2}{3} R_3, \quad R_3 \to R_3 - R_2, \quad R_3 \to \tfrac{15}{14} R_3\,.$$

The elementary matrices start

$$\mathbf{e}_1 \;=\; \begin{pmatrix} -\frac{1}{4} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{e}_2 \;=\; \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{e}_3 \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{2}{5} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{e}_4 \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

and these have inverses

$$\mathbf{e}_1^{-1} \;=\; \begin{pmatrix} -4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{e}_2^{-1} \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{e}_3^{-1} \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{5}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{e}_4^{-1} \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

The product of inverses gives $\mathbf{L}$ and the LU decomposition

$$\mathbf{E} \;=\; \begin{pmatrix} -4 & 0 & 0 \\ 1 & \frac{5}{2} & 0 \\ -1 & \frac{3}{2} & \frac{7}{5} \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2} & -\frac{3}{4} \\ 0 & 1 & \frac{19}{10} \\ 0 & 0 & 1 \end{pmatrix}. \qquad (19a)$$

It will be observed that $\mathbf{L}$ is made from the non-zero elements of the $\mathbf{e}_j^{-1}$ and this provides a short cut to writing $\mathbf{L}$ once $\mathbf{U}$ has been found. Note also that since none of the $\mathbf{e}_j$ has changed the sign of a row, the signs of the diagonal elements of $\mathbf{L}$, namely -4, $5/2$, $7/5$, give the signs of the three eigenvalues, one negative, two positive. Their product is -14, the determinant of $\mathbf{E}$.

It is possible to factor $\mathbf{L}$ itself into a lower triangular matrix $\mathbf{L}'$ with 1s on the diagonal and a diagonal matrix $\mathbf{d}$. Observe that

$$\begin{pmatrix} -\frac{1}{4} & 0 & 0 \\ 0 & \frac{2}{5} & 0 \\ 0 & 0 & \frac{5}{7} \end{pmatrix} \mathbf{L} \;=\; \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 \\ \frac{1}{4} & \frac{3}{5} & 1 \end{pmatrix} \;=\; \mathbf{L}'.$$

$$\text{so} \quad \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 \\ \frac{1}{4} & \frac{3}{5} & 1 \end{pmatrix} \begin{pmatrix} -4 & 0 & 0 \\ 0 & \frac{5}{2} & 0 \\ 0 & 0 & \frac{7}{5} \end{pmatrix} \mathbf{U} \;=\; \mathbf{E}\,.$$

This is called 'LDU decomposition' though I do not see that it offers any advantage over the two-product LU version. Despite a superficial resemblance to a similarity transformation, the above

product is not similar to $\mathbf{E}$ because $\mathbf{L'}^{-1} \neq \mathbf{U}$ and the eigenvalues -4, $5/2$ and $7/5$ are not the eigenvalues of $\mathbf{E}$. It does, however, point to an alternative LU factorisation as $\mathbf{L'}.(\mathbf{DU})$:

$$\begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 \\ \frac{1}{4} & \frac{3}{5} & 1 \end{pmatrix} \begin{pmatrix} -4 & -2 & 3 \\ 0 & \frac{5}{2} & \frac{19}{4} \\ 0 & 0 & \frac{7}{5} \end{pmatrix} = \mathbf{E}\,. \tag{19b}$$

Since the eigenvector to λ_1 of $\mathbf{E}$ was fairly close in direction to $(1, 1, 1)$, I will choose the starting vector for $1/\lambda_3$ to be orthogonal, namely $\mathbf{v}^{(0)} = (1, -2, 1)$. The first iteration runs as follows. Solving $\mathbf{Lw} = (1, -2, 1)$ gives $\mathbf{w} = (-\frac{1}{4}, -\frac{7}{10}, \frac{9}{7})$. Next, solving $\mathbf{Uv} = \mathbf{w}$ gives $\mathbf{v} = \frac{1}{7}(16, -22, 9)$. The normalised first iterate is therefore $\mathbf{v}^{(1)} = (1 \cdot 7778, -2 \cdot 4444, 1)$. Continuing this two-step process, the second and third iterations give $(1 \cdot 665, -2 \cdot 538, 1)$ and $(1 \cdot 706, -2 \cdot 540, 1)$ and the estimated (reciprocal) eigenvalue is $1 \cdot 326$. At this point we have at least two options:

Option 1 : Continue with the inverse power method for a few more iterations and fit a double geometric series to the sequence of eigenvalue estimates. This sequence runs

$$1 \cdot 2857143, \quad 1 \cdot 2539683, \quad 1 \cdot 3264014, \quad 1 \cdot 3153179, \quad 1 \cdot 3191361, \quad 1 \cdot 3182819$$

with differences δ

$$-0 \cdot 0317460, \quad 0 \cdot 0724332, \quad -0 \cdot 0110835, \quad 0 \cdot 0038182, \quad -0 \cdot 0008542\,.$$

Fitting a double geometric series to the last four values of δ gives first terms $\mathcal{C} = 0 \cdot 052984$, $\mathcal{D} = 0 \cdot 019449$ and common ratios $r = -0 \cdot 2567216$, $s = 0 \cdot 12950$. (Refer to §4.1 and Appendix 1.) The projected sum to infinity is $0 \cdot 0645030$. Adding this to the second λ iterate above gives $1 \cdot 31873 = 1/0 \cdot 758454$. This is the third eigenvalue with error of 1 in the 6th decimal place (should be 5). The ratio r points to $\lambda_2 \approx -2 \cdot 92$; the true value is $-2 \cdot 97$.

Option 2 : The agreement between $\mathbf{v}^{(2)}$ and $\mathbf{v}^{(3)}$ is already probably sufficient for us to take $1 \cdot 326$ as an estimate of the eigenvalue and consider a shift β given by (Trace$-1 \cdot 326)/2$ to speed convergence. Unfortunately this is not straightforward since we know neither $\mathbf{E}^{-1}$ not its trace[5]. Instead we can shift the diagonal of $\mathbf{E}$. The current estimate is $\lambda_3 \approx 1/1 \cdot 326 = 0 \cdot 754$. Subtract this from the diagonal of $\mathbf{E}$ and the shifted eigenvalue will be close to zero and its reciprocal large. To apply the inverse power method we now need the LU decomposition of

$$\begin{pmatrix} -4 \cdot 754 & -2 & 3 \\ 1 & 2 \cdot 246 & 4 \\ -1 & 1 & 4 \cdot 246 \end{pmatrix} = \begin{pmatrix} -4 \cdot 7540 & 0 & 0 \\ 1 & 1 \cdot 8253 & 0 \\ -1 & 1 \cdot 4207 & 0 \cdot 0104 \end{pmatrix} \begin{pmatrix} 1 & 0 \cdot 4207 & -0 \cdot 6310 \\ 0 & 1 & 2 \cdot 5371 \\ 0 & 0 & 1 \end{pmatrix}.$$

Use the last estimate of the eigenvector, $(1 \cdot 706, -2 \cdot 540, 1)$, and convergence is very fast. In two iterations the eigenvalue estimate is $224 \cdot 4467$ and in three it is $224 \cdot 4463$, giving λ_3 of $\mathbf{E}$ to be $0 \cdot 75845540876$ which is correct to 9 decimal places. The eigenvector is similarly precise. There is little value in using geometric series here as convergence through shifting is very rapid..

The effort in the inverse power method is largely in calculating the LU decomposition, though this need be done only once, and then using it with back or forwards substitution to solve for

[5] The trace of $\mathbf{E}^{-1}$ bears no simple relation to the traces of $\mathbf{L}$ and $\mathbf{U}$, and finding $\mathbf{E}^{-1}$ would require finding both $\mathbf{L}^{-1}$ and $\mathbf{U}^{-1}$.

the next estimate of the eigenvector. It can be used judiciously with the direct power method to improve convergence once the direct method has given a rough value for an eigenvalue, say λ_1. The procedure would be to take β equal to the λ_1 estimate, subtract it from the diagonal of $\mathbf{E}$, find the LU decomposition of the resulting matrix and operate with this on the best estimate so far of the eigenvector.

5.2 Rayleigh quotient iteration

In Option 2 of the previous subsection we applied a constant shift of $0 \cdot 754$ to $\mathbf{E}$ to make the reciprocal of the shifted λ_3 large. We might suspect that, as the iteration sequence converges towards an eigenvalue, the shift could be adjusted at each step to accelerate convergence. The natural choice of shift is the most recent estimate of λ. This is known as Rayleigh quotient iteration because the last estimate of λ is given by the Rayleigh quotient defined in §2, item 3 for symmetric matrices as the quotient of two scalar products, $\mathbf{p}^T\mathbf{E}\mathbf{p}\,/\,|\mathbf{p}|^2$.

To see what improvement it makes, here is the inverse power method applied to finding λ_3 as in the previous subsection, but modified by continual shifting. I start with zero shift and the starting vector used previously, $\mathbf{v}^{(0)} = (1, -2, 1)$. The first iteration gives reciprocal eigenvalue estimate 9/7 and hence $\lambda_3^{(1)} = 7/9 = 0 \cdot 7778$. So 7/9 is subtracted from the diagonal of $\mathbf{E}$; call this $\mathbf{E_1}$. The LU decomposition of this is now found and $\mathbf{E_1}^{-1}\mathbf{v}^{(1)}$ calculated. Its value is $(-84 \cdot 007,\ 125 \cdot 662,$ $-49 \cdot 4216\ )$ from which the estimate of the eigenvalue of this shifted matrix $\mathbf{E_1}$ is $-49 \cdot 4216 = -1/0 \cdot 020234$. Normalising the vector, $\mathbf{v}^{(2)} = (1 \cdot 6998,\ -2 \cdot 54265,\ 1\ )$. The current estimate of λ_3 is $0 \cdot 7778 - 0 \cdot 020234\ = 0 \cdot 75754$. So $\mathbf{E}$ is shifted by this amount to form $\mathbf{E_2}$ and the next cycle carried out. This gives a change in λ_3 of $1/1096 \cdot 8$ and the next iteration adds a further change of $-1/(1 \cdot 9156 \times 10^7)$. The total shift is now $0 \cdot 758455408744$, which is λ_3 correct to 12 decimal places. The eigenvector is similarly accurate at $(1 \cdot 69907005196,\ -2 \cdot 54247453929,\ 1)$.

As we expected, convergence has been remarkably fast, but it has required a new LU decomposition at every iteration to solve the inverse matrix multiplication. It will be a matter of judgement how to trade between convergence rate and the efforts of LU decomposition.

Rayleigh quotient iteration is essentially the above process of shifting the diagonal at each iteration, except that the shift λ_j is calculated by the Rayleigh quotient rather than simple multiplication by $\mathbf{E}$. The formula for the next iteration is

$$\lambda_j = \frac{\mathbf{v}_j^T\mathbf{E}\mathbf{v}_j}{\mathbf{v}_j^T\,\mathbf{v}_j}, \qquad \mathbf{v}_{j+1} \;=\; \frac{G}{|G|}, \qquad G = (\mathbf{E} \,-\, \lambda_j\mathbf{I})^{-1}\,\mathbf{v}_j \tag{20}$$

where $_j$ refers to the iteration index.

In Appendix 2, §10 I illustrate a combined direct-inverse power method with diagonal shifting and geometric series projection all being applied to a challenging 6×6 matrix which has some very close eigenvalues.

6 Matrix deflation and reduction

The direct and indirect versions of the power method essentially just determine the eigen pair with the largest absolute value. The largest eigenvalue has a magnetic effect on iteration schemes, pulling successive iterations towards itself even if the intent is to determine a different eigen pair. It is

necessary, therefore, to have some way for setting aside the largest pair as they are found so that the next largest eigenvalue/vector pair can be determined. Several ways of eliminating an eigenvalue from a given matrix have been proposed, and the process is known as 'matrix deflation'. I suppose whoever invented the term pictured the matrix like a punctured balloon or tyre losing air and collapsing. There are also algorithms to decreased the order of the matrix and leave an $(n-1)\times(n-1)$ matrix with the same eigenvalues as $\mathbf{E}$ except λ_1.

6.1 Hotelling deflation

The most simple scheme is probably that known as Hotelling deflation after Harold Hotelling, an economics professor late of Stanford University. It has the great advantage that the other eigenvalues and eigenvector are not changed. This is my own account.

Consider the given matrix $\mathbf{E}$ subject to a similarity transformation (§2, item 11, Eq 4) which converts it to the diagonal matrix $\mathbf{D}$. The transformation matrix $\mathbf{P}$ is constructed from the column eigenvectors $\mathbf{p}_j$ of $\mathbf{E}$. Now the diagonal entries of $\mathbf{D}$ are the eigenvalues and $\mathbf{D}$ can be split into a sum of matrices each of which involves only one eigenvalue. I illustrate it for a 3 by 3 matrix.

$$\mathbf{D} \;=\; \lambda_1\mathbf{m_1} + \lambda_2\mathbf{m_2} + \lambda_3\mathbf{m_3}\,, \tag{21}$$

$$\mathbf{m_1} \;=\; \begin{pmatrix}1&0&0\\0&0&0\\0&0&0\end{pmatrix}, \qquad \mathbf{m_2} \;=\; \begin{pmatrix}0&0&0\\0&1&0\\0&0&0\end{pmatrix}, \qquad \mathbf{m_3} \;=\; \begin{pmatrix}0&0&0\\0&0&0\\0&0&1\end{pmatrix}.$$

The essential step is to form $\mathbf{D} - \lambda_1\mathbf{m_1}$ which clearly has eigenvalues 0, λ_2, λ_3; that is, λ_1 has been replaced by zero. The inverse of Eq 9 is $\mathbf{E} = \mathbf{PDP}^{-1}$ and this carries through to the sum of one-eigenvalue terms:

$$\mathbf{E} - \lambda_1\mathbf{P}\mathbf{m_1}\mathbf{P}^{-1} \;=\; \lambda_2\mathbf{P}\mathbf{m_2}\mathbf{P}^{-1} + \lambda_3\mathbf{P}\mathbf{m_3}\mathbf{P}^{-1} \tag{22}$$

and this too has eigenvalues 0, λ_2, λ_3 while the eigenvectors $\mathbf{p}_2$ and $\mathbf{p}_3$ have not been changed. The matrix has been deflated.

True though the above is, it is not much help to the power method as it stands because only one eigenvector, $\mathbf{p}_1$ would have been found; the matrix $\mathbf{P}$ requires $\mathbf{p}_2$ and $\mathbf{p}_3$ as well. Fortunately there is a route to $\mathbf{P}\mathbf{m_1}\mathbf{P}^{-1}$ which does not require $\mathbf{p}_2$ and $\mathbf{p}_3$. We use the properties of the transposed matrix $\mathbf{E}$ and its eigenvectors briefly introduced at item 9 in §2 and at the end of the example in §3.1. It aids the analysis greatly if we change the normalisation of the eigenvectors. To avoid confusion I will retain $\mathbf{p}_j$ as the eigenvectors of $\mathbf{E}$ normalised with the final vector component set to 1, as so far in this article. I introduce $\mathbf{x}_j$ to denote the same eigenvectors but normalised to be unit vectors: $|\mathbf{x}_j| = 1$.

Let $\mathbf{y}_j$ be the eigenvectors of $\mathbf{E}^T$. Here is a proof that $\mathbf{x}_j$ and $\mathbf{y}_j$ are orthogonal. We have

$$\mathbf{E}\mathbf{x}_j = \lambda_j\mathbf{x}_j \quad \text{and} \quad \mathbf{E}^T\mathbf{y}_k = \mu_k\mathbf{y}_k\,.$$

$$\text{Hence} \qquad \mathbf{y}_k^T\mathbf{E}\mathbf{x}_j = \lambda_j\mathbf{y}_k^T\mathbf{x}_j \quad \text{and} \quad \mathbf{x}_j^T\mathbf{E}^T\mathbf{y}_k = \mu_k\mathbf{x}_j^T\mathbf{y}_k\,.$$

Use the reversing property of the matrix transpose operator to obtain

$$(\mathbf{x}_j^T\mathbf{E}^T\mathbf{y}_k)^T \;=\; \mathbf{y}_k^T\mathbf{E}\mathbf{x}_j \;=\; \mu_k\mathbf{y}_j^T\mathbf{x}_j$$

which is identical to the line above and means, with $j = k$, that $\mu_k = \lambda_k$. In words, a matrix and its transpose share the same eigenvalues. Where the eigenvalues are all different (no degeneracy),

$\lambda_k \neq \lambda_j$, $j \neq k$, implies that $\mathbf{y}_k^T\mathbf{x}_j = \mathbf{x}_j^T\mathbf{y}_k = 0$. This proves orthogonality. For $j = k$ it will be expedient to normalise $\mathbf{y}_j$ so that the dot product $\mathbf{y}_j^T\mathbf{x}_j = 1$. (This does not generally make $\mathbf{y}_j$ into a unit vector.) Now form the transformations matrix $\mathbf{X}$ from the columns $\mathbf{x}_j$, $j = 1,\ 3$. This is the equivalent of $\mathbf{P}$. Also form $\mathbf{Y}$ from the columns $\mathbf{y}_j$. The orthogonality of the vector pairs $\mathbf{x}_j$ and $\mathbf{y}_k$ leads to $\mathbf{X}$ and $\mathbf{Y}$ being related by $\mathbf{Y}^T\mathbf{X} = \mathbf{I}$, the identity matrix. Therefore

$$\mathbf{Y}^T \;=\; \mathbf{X}^{-1}\,. \tag{23}$$

Revisit Eq 19 with the renormalised eigenvectors.

$$\mathbf{E} - \lambda_1\mathbf{X}\mathbf{m_1}\mathbf{X}^{-1} \;=\; \lambda_2\mathbf{X}\mathbf{m_2}\mathbf{X}^{-1} + \lambda_3\mathbf{X}\mathbf{m_3}\mathbf{X}^{-1}$$

$$\mathbf{E} - \lambda_1\mathbf{X}\mathbf{m_1}\mathbf{Y}^T \;=\; \lambda_2\mathbf{X}\mathbf{m_2}\mathbf{Y}^T + \lambda_3\mathbf{X}\mathbf{m_3}\mathbf{Y}^T\,.$$

The final step is to see that $\mathbf{X}\mathbf{m_1}\mathbf{Y}^T = \mathbf{x}_1\mathbf{y}_1^T$, a 3×3 matrix. If we write $\mathbf{X}$ and $\mathbf{X}$ as block matrices,

$$\mathbf{X}\mathbf{m_1}\mathbf{Y}^T \;\equiv\; \left(\begin{array}{c|c|c} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{array}\right)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{y}_1^T \\ \hline \mathbf{y}_2^T \\ \hline \mathbf{y}_3^T \end{pmatrix} \;=\; \left(\begin{array}{c|c|c} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{array}\right)\begin{pmatrix} \mathbf{y}_1^T \\ \hline \mathbf{0} \\ \hline \mathbf{0} \end{pmatrix} \;=\; \mathbf{x}_1\mathbf{y}_1^T\,.$$

The matrix $\mathbf{m_1}$ ensures that neither $\mathbf{x}_2$ nor $\mathbf{x}_3$ features. The consequence for the power method is that once λ_1 and $\mathbf{x}_1$ ($\equiv \mathbf{p}_1$) have been found, $\mathbf{E}$ can be deflated by removal of λ_1 provided we invest in also determining $\mathbf{y}_1$ for $\mathbf{E}^T$. For a symmetric matrix $\mathbf{E}^T = \mathbf{E}$ and there is no extra work involved. For a general matrix the inverse power method should allow $\mathbf{y}_1$ to be found quickly since we can subtract, say, $0\cdot 98\lambda_1$ from the diagonal of $\mathbf{E}$ to make the reciprocal shifted eigenvalue very large.

Here is an illustration for the matrix $\mathbf{E}$ of Eq 7, previously examined at length. We pick up from §3 where λ_1 and $\mathbf{p}_1$ were found. $\mathbf{p}_1$ is renormalised to $\mathbf{x}_1 = (0\cdot 0292438,\ 0\cdot 7828865,\ 0\cdot 6214769)$. To find $\mathbf{y}_1$ subtract about $0\cdot 98\lambda_1 = 6\cdot 1$ from the diagonal and carry out LU decomposition on $\mathbf{E}^T$. I have used $(1, -2, 1)$ as the starting vector. After eight iterations of the inverse power method the shifted reciprocal eigenvalue has changed by only $2\cdot 6\times 10^{-9}$ and settled at $8\cdot 87594597336$. Then $6\cdot 1 + 1/8\cdot 87594597336$ is λ_1 correct to 11 decimal places. The corresponding eigenvector is equally precise; its first few digits are

$$\begin{pmatrix} -0\cdot 0635643 \\ 0\cdot 3508392 \\ 1 \end{pmatrix} \quad \text{so} \quad \mathbf{y} = \begin{pmatrix} -0\cdot 0710783 \\ 0\cdot 3923124 \\ 1\cdot 1182114 \end{pmatrix}, \quad \mathbf{y}_1^T\mathbf{x}_1 = 1.$$

$$\mathbf{x}\mathbf{y}^T \;=\; \begin{pmatrix} -0\cdot 0020786 & 0\cdot 0114727 & 0\cdot 0327008 \\ -0\cdot 0556463 & 0\cdot 307136 & 0\cdot 8754326 \\ -0\cdot 0441735 & 0\cdot 2438131 & 0\cdot 6949425 \end{pmatrix},$$

$$\mathbf{E} - 6\cdot 2126640\,\mathbf{x}\mathbf{y}^T \;=\; \begin{pmatrix} -3\cdot 987086 & -2\cdot 0712762 & 2\cdot 796841 \\ 1\cdot 3457115 & 1\cdot 0918666 & -1\cdot 4387686 \\ -0\cdot 7255647 & -0\cdot 5147288 & 0\cdot 6825557 \end{pmatrix}. \tag{24}$$

This is the deflated matrix with eigenvalues 0, λ_2, λ_3 and trace $4 - \lambda_1 = -2\cdot 212664$. It is now a singular matrix without an inverse so it cannot be used as it stands in the inverse power method. It is, however, all right in the direct power method and can be used to find a precise value for λ_2 and particularly the eigenvector $\mathbf{p}_2$. (In truth λ_2 is already known to high precision from the Trace-

$\lambda_1 - \lambda_3$, all now determined.) In §4.1 Table 4 the geometric series projected value of λ_2 (from the ratio r) was $2\cdot9711182$ and $\mathbf{p}_2$ was $(6\cdot273,\ -1\cdot726,\ 1)$. When this is used as starting vector in the direct power method with the deflated matrix Eq 21, convergence is rapid, and can be made even more rapid by shifting by about $(-2\cdot212664+2\cdot9711182)/2 = \lambda_3/2 \approx 0\cdot4$ and/or using the geometric series projection.

I have quoted several iterations of λ and $\mathbf{p}$ to high precision to emphasise that such precision is necessary where several eigen pairs are to be found, because rounding errors in the first few pairs accumulate and propagate to later pairs. A constant check is that the sum of eigenvalues should equal the trace of the original matrix.

There is a more general deflation process named after Wielandt. It has the advantage that is does not require the eigenvector $\mathbf{y}$ of $\mathbf{E}^T$ to be found, but suffers because the eigenvectors $\mathbf{x} \equiv \mathbf{p}$ are not those of the original matrix. It is therefore a way for finding eigenvalues only. The eigenvectors would have to be found in a separate operation by solving $(\mathbf{E} - \lambda\mathbf{I})\mathbf{p} = \mathbf{0}$.

6.2 Matrix order reduction

I found the scheme described below in an old book on numerical methods by Louis G. Kelly, Addison-Wesley 1967, page 134. It is a way of simultaneously eliminating one chosen eigen pair and reducing the matrix order by one. It is simple to use, but has the limitation of Wielandt deflation that the eigenvectors are not preserved. I will describe the method first through an example then give a proof.

Let us remove λ_1, $\mathbf{p}_1$ from matrix $\mathbf{E}$ of Eq 7. The matrix is partitioned to separate the right-most column and the bottom row. The eigenvector, normalised so its last component is 1, is also partitioned:

$$\mathbf{E} = \left(\begin{array}{c|c} B & r \\ \hline s & c \end{array}\right) = \left(\begin{array}{cc|c} -4 & -2 & 3 \\ 1 & 3 & 4 \\ \hline -1 & 1 & 5 \end{array}\right), \qquad \mathbf{p}_1 = \begin{pmatrix} w_1 \\ \hline 1 \end{pmatrix} = \begin{pmatrix} 0\cdot047055 \\ 1\cdot259719 \\ \hline 1 \end{pmatrix}.$$

The reduced matrix is

$$\mathbf{E_1} = \mathbf{B} - \mathbf{w}_1\mathbf{s} = \begin{pmatrix} -4 & -2 \\ 1 & 3 \end{pmatrix} - \begin{pmatrix} 0\cdot047055 \\ 1\cdot259719 \end{pmatrix}\begin{pmatrix} -1 & 1 \end{pmatrix} = \begin{pmatrix} -3\cdot952945 & -2\cdot047055 \\ 2\cdot259719 & 1\cdot740281 \end{pmatrix}. \qquad (25)$$

The eigenvalues of $\mathbf{E_1}$ are λ_2 and λ_3 of $\mathbf{E}$. Its eigenvectors are respectively $(-2\cdot0849491, 1)$ and $(-0\cdot4344898, 1)$.

Whoever thought up this ingenious device probably had in mind to effect a partial diagonalisation of $\mathbf{E}$ so that the eigenvalue of interest is isolated on the diagonal by being alone in a row or column of zeros. Just as a full diagonalisation is effected by a similarity transformation using a matrix whose columns are the n eigenvectors of the n eigenvalues, so partial diagonalisation for one eigen pair may be achieved by a similarity transformation in which the transformation matrix uses the one known eigenvector. The partitioning of $\mathbf{E}$ and $\mathbf{p}_1$ as above anticipate the structure of the partially diagonalised matrix $\mathbf{E_1}$. The transformation matrix, $\mathbf{T}$ may originally have been an intuitive guess:

$$\mathbf{T} = \left(\begin{array}{c|c} I & w_1 \\ \hline 0 & 1 \end{array}\right), \qquad \mathbf{T}^{-1} = \left(\begin{array}{c|c} I & -w_1 \\ \hline 0 & 1 \end{array}\right).$$

Here I denotes a block whose elements are those of the identify matrix of order $n-1$. The similarity transformation is

$$\mathbf{T}^{-1}\mathbf{E}\,\mathbf{T} \;=\; \left(\begin{array}{c|c} B - w_1 s & Bw_1 + r - w_1(w_1 s + c) \\ \hline s & w_1 s + c \end{array}\right).$$

From this we pick out elements of the eigenvalue equation

$$\mathbf{E}\,\mathbf{p}_1 \;=\; \lambda_1 \mathbf{p}_1 \quad \text{equivalent to} \quad \left(\begin{array}{c|c} B & r \\ \hline s & c \end{array}\right)\begin{pmatrix} w_1 \\ 1 \end{pmatrix} \;=\; \begin{pmatrix} Bw_1 + r \\ \hline w_1 s + c \end{pmatrix} \;=\; \begin{pmatrix} \lambda_1 w_1 \\ \hline \lambda_1 \end{pmatrix}.$$

The upper right block in $\mathbf{T}^{-1}\mathbf{E}\,\mathbf{T}$ is zero and the partially diagonalised matrix is

$$\mathbf{H} \;=\; \left(\begin{array}{c|c} B - w_1 s & 0 \\ \hline s & \lambda_1 \end{array}\right). \tag{26}$$

Being a similarity transformation, this has exactly the same eigenvalues as the original matrix though of course the eigenvectors are different because the transformation is geometrically equivalent to rotation and stretching of the co-ordinate axes.

What about the eigenvectors of $\mathbf{E}_1$, which would be found if, say, the Power Method were applied to it? I find that it is possible to wind the deflation and matrix reduction process backwards and recover the eigenvector $\mathbf{p}_2$ once λ_2 has been found from $\mathbf{E}_1$. The steps are:

1. Find the eigenvalue λ_2 of $\mathbf{E}_1$ and its $n-1$ row eigenvector which I will call $\mathbf{q}_2$. The components of this satisfy the same $n-1$ simultaneous equations as they would in the $n \times n$ matrix $\mathbf{H}$ since the coefficient of the last column in $\mathbf{H}$ is 0.

2. Let the components of $\mathbf{q}_2$ be multiplied by a constant β. Take the dot product of the bottom row of $\mathbf{H} - \lambda_2\mathbf{I}$ with $\beta\mathbf{q}_2$ and solve for the β that makes this dot product zero. The eigenvector of $\mathbf{H}$ corresponding to $\mathbf{q}_2$ is the block matrix $(\beta\mathbf{q}_2, 1)^T$.

3. Form an $n \times n$ matrix in which this column eigenvector is the first column and the other columns are all 0 except for their last rows which are 1. Call this matrix $\mathbf{G}$.

4. Apply the reverse similarity transformation to $\mathbf{G}$. The first column of $\mathbf{T}\,\mathbf{G}\,\mathbf{T}^{-1}$ will be the required eigenvector $\mathbf{p}_2$ of $\mathbf{E}$. The second column will recover the eigenvector $\mathbf{p}_1$.

This is how it works for $\mathbf{E}_1$ of Eq 23. $\lambda_2 = -2 \cdot 971119$ and $\mathbf{q}_2 = (-2 \cdot 084949, 1)$ have been found. Refer to the last row of $\mathbf{H} - \lambda_2\mathbf{I_3}$ and solve the dot product

$$-(-2 \cdot 084949)\,\beta + \beta + [6 \cdot 212664 - (-2 \cdot 971119)] \;=\; 0\,.$$

The solution is $\beta = -2 \cdot 976964$ so the eigenvector in $\mathbf{H}$ is $(6 \cdot 206819, -2 \cdot 976964, 1)$. The matrix $\mathbf{G}$ is therefore

$$\begin{pmatrix} 6 \cdot 206819 & 0 & 0 \\ -2 \cdot 976964 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \qquad \text{and} \qquad \mathbf{T}\,\mathbf{G}\,\mathbf{T}^{-1} \;=\; \begin{pmatrix} 6 \cdot 253875 & 0 \cdot 047055 & -0 \cdot 306500 \\ -1 \cdot 717245 & 1 \cdot 259719 & -0 \cdot 246368 \\ 1 & 1 & -0 \cdot 306775 \end{pmatrix}.$$

The first column is $\mathbf{p}_2$, the second is $\mathbf{p}_1$ as given at Eq 8. I have not seen this recovery of the eigenvector describes in the literature, but again I assume it to be well known.

Sections §4, 5 and 6 plus Appendix 1 have presented several tools that can be tried to determine the real eigen pairs of a given matrix. It seems likely that no one approach will solve all eigenvalues of all matrices so some informed thinking may be needed to shape the approach in each case. As a test case and example of an overall strategy, Appendix 2 §10 sets out my attempt to solve a 5×5 non-symmetric matrix with real, fairly well spaced eigenvalues. This should not prove too difficult. A much more challenging matrix is tackled in Appendix 3, §11.

7 Jacobi's method for symmetric matrices

In 1846 the mathematical prodigy Carl Jacobi published a method equivalent to finding the eigenvalues and eigenvectors of a symmetric matrix by rotating the axes to reduce the off-diagonal elements to zero. This is achieved by a similarity transformation so the diagonalised matrix has the eigenvalues of the original one down its diagonal. The method was picked up again by Wallace Givens and others in the 1950s.

To understand the method, first consider a 2×2 symmetric real matrix $\mathbf{A}$ as describing some symmetric physical quantity. Item 20 in §2 explained that diagonalising $\mathbf{A}$ is equivalent to rotating it to align its principal axes with the co-ordinate frame being used to measure the elements of $\mathbf{A}$. Rotation is effected by multiplication by a matrix of the form

$$\mathbf{P} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}. \tag{27}$$

Since we want a similarity transformation to present the eigenvalues, the rotation $\mathbf{P}$ and its inverse are used as a pair. Writing c for $\cos\theta$, s for $\sin\theta$, the calculation runs

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} \equiv \begin{pmatrix} c & -s \\ s & c \end{pmatrix}\begin{pmatrix} f & g \\ g & h \end{pmatrix}\begin{pmatrix} c & s \\ -s & c \end{pmatrix}$$

$$= \begin{pmatrix} -g\sin 2\theta - \frac{1}{2}(f-h)\cos 2\theta - \frac{1}{2}(f+h) & g\cos 2\theta + \frac{1}{2}(f-h)\sin 2\theta \\ g\cos 2\theta + \frac{1}{2}(f-h)\sin 2\theta & g\sin 2\theta - \frac{1}{2}(f-h)\cos 2\theta + \frac{1}{2}(f+h) \end{pmatrix}.$$

This is symmetric. The off-diagonal elements can be made zero by suitable choice of θ, namely

$$\tan 2\theta = \frac{2g}{h-f}. \tag{28}$$

From this c and s can be found by taking nested square roots:

$$c = \pm\sqrt{\tfrac{1}{2}(1\pm R)}, \qquad s = \pm\sqrt{\tfrac{1}{2}(1\mp R)}, \qquad R^2 = \frac{1}{T^2+1}, \quad , \quad T = \frac{2g}{h-f}.$$

Note that R^2 is always positive so the root is a real number. It is necessary to choose the outer and inner signs, − or +, according to the signs of g and $h-f$. I find that this choice works:

g	$h-f$	c outer	c inner	s outer	s inner
>0	>0	+	+	+	−
<0	<0	+	−	−	+
>0	<0	+	−	+	+
<0	>0	+	+	−	−

This concept is now extended to an $n\times n$ symmetric matrix, $n>2$. This is pictured as a symmetric quantity in n dimensions being rotated about one axis at a time. If rotation involves the j^{th} row and the k^{th} column, $\mathbf{P}$ is the identity matrix with a 2×2 rotation inserted:

$$\begin{pmatrix} 1 & 0 & . & 0 & . & 0 & . & 0 & . & 0 & 0 \\ 0 & 1 & . & 0 & . & 0 & . & 0 & . & 0 & 0 \\ . & . & . & . & . & 0 & . & . & . & . & . \\ 0 & 0 & . & a_{jj} & . & 0 & . & a_{jk} & . & 0 & 0 \\ . & . & . & . & . & 1 & . & . & . & . & . \\ 0 & 0 & . & a_{kj} & . & 0 & . & a_{kk} & . & 0 & 0 \\ . & . & . & . & . & 0 & . & . & . & . & . \\ 0 & 0 & . & 0 & . & 0 & . & 0 & . & 1 & 0 \\ 0 & 0 & . & 0 & . & 0 & . & 0 & . & 0 & 1 \end{pmatrix}.$$

We identify a_{jj} with f, $a_{kj} = a_{jk}$ with g and a_{kk} with h. $\mathbf{P}^{-1}$ is obtained from $\mathbf{P}$ simply by reversing the signs of the two s. When $\mathbf{P}^{-1}\mathbf{AP}$ is formed, elements $a_{kj} = a_{jk} \to 0$.

The algorithm is iterative. At each iteration $a_{kj} = a_{jk}$ is chosen to be the off-diagonal element with largest absolute value. At the end of that iteration these elements are both 0, but in the next iteration they will in general again become non-zero. However, by stages the whole matrix is nudged towards becoming diagonal. As an illustration, I wrote a computer program and applied the method to the 4×4 symmetric matrix $\mathbf{B}$ from §3.2. The convergence can be tracked by watching the decrease in off-diagonal elements towards zeros, either by noting the largest absolute element used for choosing the rotation matrix, or by the absolute sum of off-diagonals. Figure 4 plots both on a logarithmic scale. (The sum in red points is of the upper off-diagonals only.) Convergence is slow to start, but accelerates. The red curve has a monotonic decrease. The non-linear graph shows that it converges faster than a geometric series. I set the finish criterion to be that the largest off-diagonal element is $< 10^{-8}$ and this was achieved at iteration 17. All eigenvalues were then correct to at least 14 places of decimal. There is also advantage in calculating the eigenvalues of the 2×2 submatrices down the diagonal, as was done with the QR method in Example 3 of §7.2 and in §7.3. For matrix $\mathbf{B}$ the three pairs of roots for the three 2×2 submatrices have stabilised by iteration 13 to $\{-1 \cdot 96422818, -7 \cdot 10569674\}$, $\{1 \cdot 02755183, -1 \cdot 96422818\}$, $\{7 \cdot 04237309, 1 \cdot 02755183\}$ respectively. These are the eigenvalues correct to 8 decimal places. The eigenvalues of these submatrices therefore anticipate the eigenvalues of the whole matrix.

Jacobi's method has the benefit of producing all eigenvalues simultaneously to about the same accuracy – unlike the Power and QR methods, it does not seem to favour one eigenvalue. A further advantage is that all the eigenvectors are obtained simultaneously from the product of the individual rotation matrices $\mathbf{P_0}\,\mathbf{P_1}\,\mathbf{P_2}\,....\mathbf{P_N}$; each column is the eigenvector of the eigenvalue in the corresponding column of the diagonalised matrix. The method clearly has much to recommend it, the main shortcoming being that in the form above it applies only to symmetric matrices. No doubt it has been extended to non-symmetric ones. One thought of my own was to try to convert the given matrix to upper triangular form by nested elementary rotations. Suppose the proto-2×2 matrix is

$$\begin{pmatrix} f & g \\ d & h \end{pmatrix}.$$

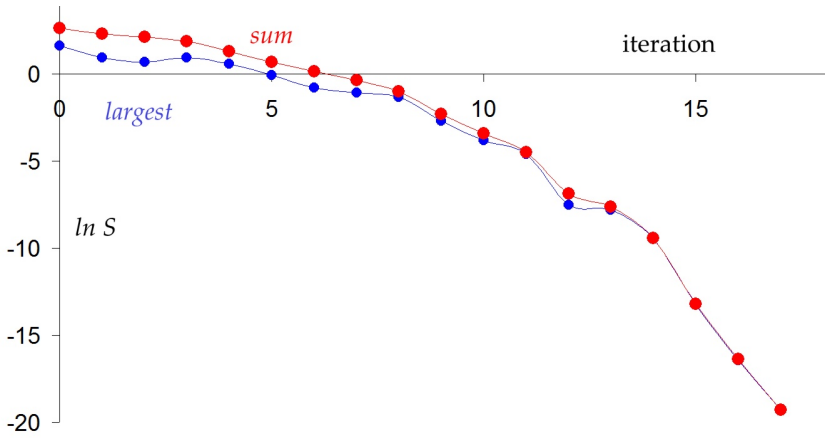

Figure 3: Convergence of the Jacobi method for matrix $\mathbf{B}$. Red: $\log_e$(sum of upper off-diagonal elements). Blue: $\log_e$(largest upper off-diagonal element).

I find that d can be set to zero by rotation through angle θ if

$$c \;=\; \frac{(d+g)s^2 - d}{(f-h)s} \;=\; 0\,, \quad c = \cos\theta, \quad s = \sin\theta.$$

This can be solved as a quadratic in s^2 to give

$$s^2 \;=\; \frac{2d(d+g) + (f-h)^2 \pm (f-h)R}{2D}\,, \qquad c^2 \;=\; \frac{2g(d+g) + (f-h)^2 \mp (f-h)R}{2D}\,,$$

$$R \;=\; \sqrt{(f-h)^2 + 4dg}\,, \qquad D = (d+g)^2 + (f-h)^2\,. \tag{29}$$

The result, however, is real only provided $(f-h)^2 + 4dg \geq 0$, otherwise s and c are complex. Shifting the diagonal does not overcome this problem because $f-h$ remains the same. I have written a computer program to effect this variant of Jacobi's method. To deal in some limited way with iterations in which R would be complex, if the largest absolute element in the lower off-diagonals gives R^2 negative, the second largest element is substituted for the first. If that too gives R^2 negative, the third largest is used, and if that too gives R^2 negative, the program terminates.

The program runs to correct completion with the 3×3 matrix $\mathbf{E}$ of §3.1 using at each iteration only the largest off-diagonal element. In this matrix the lower off diagonals are fairly small compared with the diagonal elements, and this favours R^2 being positive. At iteration 13 the sum of lower off-diagonal elements was $1\cdot4\times10^{-11}$ and all three eigenvalues were to at least 10 decimal places. In addition the eigenvector of the eigenvalue at position (1, 1) was correctly given. This bonus is granted because the upper triangular matrix in locally diagonalised in its first column.

The method failed to converge with both test matrices in Appendices 2 and 3. I examined several 6×6 matrices which were diagonally dominated. With most the algorithm also failed to converge. It seems that unless R^2 is positive for the largest element at every iteration, the lower triangle never reduces towards zero, but instead the sum of lower off-diagonal elements can increase or decrease in an apparently random way. One matrix for which it did converge, though only after over 60 iterations, was

$$\begin{pmatrix} 15 & 1 & 2 & -3 & 3 & -1 \\ 2 & 8 & 3 & -1 & -1 & 2 \\ -1 & 0 & -11 & 3 & -2 & -3 \\ -3 & -1 & 1 & 7 & 2 & 1 \\ 2 & 2 & 1 & -3 & -8 & 3 \\ -1 & 0 & -3 & 3 & -2 & 1 \end{pmatrix}.$$

Note that all diagonal elements are relatively large except in the last row. At iteration 69 the largest off-diagonal element was $9\cdot5\times10^{-9}$ and the eigenvalues given as

$$\begin{gathered} -10\cdot75526106 \\ -8\cdot35418432 \\ 0\cdot726374653 \\ 6\cdot49865301 \\ 7\cdot33031404 \\ 16\cdot5541036824 \end{gathered}$$

in that order down the diagonal. The first eigenvector is also given, but not the others. At every iteration R^2 was positive for the largest element; this may well be a criterion for convergence. Clearly, my extension of Jacobi's method is not of wide applicability and does not converge particularly quickly. However R^2 becoming negative does tell you that the method is not applicable.

8 Eigenvalues by QR-Schur decomposition

We now turn to another and more widely applicable iterative algorithm which also solves for all N eigenvalues at the same time. The QR algorithm was developed independently by John Francis in England and Vera Kublanovskaya in Russia in 1959 to 1961, though only Francis implemented it on a computer. The name Schur is also associated with QR decomposition. Issai Schur was of Russian Jewish descent, but studied under Frobenius in Berlin then worked all his adult life in Germany until forced to flee in 1939. He died in Tel Aviv in 1941.

8.1 A rationale for the basic QR algorithm

The first paper by John Francis[6] describes some of his thinking in developing this algorithm. He was building on an iteration scheme developed a year on so earlier by Heinz Rutishauser involving LU decomposition, the so-called LR transformation algorithm. Here is my own invented rationale of how we might imagine the algorithm coming about.

The aim is to develop an algorithm which will find all eigenvalues of a matrix $\mathbf{A}$ simultaneously. (The eigenvectors can be ignored for the time being.) We know that in a diagonal matrix all eigenvalues appear with equal status along the main diagonal, so it would be attractive to devise a scheme which will diagonalise our given matrix $\mathbf{A}$. We need a similarity transformation involving a square matrix $\mathbf{Q}$ such that $\mathbf{Q}^{-1}\mathbf{A}\,\mathbf{Q}$ is diagonal and has the same eigenvalues as $\mathbf{A}$. In fact we do not need a diagonal matrix – that is too demanding – since an upper or lower triangular matrix also has its eigenvalues down the diagonal (§2 item 13). So the concept is, by iteration, to effect a similarity transformation $\mathbf{A} \equiv \mathbf{A_0} \to \mathbf{A}_\infty$ where $\mathbf{A}_\infty$ is upper triangular. In fact there is a theorem proved by Schur that any square matrix is similar to an upper triangular matrix, the transformation matrix being orthogonal (or unitary in the complex case).

Suppose that the overall similarity transformation is compounded from a sequence of subsimilarity transformations of the form $\mathbf{A_{k+1}} = \mathbf{Q_k}^{-1}\mathbf{A_k}\,\mathbf{Q_k}$. As the iteration process draws to its conclusion, $\mathbf{A_k} \approx \mathbf{A_{k+1}} \approx \mathbf{A}_\infty$. That implies that $\mathbf{Q_k} \to \mathbf{I}$, the identity matrix and that $\mathbf{A_k}$ is almost triangular. How might an iteration be engineered? Suppose we factorise $\mathbf{A_{k+1}}$ as follows:

$$\mathbf{A_{k+1}} \;\equiv\; \mathbf{Q_k}^{-1}\mathbf{A_k}\,\mathbf{Q_k} \;=\; \mathbf{R_k}\,\mathbf{Q_k} \quad \text{where} \quad \mathbf{R_k} = \mathbf{Q_k}^{-1}\mathbf{A_k} \quad \text{so} \quad \mathbf{A_k} = \mathbf{Q_k}\mathbf{R_k}\,.$$

The iteration would then involve the two steps of factorising $\mathbf{A_k}$ into $\mathbf{Q_k}\mathbf{R_k}$ and then reverse multiplying $\mathbf{R_k}\mathbf{Q_k} = \mathbf{A_{k+1}}$. $\mathbf{Q_k}$ will be made orthogonal (or unitary) and $\mathbf{R_k}$ made upper triangular, consistent with the limiting behaviour of $\mathbf{A_k}$.

Can a scheme along these lines actually converge? To answer this let us examine the result of multiplying an almost triangular matrix by an almost identity matrix. I will illustrate this for 3×3 matrices. To fit with conventional notation let

$$\mathbf{Q_k} \;=\; \begin{pmatrix} 1+\varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & 1+\varepsilon_{22} & \varepsilon_{32} \\ \varepsilon_{31} & \varepsilon_{32} & 1+\varepsilon_{31} \end{pmatrix}, \qquad \mathbf{R_k} \;=\; \begin{pmatrix} \lambda_1+\alpha_1 & b_{12} & b_{13} \\ 0 & \lambda_2+\alpha_2 & b_{23} \\ 0 & 0 & \lambda_3+\alpha_3 \end{pmatrix}$$

where the ε_{jk} and α_j are small deviations from zero. The b_{jk} may be more substantial. We need both the left and right products:

$$\mathbf{Q_k}\mathbf{R_k} \;=$$

[6] 'The QR transformation, a unitary analogue to the LR transformation - Part 1' Computer Journal, Vol 4, 265-271, 1961.

$$\begin{pmatrix} \lambda_1(1+\varepsilon_{11})+\alpha_1+... & b_{12}(1+\varepsilon_{11})+\lambda_2\varepsilon_{12}+... & b_{13}(1+\varepsilon_{11})+\lambda_3\varepsilon_{13}+b_{23}\varepsilon_{12}+\ ... \\ (\lambda_1+\alpha_1)\varepsilon_{21} & \lambda_2(1+\varepsilon_{22})+b_{12}\varepsilon_{21}+\alpha_2+\ ... & b_{23}(1+\varepsilon_{22})+\lambda_3\varepsilon_{23}+b_{13}\varepsilon_{21}+\ ... \\ (\lambda_1+\alpha_1)\varepsilon_{31} & b_{12}\varepsilon_{31}+\lambda_2\varepsilon_{32}+\ ... & \lambda_3(1+\varepsilon_{33})+b_{23}\varepsilon_{32}+b_{13}\varepsilon_{31}+\alpha_3+\ ... \end{pmatrix}.$$

$$\mathbf{R_k Q_k} \ =$$

$$\begin{pmatrix} \lambda_1(1+\varepsilon_{11})+b_{12}\varepsilon_{21}+b_{13}\varepsilon_{31}+\alpha_1... & b_{12}(1+\varepsilon_{22})+b_{13}\varepsilon_{32}+\lambda_1\varepsilon_{12}... & b_{13}(1+\varepsilon_{33})+\lambda_1\varepsilon_{13}+b_{12}\varepsilon_{23}... \\ \lambda_2\varepsilon_{21}+b_{23}\varepsilon_{31}+\ ... & \lambda_2(1+\varepsilon_{22})+b_{23}\varepsilon_{32}+\alpha_2+\ ... & b_{23}(1+\varepsilon_{33})+\lambda_2\varepsilon_{23}+\ ... \\ (\lambda_3+\alpha_3)\varepsilon_{31} & (\lambda_3+\alpha_3)\varepsilon_{32} & \lambda_3(1+\varepsilon_{33})+\alpha_3+\ ... \end{pmatrix}.$$

The dots ... mean that a term made of a product of two small quantities has been omitted. Observe the following differences between these two products. In the first an eigenvalue appears in only one column, but in the second it appears in only one row. As a result in position (3,1) λ_1 has been replaced by λ_3. Similarly in position (2,1) λ_1 has been replaced by λ_2 and in position (3,2) λ_2 has been replaced by λ_3. There is the complication that at (2,1) the significant quantity b_{23} appears in **RQ**, but it is multiplied by ε_{31} which diminishes at each iteration. If $|\lambda_1| > |\lambda_2| > |\lambda_3|$, **RQ** is closer to being triangular than **QR** because all elements below the diagonal are smaller in magnitude. If **QR** were replaced by **RQ** at each iteration, the elements in positions (2,1), (3,1) and (3,2) would converge to zero at the rates of λ_2/λ_1, λ_3/λ_1 and λ_3/λ_2 per iteration respectively.

A part of the concept which is still missing is how to determine the near-identity matrix $\mathbf{Q_k}$ and the triangular matrix $\mathbf{R_k}$ during an early stage of the iteration. Intuition may have led the inventor of this algorithm to consider that $\mathbf{Q_k}$ should be an orthogonal matrix – one whose rows and columns are all unit vectors and whose dot product with other rows or columns is zero. It is related to the identity matrix by rotations and reflections and is a unit matrix in the sense that its determinant is either +1 or −1. The QR algorithm is a way of factorising the given matrix **A** into the product **QR** were **Q** is an orthogonal matrix and **R** is upper triangular. The method is well described in textbooks and on the internet so I will only outline it. We regard the columns of **A** as vectors spanning an N-dimensional space. The Gram-Schmidt algorithm is used to find an orthonormal set (perpendicular unit vectors) which span the same space. The algorithm takes the vector $\mathbf{a_1}$ in column 1 as a starting vector, which we label $\mathbf{u_1}$. The unit vector along this axis is $\mathbf{e_1} = \mathbf{u_1}/|u_1|$ and this is the first column of **Q**. Moving to the second column's vector $\mathbf{a_2}$, this will have a component perpendicular to $\mathbf{e_1}$ and a component normal to it. Calling the normal component $\mathbf{u_2}$,

$$\mathbf{u_2} \ = \ \mathbf{a_2} - (\mathbf{a_2}\cdot\mathbf{e_1})\,\mathbf{e_1} \quad \text{and} \quad \mathbf{e_2} = \frac{\mathbf{u_2}}{|u_2|}\,.$$

$\mathbf{e_2}$ becomes the second column of **Q**. The third column is a unit vector normal to both $\mathbf{e_1}$ and $\mathbf{e_2}$, made by subtracting the components of $\mathbf{a_3}$ which are parallel to $\mathbf{e_1}$ and $\mathbf{e_2}$. For the k^{th} column

$$\mathbf{u_k} \ = \ \mathbf{a_k} - (\mathbf{a_k}\cdot\mathbf{e_1})\,\mathbf{e_1} \ - \ (\mathbf{a_k}\cdot\mathbf{e_2})\,\mathbf{e_2} \ - \ ..... \ - (\mathbf{a_k}\cdot\mathbf{e_{k-1}})\,\mathbf{e_{k-1}}\,. \tag{30}$$

$$\text{Thus} \quad \mathbf{A} = [\mathbf{a_1} \mid \mathbf{a_2} \mid .... \mid \mathbf{a_N}] \qquad \text{maps to} \qquad \mathbf{Q} = [\mathbf{e_1} \mid \mathbf{e_2} \mid .... \mid \mathbf{e_N}]\,.$$

The upper elements of **R** are the projected lengths of the $\mathbf{a_j}$ on the N unit vectors $\mathbf{e_k}$.

$$\mathbf{R} \ = \ \begin{pmatrix} \mathbf{a_1}\cdot\mathbf{e_1} & \mathbf{a_2}\cdot\mathbf{e_1} & \dots & \mathbf{a_N}\cdot\mathbf{e_1} \\ 0 & \mathbf{a_2}\cdot\mathbf{e_2} & \dots & \mathbf{a_N}\cdot\mathbf{e_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{a_N}\cdot\mathbf{e_N} \end{pmatrix}. \tag{31}$$

So QR factorises **A** into **Q** containing directional information and **R** with length information.

Drawing all this together, the QR-Schur iteration has these two steps repeated in each cycle:

1. factorise $\mathbf{A_k} = \mathbf{Q_k}\,\mathbf{R_k}$,

2. reverse multiply to obtain $\mathbf{A_{k+1}} = \mathbf{R_k}\,\mathbf{Q_k}$.

From step 1 $\mathbf{R_k} = \mathbf{Q_k}^{-1}\mathbf{A_k}$ so $\mathbf{A_{k+1}} = \mathbf{Q_k}^{-1}\mathbf{A_k}\,\mathbf{Q_k}$, a similarity transformation. Propagating this back to $\mathbf{A_0}$

$$\mathbf{A_{k+1}} \;=\; \mathbf{Q_k}^{-1}\,\mathbf{Q_{k-1}}^{-1} \;....\; \mathbf{Q_1}^{-1}\mathbf{Q_0}^{-1}\mathbf{A_0}\,\mathbf{Q_0}\,\mathbf{Q_1} \;....\; \mathbf{Q_{k-1}}\,\mathbf{Q_k}\,.$$

Using the order-reversing property of the inverse operator

$$\mathbf{Q_k}^{-1}\,\mathbf{Q_{k-1}}^{-1} \;....\; \mathbf{Q_1}^{-1}\mathbf{Q_0}^{-1} = (\mathbf{Q_0}\,\mathbf{Q_1} \;....\; \mathbf{Q_{k-1}}\,\mathbf{Q_k})^{-1}$$

so $\mathbf{A_{k+1}}$ is similar to $\mathbf{A_0}$ and therefore has the same eigenvalues. In most cases the sequence will converge $\mathbf{Q_k} \to \mathbf{I}$ and $\mathbf{A_k} \to \mathbf{A}_\infty$, an upper triangular matrix, at rates dependent on λ_j/λ_1. Then

$$\mathbf{A} \;\equiv\; \mathbf{A_0} \;=\; \mathcal{Q}\mathbf{A}_\infty\,\mathcal{Q}^{-1}, \qquad \mathcal{Q} \;=\; \mathbf{Q_0}\,\mathbf{Q_1} \;....\; \mathbf{Q_{k-1}}\,\mathbf{Q_k}\,. \tag{32}$$

$\mathcal{Q}$ is orthogonal. This is called the Schur form or Schur factorisation or Schur decomposition of $\mathbf{A}$. The required eigenvalues are read from the diagonal of $\mathbf{A}_\infty$ where they will be arranged in descending order of magnitude.

With many matrices the algorithm does work out just as described. However with others $\mathcal{Q}$ does not converge to the identity but to a variant with +1 or −1 in each row and each column and 0s elsewhere. The examples below and in Appendix 3 illustrate this.

8.2 Three numerical examples

I wrote a computer program to carry out Gram-Schmidt orthogonalisation and form the matrix $\mathbf{R}$ of scalar products. This was then placed in an iterative loop which would terminate when all values down the diagonal of the reverse product matrix $\mathbf{RQ}$ changed by less than 10^{-7}.

Example 1 : Here is the procedure applied to the 4×4 matrix $\mathbf{E} = \mathbf{AB}$ of §3.3. The change during iteration 3 is

$$\mathbf{E_3} \;=\; \begin{pmatrix} 49\cdot0191 & -6\cdot6409 & 2\cdot2392 & 6\cdot9241 \\ 0\cdot8493 & 3\cdot3108 & -38\cdot0046 & -3\cdot1015 \\ 3\cdot1514 & -12\cdot3166 & 25\cdot9724 & -6\cdot0609 \\ -0\cdot0016 & 0\cdot0424 & 0\cdot0502 & 0\cdot6977 \end{pmatrix}$$

$$= \begin{pmatrix} 0\cdot9978 & 0\cdot0568 & 0\cdot0344 & -0\cdot0004 \\ 0\cdot01729 & 0\cdot2783 & -0\cdot9603 & 0\cdot0046 \\ 0\cdot06415 & -0\cdot9588 & -0\cdot2767 & 0\cdot0049 \\ -3\cdot258_{E-5} & 0\cdot0034 & 0\cdot0058 & 1\cdot0000 \end{pmatrix} \begin{pmatrix} 49\cdot1277 & -7\cdot3590 & 3\cdot2432 & 6\cdot4663 \\ 0 & 12\cdot3535 & -35\cdot3519 & 5\cdot3438 \\ 0 & 0 & 29\cdot3879 & 4\cdot8979 \\ 0 & 0 & 0 & 0\cdot6512 \end{pmatrix}.$$

$$\mathbf{R_3Q_3} \;=\; \mathbf{E_4} \;=\; \begin{pmatrix} 49\cdot0997 & -2\cdot3443 & 7\cdot8986 & 6\cdot4301 \\ -2\cdot0543 & 37\cdot3518 & -2\cdot0509 & 5\cdot2284 \\ 1\cdot8850 & -28\cdot1606 & -8\cdot1028 & 5\cdot0411 \\ -2\cdot123_{E-5} & 0\cdot0022 & 0\cdot0038 & 0\cdot6512 \end{pmatrix}.$$

It takes to iteration 53 to meet the convergence criterion. The last iteration reads

$$\mathbf{E_{52}} \;=\; \begin{pmatrix} 50\cdot4314246 & -3\cdot5962 & 0\cdot4729 & 5\cdot9699 \\ -1\cdot14_{E-6} & 37\cdot3413269 & -26\cdot5469 & 3\cdot1906 \\ 3\cdot67_{E-36} & 0 & -9\cdot42701689 & -6\cdot9489 \\ -7\cdot96_{E-98} & 0 & 0 & 0\cdot6542654 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 2\cdot 269_{E-8} & 0 & 0 \\ -2\cdot 27_{E-8} & 1 & 0 & 0 \\ 7\cdot 29_{E-38} & 0 & -1 & 0 \\ -1\cdot 58_{E-99} & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 50\cdot 4314246 & -3\cdot 5962 & 0\cdot 4729 & 5\cdot 9699 \\ 0 & 37\cdot 3413268 & -26\cdot 5469 & 3\cdot 1906 \\ 0 & 0 & 9\cdot 42701689 & 6\cdot 9489 \\ 0 & 0 & 0 & 0\cdot 6542654 \end{pmatrix}$$

$$\mathbf{R_{52}Q_{52}} = \mathbf{E_{53}} = \begin{pmatrix} 50\cdot 4314247 & -3\cdot 5962 & -0\cdot 4729 & 5\cdot 9699 \\ -8\cdot 47_{E-7} & 37\cdot 3413268 & 26\cdot 5469 & 3\cdot 1906 \\ 6\cdot 87_{E-37} & 0 & -9\cdot 4270169 & 6\cdot 9489 \\ -1\cdot 03_{E-99} & 0 & 0 & 0\cdot 6542654 \end{pmatrix}$$

Observe the following points

- **Q** does not converge to **I**, but to a unit matrix with -1 in position (3, 3). The eigenvalue λ_3 is correctly given in $\mathbf{E_{53}}$ as $-9\cdot 427...$ (see §3.3). This is somewhat at odds with the rationale for the algorithm in §7.1.
- The eigenvalues are arranged down the order in descending order of magnitude.
- The ratio of elements in position (4,1) between iterations 3 and 4 is $2\cdot 123/160\cdot 05 \approx \lambda_4/\lambda_1$. Similar expected ratios are found in positions (2,1), (3,1), (4,2).
- As a consequence the overall convergence to a triangular matrix is determine by the ratio λ_2/λ_1. The last eigenvalues to converge are λ_2 and λ_1.
- The sum of the diagonal elements is $79\cdot 000...$, the trace of the original matrix.
- The above-diagonal elements in **R** and **E** are of the same order of magnitude as the eigenvalues.

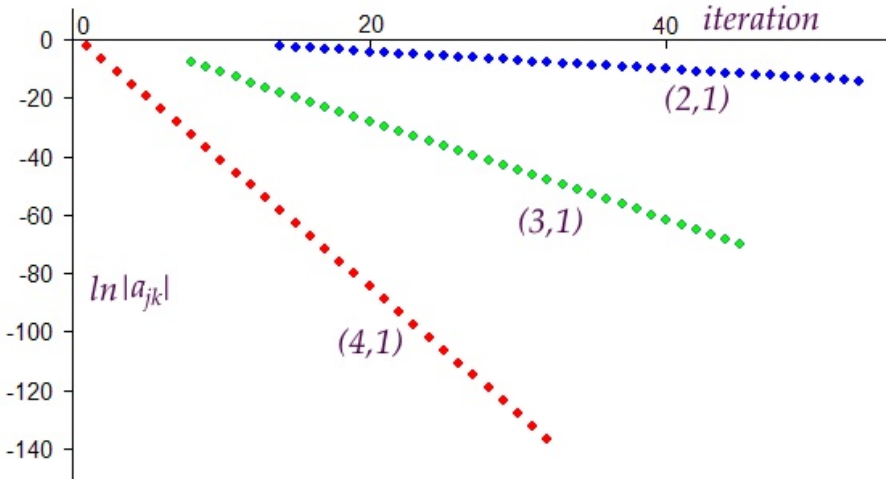

Figure 4: Convergence of matrix elements (2,1), (3,1) and (4,1) to zero with continued iteration.

To emphasise how convergence depends on the eigenvalue ratios Figure 3 plots the absolute value of the elements in positions (2,1), (3,1) and (4.1) of $\mathbf{E_k}$ on a logarithmic scale against iteration number k. The gradients settle to these values: $-0{\cdot}3005$, $-1{\cdot}6770$, $-4{\cdot}3446$ respectively corresponding to ratios $0\cdot 7404$, $0\cdot 1869$, $0\cdot 0130$. If these are multiplied by $\lambda_1 = 50\cdot 4314$, they give $37\cdot 3414$, $9\cdot 4275$, $0\cdot 6545$ in agreement with λ_2, $|\lambda_3|$ and λ_4.

Example 2 : Though in this article I have not concerned myself with complex eigenvalues, it is illuminating to see what happens in QR iterations when two eigenvalues are a conjugate pair. I chose

$$\mathbf{C} = \begin{pmatrix} 1 & 3 & -3 \\ 5 & -2 & 1 \\ 1 & 2 & 1 \end{pmatrix}$$

and ran the computer program through over 50 iterations with no diagonal shift. The trace is 0. **QW** does not converge; rather is tends to the form

$$\mathbf{Q} \rightarrow \begin{pmatrix} -1 & 0 & 0 \\ 0 & s & -t \\ 0 & t & s \end{pmatrix}, \qquad s^2 + t^2 = 1\,.$$

s and t jump in value from iteration to iteration, but $s + it$ always lies on the unit circle in the complex plane. The reverse multiplied matrix tends to

$$\mathbf{RQ} \rightarrow \begin{pmatrix} -5\cdot 1967535 & b_{12} & b_{13} \\ 0 & b_{22} & b_{23} \\ 0 & b_{32} & b_{33} \end{pmatrix},$$

where $-5{\cdot}1967535$ is the real eigenvalue. The characteristic equation of the lower right 2×2 submatrix tends to $\lambda^2 - 5\cdot 1967535\lambda + 10\cdot 0062471 = 0$ with roots $2\cdot 5983768 \pm 1\cdot 8040747\,i$, the two complex eigenvalues. Convergence is alternating and does not fit a geometric series.

Example 3 : The third example is the symmetric 4×4 matrix $\mathbf{B}$ of §3.2. This has real eigenvalues of which two have opposite signs but close absolute values, namely $-7\cdot 1056967$ and $7\cdot 0423731$. The **RQ** matrices show features similar to Example 2. At iteration 36 convergence is well advanced and the matrices are

$$\mathbf{B_{34}} = \begin{pmatrix} -0\cdot 76673 & 7\cdot 03574 & 1\cdot 99_{E-10} & 2\cdot 01_{E-11} \\ 7\cdot 03574 & 0\cdot 70341 & 5\cdot 82_{E-10} & -8\cdot 27_{E-11} \\ 4\cdot 01_{E-18} & 0 & -1\cdot 96423 & -1\cdot 49_{E-9} \\ -1\cdot 68_{E-28} & 0 & -1\cdot 49_{E-9} & 1\cdot 02755 \end{pmatrix}$$

$$= \begin{pmatrix} -0\cdot 10834 & 0\cdot 99411 & 0 & 0 \\ 0\cdot 99411 & 0\cdot 10834 & 0 & 0 \\ 5\cdot 66_{E-19} & 0 & -1 & -7\cdot 59_{E-10} \\ -2\cdot 38_{E-29} & 0 & -7\cdot 59_{E-10} & 1 \end{pmatrix} \begin{pmatrix} 7\cdot 07739 & -0\cdot 06295 & 5\cdot 57_{E-10} & -8\cdot 44_{E-11} \\ 0 & 7\cdot 07053 & 2\cdot 60_{E-10} & 1\cdot 11_{E-11} \\ 0 & 0 & 1\cdot 96423 & 7\cdot 11_{E-10} \\ 0 & 0 & 0 & 1\cdot 02755 \end{pmatrix}.$$

$$\mathbf{R_{34}Q_{34}} = \mathbf{B_{35}} = \begin{pmatrix} -0\cdot 82931 & 7\cdot 02892 & -5\cdot 57_{E-10} & -8\cdot 44_{E-11} \\ 7\cdot 02892 & 0\cdot 76599 & -2\cdot 60_{E-10} & 1\cdot 11_{E-11} \\ 1\cdot 11_{E-18} & 0 & -1\cdot 96423 & -7\cdot 80_{E-10} \\ -2\cdot 44_{E-29} & 0 & -7\cdot 80_{E-10} & 1\cdot 02755 \end{pmatrix}.$$

Eigenvalues $\lambda_3 = -1\cdot 96423$ and $\lambda_4 = 1.02755$ are correct, but we have a 2×2 symmetric submatrix in the first two rows with dominant off-diagonal elements. This is because $\mathbf{Q}$ is converging to $\left(\begin{smallmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{smallmatrix}\right)$. Note that the other off-diagonal elements of $\mathbf{B}$ are almost zero. The characteristic equation of the submatrix of $\mathbf{B_{34}}$ is $\lambda^2 + 0\cdot 0633236\lambda - 50\cdot 0409675$ with roots $-7\cdot 105697$, $7\cdot 0423731$, the two other eigenvalues. Thus numerically close eigenvalues appear in 2×2 submatrices, similar to complex eigenvalues. In Example 1 the eigenvalues down the diagonal can be regarded as lying on the diagonals of a sequence of 2×2 upper triangular submatrices stepping down the diagonal. These are special cases of the general 2×2 submatrices seen in Examples 2 and 3, and suggest that in the QR algorithm attention should be given to the 2×2 submatrices, especially when the orthogonal matrix $\mathbf{Q}$ is not converging to a diagonal unit matrix.

8.3 QR's convergence rate and diagonal shifting

Convergence of the classic QR algorithm with real eigenvalues is convergence to an upper triangular matrix. It is the tending to zero of the below-diagonal elements which tells us that the diagonal elements are tending to the eigenvalues. The rate of below-diagonal convergence is determined by the ratios λ_j/λ_1, $j = 2, 3, 4, ..., N$ and in this respect it is similar to the Power Method. We might hope that improved convergence would be obtained by the two devices of a) shifting the matrix diagonal values and b) summing a geometric series, both of which are effective in the Power Method. To be clear, §7.1 has *not* shown at what rate the diagonal elements approach the eigenvalues. A visual comparison of the diagonals of **QR** and **RQ** at §7.1 shows how the b_{jk} become mixed in with the λ and convergence is not obvious. Therefore we have no grounds to expect geometric series on the diagonal elements even though they have been shown to occur below the diagonal. Nevertheless, geometric series do seem to occur in some cases, and solving the 2×2 submatrices then fitting geometric series to the iterated eigenvalue estimates may be fruitful as the case below illustrates.

Look again at the matrix **B** of §3.2 and in example 2 of §8.2. At the end of iteration 3 the reverse-multiplied matrix $\mathbf{B_4}$ is

$$\begin{pmatrix} 1.1813 & 6.9302 & 0.7336 & -0.0152 \\ 6.9302 & -1.2364 & -0.0093 & 0.1058 \\ 0.7336 & -0.0093 & -1.776 & -0.7293 \\ -0.0152 & 0.1058 & -0.7293 & 0.8336 \end{pmatrix}.$$

Q is already converging to $\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ and the upper left 2×2 submatrix already shows equally dominant elements off the diagonal. The characteristic equation of this submatrix is $\lambda^2 + 0.055039\lambda - 49.487961$ with roots $-7 \cdot 062342$ and $7 \cdot 007302$. Do the same with the lower right 2×2 submatrix; it has characteristic equation $\lambda^2 + 0 \cdot 944961\lambda - 2 \cdot 014381$ with roots $-1 \cdot 968348$ and $1 \cdot 023387$. These four roots of two quadratic equations are approximations to the eigenvalues of **B**. Table 5 lists the corresponding values at the close of iterations 4 and 5. In Table 6 for each eigenvalue I show the differences $\delta_4 = \lambda^{(4)} - \lambda^{(3)}$, $\delta_5 = \lambda^{(5)} - \lambda^{(4)}$, their ratio r, and the sum to infinity of their geometric series according to the formula $\delta_5/(1 - r)$. This sum Σ_∞ is added to the estimate $\lambda^{(4)}$ to give a projected final value for λ. The errors in the last column show remarkable agreement with the precise eigenvalues, and all after only five iterations. Clearly these projected values are good enough to be used as diagonal shifts in, say, the inverse Power Method, and we can expect rapid convergence of both the eigenvalues and eigenvectors.

Iteration	3	4	5
λ_1	$-7 \cdot 062342$	$-7 \cdot 102659$	$-7 \cdot 105471$
λ_2	$7 \cdot 007302$	$7 \cdot 039693$	$7 \cdot 042166$
λ_3	$-1 \cdot 968348$	$-1 \cdot 964499$	$-1 \cdot 964245$
λ_4	$1 \cdot 023387$	$1 \cdot 027464$	$1 \cdot 027550$

Table 5: Eigenvalue estimates at iterations 3, 4 , 5 for **B**, §3.2.

As with the Power Method, I have not seen this approach described in the literature, though it seems so obvious that others must have developed it. It is probably similar to Aiken's acceleration technique, which is know to offer benefit only where the series converges approximately as a geometric series. To be valid the ratios of consecutive pairs of differences δ_k must be almost equal and have an absolute value < 1 and preferably less than about $0 \cdot 6$. It may be that this is not achieved often

	δ_4	δ_5	r	Σ_∞	λ projected	error
λ_1	$-0\cdot0403170$	$-0\cdot0028122$	$0\cdot0697531$	$-0\cdot0030231$	$-7\cdot1056816715$	$1\cdot5_{E-5}$
λ_2	$0\cdot0323913$	$0\cdot0024727$	$0\cdot0763378$	$0\cdot0026770$	$7\cdot0423703793$	$-2\cdot7_{E-6}$
λ_3	$0\cdot0038487$	$0\cdot00025364$	$0\cdot0659012$	$0\cdot0002715$	$-1\cdot9642273288$	$8\cdot6_{E-7}$
λ_4	$0\cdot0040770$	$8\cdot59182_{E-5}$	$0\cdot0210738$	$8\cdot77678_{E-5}$	$1\cdot0275518526$	$2\cdot1_{E-8}$

Table 6: Geometric series fitted to differences δ, and projected eigenvalues.

enough for geometric series to be a worthwhile part of the QR algorithm. Appendix 3 describes QR applied to a challenging 6×6 matrix.

In §4.3, Eq 16 we saw how shifting the diagonal elements by the same constant will greatly change the rate of convergence in the Power Method. Some empirical evidence of the effect of shifting in given in Figure 5 for the symmetric matrix **B** of §3.2 studied above. A fixed shift (the horizontal axis) was applied for all iterations and the calculation continued until all λ estimates down the diagonal by less than 10^{-7} from one iteration to the next. Figure 5 plots $\log_{10}$ of the number of iterations required for this to be met as a function of the shift value β subtracted from the diagonal. The eigenvalues are $-7\cdot10266$, $-1\cdot9645$, $1\cdot0275$ and $7\cdot0397$. What is most noticeable is the complete failure to converge at six values corresponding to the spikes in Figure 5[7]. These peaks correspond with the arithmetic means of pairs of eigenvalues, these being listed in the table below.

eigenvalues	$-7\cdot10266$	$-1\cdot9645$	$1\cdot0275$
$-1\cdot9645$	$-4\cdot53358$		
$1\cdot0275$	$-3\cdot03758$	$-0\cdot4685$	
$7\cdot0397$	$-0\cdot03148$	$2\cdot5376$	$4\cdot0336$

When the shift is a mean value $(\lambda_i+\lambda_j)/2$, there is no reduction in the off-diagonal elements between **QR** and **RQ** so the algorithm never converges. When the algorithm did converged, **Q** was

$$\begin{pmatrix}-1&0&0&0\\0&1&0&0\\0&0&-1&0\\0&0&0&1\end{pmatrix}\quad\text{or}\quad\begin{pmatrix}-1&0&0&0\\0&-1&0&0\\0&0&1&0\\0&0&0&-1\end{pmatrix}\quad\text{or}\quad\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&-1&0\\0&0&0&1\end{pmatrix}$$

and the product **RQ** had all elements virtually zero except in the upper left 2×2 submatrix and on the diagonal.

Let us now look further at what happens when the shift β is close to an eigenvalue. With $\beta = 1\cdot03$ the reverse-multiplied matrix at the end of iteration 2 is

$$\begin{pmatrix}-3\cdot775&4\cdot831&4\cdot373&-6\cdot1_{E-7}\\4\cdot830&0\cdot050&2\cdot258&3\cdot1_{E-6}\\4\cdot373&2\cdot258&-1\cdot393&-2\cdot1_{E-6}\\-6\cdot1_{E-7}&3\cdot1_{E-6}&-2\cdot1_{E-6}&-0\cdot002448\end{pmatrix}.$$

Observe that the last row and the last column are already almost zero apart from the shifted eigenvalue approximation in position (4, 4). By iteration 4 all off-diagonal elements in this row and this column are less than 3×10^{-12}. In other words, the eigenvalue close to $1\cdot03$ has been found precisely as $-0\cdot0024481687+1\cdot03 = 1\cdot02755183127$ in only 4 iterations. The matrix can now be deflated simply

[7] At four of these spikes, $-4\cdot53$, $-3\cdot04$, $2\cdot54$ and $4\cdot03$, I have plotted a nominal 1400 iterations because the program ran indefinitely.

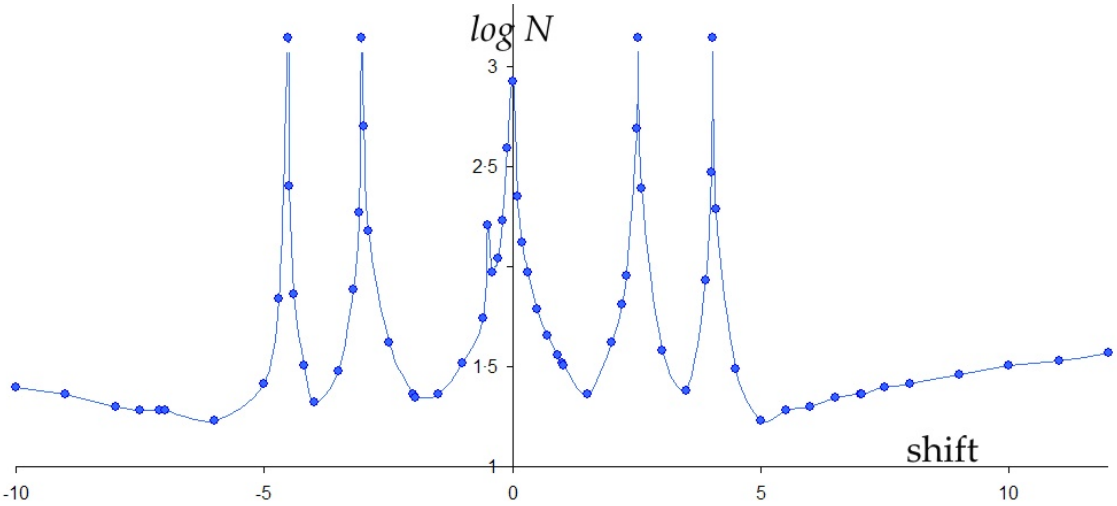

Figure 5: $\log_{10} N$ where is N is the number of iterations for the convergence criterion to be met as a function of shift β subtracted form the diagonal of matrix $\mathbf{B}$.

by deleting the bottom row and last column to leave a 3×3 matrix with the same three remaining eigenvalues. We then have a smaller matrix which will converge to its smallest absolute eigenvalue if given a close enough shift.

So how is a good estimate of an eigenvalue to be made to start the shifting process? The bounds of Wolkowicz and Styan determined in §3.2 for $\mathbf{B}$ are one starting place, as is the method of counting positive and negative pivots under various shifts. Another approach could be to apply a few iterations of the direct power method, which will point to the largest absolute eigenvalue. Following §3.2 suppose initially we take $\beta = 2$ and run the QR algorithm for a few iterations, calculating at each iteration the eigenvalues of the 2×2 submatrices down the diagonal. At iteration 4 with $\beta = 2$ these pairs are $(1 \cdot 23, -6 \cdot 78)$, $(6 \cdot 46, -1 \cdot 97)$, $(3 \cdot 52, 1 \cdot 02516)$ respectively, so here already is $1 \cdot 025$ giving a good estimate of the smallest eigenvalue. Now run the algorithm with a shift of $1 \cdot 02516$ for only one iteration and evaluate the last (bottom right) 2×2 submatrix. Its eigenvalues are $0 \cdot 00238089$ and $-1 \cdot 789406$, so with the shift restored these are $1 \cdot 027545$ and $-0 \cdot 76424$. Now take $1 \cdot 027545$ as a shift and run the algorithm for one further iteration. At this stage the bottom (fourth) row and last column have all elements almost zero; the matrix has become almost singular. The appropriate eigenvalue of the last 2×2 submatrix is $6 \cdot 790257 \times 10^{-6}$ which, when shifted back by the latest β, gives a good eigenvalue approximation of $1 \cdot 02755183$. We saw this strategy of using the latest best estimate of the eigenvalue as the shift β when assessing the Power Method, and clearly it works very well with the QR method. Continual shifting was used by Francis as explained in §10.

8.4 Note on Krylov subspaces

As noted in the Introduction, §1, the **QR** method is effective for dense matrices (ones with no or only a few zero elements), but has computer storage difficulties when matrices are very large, as are most from finite element models. The Power Method, in contrast, requires only that the given matrix $\mathbf{E}$ (or a diagonally shifted copy) repeatedly multiply a vector. $\mathbf{E}$ is not changed in the process. This makes it a potentially useful starting point for developing algorithms for very large sparse matrices which can be stored only as an indexed list of their non-zero elements.

Recall that the Direct Power Method is started by making a guess at an eigenvector, $\mathbf{v}_0$.

This is multiplied several times by $\mathbf{E}$, and $\mathbf{v}_k = \mathbf{E}^k\mathbf{v}_0$ in many cases converges to the dominant eigenvector, simultaneously revealing the associated eigenvalue. I show in §4.2 and Appendix 1 that if the sequence of $\mathbf{v}_k,\ k = 1, 2, ..., n$ is analysed as a geometric series, the eigenvector can be predicted and hence convergence accelerated. The set of $\mathbf{v}_k$ so generated span a vector space called a 'Krylov subspace'after the Russian engineer Alexei Krylov who discussed their use in 1931. Several methods for large sparse matrices manipulate the Krylov subspace to find one or more eigen pairs. Since as k increases, the $\mathbf{v}_k$ become closer and closer to each other and to the dominant eigenvector and so lose their independence, several of these algorithms first produce an orthonormal spanning set of independent vectors using either the Gram-Schmidt procedure, §8.1, or something similar. One similar and more stable orthogonalisation scheme is the Arnoldi which produces an alternative set of orthonormal vectors spanning the Krylov subspace of dimension n. Approximate eigenvectors are obtained by finding a vector $\mathbf{p}$ within the Krylov subspace – that is, a linear combination of the orthonormal base vectors – which minimises some function which would be zero for the true solution of the eigenvalue problem, such as the modulus of $(\mathbf{E} - \lambda\mathbf{I})\mathbf{p}$.

Krylov subspace methods are also used to solve large systems of linear simultaneous equations which in matrix form are $\mathbf{Ex} = \mathbf{c}$. Formally the solution is $\mathbf{x} = \mathbf{E}^{-1}\mathbf{c}$ and we recall that by the Cayley-Hamilton theorem the $n \times n$ inverse matrix $\mathbf{E}^{-1}$ can be written as a polynomial in $\mathbf{E}$ of degree $n - 1$. In many cases a polynomial of much lower degree can give a good enough approximation. We then would have $\mathbf{x} \approx (a_1\mathbf{E}+a_2\mathbf{E}^2+a_3\mathbf{E}^3+....+a_m\mathbf{E}^m)\mathbf{c}$ and we see that the vectors $\mathbf{E}^k\mathbf{c}$ are the generators of the Krylov subspace of dimension m formed by $\mathbf{E}$ operating on the given vector $\mathbf{c}$. The approximate solution $\mathbf{x}$ is a vector lying within this Krylov subspace. The various iterative methods determine the coefficients a_k to minimise a suitable residual function.

According to Wikipedia the best known Krylov subspace methods for eigen pairs are the Arnoldi, Lanczos (for symmetric matrices), Conjugate Gradient, IDR (Induced dimension reduction), GMRES (generalized minimum residual), BiCGSTAB (biconjugate gradient stabilized), QMR (quasi-minimal residual), TFQMR (transpose-free QMR), and MINRES (minimal residual) methods. I refer the interested reader to the literature.

9 Transformation to Hessenberg form

Both the Jacobi and basic QR methods gradually reduce a symmetric matrix to diagonal form, and a general matrix to triangular form, by a sequence of similarity transformation which preserve the eigenvalues. We now look at a half-way position in which similarity transformations convert the elements in the lower left corner of a general matrix into zeros, and correspondingly, the lower left and upper right corners of a symmetric matrix into 0s. The former are called Hessenberg matrices after Karl Hessenberg. They are nearly triangular, having all entries below the first sub-diagonal zero. In fact John Francis in his seminal papers calls them 'almost triangular'. The equivalent for a symmetric matrix is the tridiagonal form, with only the central three diagonals non-zero. Intuitively, we might expect that eliminating as many elements as possible would ease the eventual determination of the eigenvalues by reducing the number of necessary floating point operations in the computer. It is an interesting consequence of Niels Abel's impossibility theorem that there cannot be an algorithm to produce a similarity transformation of a general matrix to diagonal form which does not require that the eigenvalues – that is, the diagonal elements – first be known. This is because the eigenvalues are equivalent to the roots of the characteristic polynomial, and Abel's theorem states that for polynomials of degree 5 and greater there is no closed formula or finite algorithm for the roots. However, there are algorithms which will produce Hessenberg or tridiagonal form in a finite number of steps.

The advantages of Hessenberg form will become clearer below, but here I record the number of multiplications of matrix elements necessary to evaluate some types of matrix product. An $n \times n$ matrix has n^2 elements and when two general $n \times n$ matrices are multiplied n multiplications and $n-1$ additions are required to calculate each element in the product, so n^3 floating point $\times$ operations in all are needed. In contrast, the product of two upper triangular matrices, which is also an upper triangular matrix, requires $n(n+1)(n+2)/6 \approx n^3/6 + n^2/2$ $\times$ operations. Hessenberg matrices lies between these extremes. The product of two Hessenberg matrices is a matrix with two sub-diagonals below the main diagonal. It requires $(n-1)(n+4)(n+6)/6 \approx n^3 + 3n^2/2$ $\times$ operations. The product of a Hessenberg and an upper triangular matrix is another Hessenberg matrix and requires $(n+2)(n^2+4n-3)/6 \approx n^3+n$ $\times$ operations. By careful choice of algorithm LU and QR decomposition of Hessenberg matrices can be achieved in only n^2 operations. Seeing the computational benefits of Hessenberg form, Francis assumed it as the starting point in all his analsyis.

9.1 Householder reflectors

Whilst the Jacobi method achieves diagonal form with a series of elementary rotations, we now use a series of elementary reflections. The left panel in Figure 5 is a 3-D diagram showing a vector $\mathbf{a_0}$ being reflected in a mirror plane into $\mathbf{a_1}$ lying along the x-axis. The mirror passes through the origin. In this co-ordinate system $\mathbf{a_1}$ has only one non-zero co-ordinate. $\mathbf{m}$ is a unit vector in the mirror plane, lying in the same plane as $\mathbf{a_0}$ and $\mathbf{a_1}$. Clearly $|\mathbf{a_0}| = |\mathbf{a_1}|$ and $\mathbf{m}$ is along $\mathbf{a_0} + \mathbf{a_1}$. $\mathbf{n}$ is a unit vector normal to the plane and is along $\mathbf{a_0} - \mathbf{a_1}$. We want a matrix $\mathbf{P}$ which will effect this reflection; that is, $\mathbf{P a_0} = \mathbf{a_1}$. Now $\mathbf{a_0} = (a \cdot m)\mathbf{m} + (a \cdot n)\mathbf{n}$ and $\mathbf{a_1} = (a \cdot m)\mathbf{m} - (a \cdot n)\mathbf{n}$. We therefore need $\mathbf{Pm} = \mathbf{m}$ and $\mathbf{Pn} = -\mathbf{n}$. The 'trick' here is to see that the matrix $\mathbf{P'} = \mathbf{n}\mathbf{n}^T$ will almost achieve this. $\mathbf{P'm} = \mathbf{n}(n^T \cdot m) = 0$ since $\mathbf{m}$ and $\mathbf{n}$ are perpendicular. $\mathbf{P'n} = \mathbf{n}(n^T \cdot n) = \mathbf{n}$ since $\mathbf{n}$ is a unit vector. The required matrix $\mathbf{P}$ is now seen to be $\mathbf{I} - 2\mathbf{n}\,\mathbf{n}^T$. It is called a 'Householder's reflector' after Alston Householder who developed its use in the late 1950s. The same formula for $\mathbf{P}$ holds in higher dimensions.

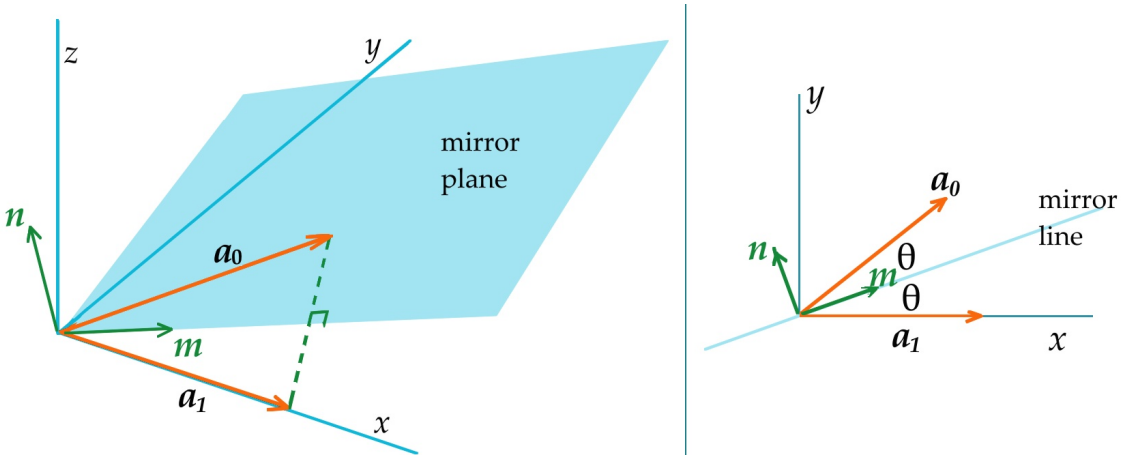

Figure 6: A vector $\mathbf{a_0}$ reflected into the x-axis as $\mathbf{a_1}$. Left: 3-D view, Right: 2-D view.

It is interesting to see the form of $\mathbf{P}$ in only 2 dimensions. This is shown in the right panel in Figure 5 where the mirror line is at angle θ to the x-axis. $\mathbf{a_0} = (a_x, a_y)$. Vector $\mathbf{m} = (c,\ s)$ where

$c = \cos\theta$, $s = \sin\theta$, and $\mathbf{n} = (-s,\ c)$. Therefore

$$\mathbf{P'} \equiv \mathbf{n}\,\mathbf{n}^T = \begin{pmatrix} -s \\ c \end{pmatrix}\begin{pmatrix} -s & c \end{pmatrix} = \begin{pmatrix} s^2 & -cs \\ -cs & c^2 \end{pmatrix},$$

$$\mathbf{P} \equiv \mathbf{I} - 2\mathbf{P'} = \begin{pmatrix} 1-2s^2 & 2cs \\ 2cs & 1-2c^2 \end{pmatrix} = \begin{pmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{pmatrix} = \frac{1}{|a|}\begin{pmatrix} a_x & a_y \\ a_y & -a_x \end{pmatrix}. \tag{33}$$

This presentation reveals the similarity with the rotation matrix used in Jacobi's method, Eq 27.

In N dimension, if the vector being reflected is $\mathbf{a_0} = (a_1, a_2, a_3, ..., a_N)$ and if $\mathbf{a_1} = (|a|, 0, 0, ..., 0)$, the unit normal is $\mathbf{n} = (a_1 - |a|, a_2, a_3, ..., a_N)/\sqrt{2|a|(|a| - a_1)}$. Calling this $(n_1, n_2, n_3, ..., n_N)$,

$$\mathbf{P} = \begin{pmatrix} 1-2n_1^2 & -2n_1n_2 & -2n_1n_3 & ... & -2n_1n_N \\ -2n_2n_1 & 1-2n_2^2 & -2n_2n_3 & ... & -2n_2n_N \\ -2n_3n_1 & -2n_3n_2 & 1-2n_3^2 & ... & -2n_3n_N \\ ... & ... & ... & ... & ... \\ -2n_Nn_1 & -2n_Nn_2 & -2n_Nn_3 & ... & 1-2n_N^2 \end{pmatrix}. \tag{34}$$

Not only is $\mathbf{P}$ symmetric but it has the very useful property of being its own inverse. Mathematically this arises in part from the normalisation of $\mathbf{n}$ to be a unit vector, but physically it corresponds to a double reflection in a mirror being identical to the original object.

A concrete numerical example may consolidate the above. Suppose that 4-D vector $\mathbf{a_0} = (3,\ 2,\ -2,\ 1\,)$ is reflected into axis 1. $|a|^2 = 18$ so the reflected vector will be $\mathbf{a_1} = (\sqrt{18},\ 0,\ 0,\ 0\,)$. $\mathbf{n} = (3 - \sqrt{18},\ 2,\ -2,\ 1\,)/3 \cdot 2472$. Then

$$\mathbf{P} = \mathbf{I_4} - 2\mathbf{n}\,\mathbf{n}^T = \begin{pmatrix} 0\cdot7071 & 0\cdot4714 & -0\cdot4714 & 0\cdot2357 \\ 0\cdot4714 & 0\cdot2413 & 0\cdot7587 & -0\cdot3794 \\ -0\cdot4714 & 0\cdot7587 & 0\cdot2413 & 0\cdot3794 \\ 0\cdot2357 & -0\cdot3794 & 0\cdot3794 & 0\cdot8103 \end{pmatrix}.$$

Direct calculation proves that $\mathbf{P}\,\mathbf{a_0} = (\sqrt{18}\,,\ 0,\ 0,\ 0,\ )$ as required.

9.2 Applying Householder reflections to obtain Hessenberg form

The example matrix I use here is

$$\mathbf{C} = \begin{pmatrix} 5 & 1 & 2 & -3 & 1 \\ 3 & 1 & -4 & 5 & 2 \\ 2 & -3 & -1 & 4 & -1 \\ -2 & 1 & 2 & -4 & 1 \\ 1 & 2 & 2 & 7 & -2 \end{pmatrix}. \tag{35}$$

This has five real eigenvalues. Observe that the first column contains the vector $\mathbf{a_0} = (3,\ 2,\ -2,\ 1\,)$ for which $\mathbf{P}$ was found in the last subsection. We want to convert this into $(\sqrt{18}\,,\ 0,\ 0,\ 0\,)$, thereby introducing three zeros into $\mathbf{C}$. Clearly $\mathbf{P}$ must be augmented to a 5×5 matrix before it can operate on $\mathbf{C}$ so we fill it out with part of the identity matrix. Call this $\hat{\mathbf{P}}$. Its block structure is

$$\hat{\mathbf{P}} = \left(\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{P} \end{array}\right)$$

where $\mathbf{P}$ is 4×4 matrix. Multiplying

$$\hat{\mathbf{P}}\,\mathbf{C} \;=\; \begin{pmatrix} 5 & 1 & 2 & -3 & 1 \\ 4\cdot 2426 & 0\cdot 7071 & 0\cdot 4714 & -0\cdot 4714 & 0\cdot 2357 \\ 0 & 0\cdot 4714 & 0\cdot 2413 & 0\cdot 7587 & -0\cdot 3794 \\ 0 & -0\cdot 4714 & 0\cdot 7587 & 0\cdot 2413 & 0\cdot 3794 \\ 0 & 0\cdot 2357 & -0\cdot 3794 & 0\cdot 3794 & 0\cdot 8103 \end{pmatrix}$$

where $4\cdot 2426 = \sqrt{18}$. This is all fine, but it is not a similarity transformation so its eigenvalues will not be those of $\mathbf{C}$. We need to post-multiply by $\hat{\mathbf{P}}^{-1} = \hat{\mathbf{P}}^{T}$. The result is

$$\hat{\mathbf{P}}\,\mathbf{C}\,\hat{\mathbf{P}}^{-1} \;=\; \begin{pmatrix} 5\cdot 0 & 3\cdot 2998 & -1\cdot 7015 & 0\cdot 7015 & -0\cdot 8508 \\ 4\cdot 2426 & -6\cdot 5000 & 5\cdot 5523 & -0\cdot 3668 & 4\cdot 6618 \\ 0 & 0\cdot 8159 & -3\cdot 0877 & -0\cdot 6487 & 1\cdot 3593 \\ 0 & -1\cdot 7587 & 2\cdot 3862 & 2\cdot 3502 & -2\cdot 2100 \\ 0 & 1\cdot 3508 & 5\cdot 0719 & 0\cdot 5599 & 1\cdot 2375 \end{pmatrix}.$$

and this does have the same eigenvalues. The zeros in the first row and first column of $\hat{\mathbf{P}}$ prevent this second multiplication from destroying the zeros which the first multiplication so obligingly introduced.

Obviously the next step in moving to Hessenberg form is to apply a Housholder reflector to the second column above. Now $\mathbf{a_0} = (0\cdot 8159,\ -1\cdot 7587,\ 1\cdot 3508\,)$ with $|a| = 2\cdot 3629$. The new unit normal is $\mathbf{n} = (-0.5721,\ -0.6504,\ 0.4996\ )$ and the new $\hat{\mathbf{P}}$ is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0\cdot 3453 & -0\cdot 7443 & 0\cdot 5717 \\ 0 & 0 & -0\cdot 7443 & 0\cdot 15389 & 0\cdot 6499 \\ 0 & 0 & 0\cdot 5717 & 0\cdot 6499 & 0\cdot 5009 \end{pmatrix}.$$

To complete this second step again post-multiply by the new $\hat{\mathbf{P}}^{-1}$ and thereby obtain

$$\begin{pmatrix} 5\cdot 0 & 3\cdot 2998 & -1\cdot 5960 & 0\cdot 8215 & -0\cdot 9429 \\ 4\cdot 2426 & -6\cdot 5000 & 4\cdot 8551 & -1\cdot 1594 & 5\cdot 2705 \\ 0 & 2\cdot 3629 & 2\cdot 8632 & 1\cdot 5368 & 0\cdot 3715 \\ 0 & 0 & 0\cdot 8462 & -4\cdot 6070 & 3\cdot 9189 \\ 0 & 0 & -0\cdot 2889 & -1\cdot 5358 & 2\cdot 2438 \end{pmatrix}.$$

A third step, operating on column 3, will complete the Hessenberg transformation. This time $\mathbf{a_0} = (0\cdot 8462,\ -0\cdot 2889)$, $\mathbf{n} = (-0\cdot 16378,\ -0\cdot 9865\,)$. We find the new

$$\hat{\mathbf{P}} \;=\; \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0\cdot 9464 & -0\cdot 3231 \\ 0 & 0 & 0 & -0\cdot 3231 & -0\cdot 9464 \end{pmatrix}.$$

The non-zero 2×2 submatrix here has the form of Eq 29 and corresponds to reflection in a mirror line inclined at $-9\cdot 4^{\circ}$ to the $(N-1)^{st}$ axis. The Hessenberg form is

$$\mathbf{C_H} \;=\; \begin{pmatrix} 5\cdot 0000 & 3\cdot 2998 & -1\cdot 5960 & 1\cdot 0821 & 0\cdot 6269 \\ 4\cdot 2426 & -6\cdot 5000 & 4\cdot 8551 & -2\cdot 8003 & -4\cdot 6131 \\ 0 & 2\cdot 3629 & 2\cdot 8632 & 1\cdot 3343 & -0\cdot 8482 \\ 0 & 0 & 0\cdot 8942 & -4\cdot 6204 & -1\cdot 5751 \\ 0 & 0 & 0 & 3\cdot 8795 & 2\cdot 2572 \end{pmatrix}. \qquad (36)$$

For later comparison, the eigenvalues of $\mathbf{C}$ and $\mathbf{C_H}$ are

$$6\cdot0028872749, \quad 3\cdot74150042346, \quad 1\cdot73064164706, \quad -3\cdot7719295645, \quad -8\cdot70309978084.$$

To calculate eigenvectors it is necessary to keep track of the matrices $\mathbf{P}$ used in the similarity transformations. Conversion to Hessenberg form is a stepping stone in crossing from the given matrix to triangular or diagonal form, so we need to know how the eigenvectors have been changed in the process. In the Jacobi method, §8, at each iteration the product was formed of the individual rotation matrices $\mathbf{P_0\,P_1\,P_2}....\mathbf{P_N}$. Each column is the eigenvector of the eigenvalue in the corresponding column of the diagonalised matrix. In converting to Hessenberg form the equivalent product of reflection matrices needs to be calculated. The result to 3 decimal places is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0\cdot707 & 0\cdot648 & -0\cdot282 & 0\cdot010 \\ 0 & 0\cdot471 & -0\cdot698 & -0\cdot435 & -0\cdot317 \\ 0 & -0\cdot471 & 0\cdot299 & -0\cdot518 & -0\cdot648 \\ 0 & 0\cdot236 & 0\cdot050 & 0\cdot680 & -0\cdot692 \end{pmatrix}.$$

The algorithm applied to a symmetric matrix transforms it to tri-diagonal form. The tri-diagonal form of $\mathbf{B}$ of §3.2 is

$$\mathbf{B} \;\rightarrow\; \begin{pmatrix} 1 & 3\cdot3166 & 0 & 0 \\ 3\cdot3166 & -0\cdot7273 & 6\cdot1065 & 0 \\ 0 & 6\cdot1065 & 0\cdot6067 & -1\cdot2232 \\ 0 & 0 & -1\cdot2232 & -1\cdot8794 \end{pmatrix}, \tag{37}$$

and the product of reflection matrices is

$$\mathbf{P_0\,P_1} \;=\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0\cdot3015 & 0\cdot8259 & 0\cdot4764 \\ 0 & 0\cdot9045 & -0\cdot0898 & -0\cdot4168 \\ 0 & -0\cdot3015 & 0\cdot5566 & -0\cdot7741 \end{pmatrix}. \tag{38}$$

This means that the eigenvectors of the tri-diagonal form must be multiplied (on the left) by $\mathbf{P_0\,P_1}$ to give the eigenvectors of the original matrix $\mathbf{B}$.

It is interesting to compare the tri-diagonal form with the iterated matrices in the Jacobi's reduction of the same matrix, each of which has the same eigenvalues. Jacobi's method gradually turns the lower left and upper right elements to zeros, so is superficially similar to Hessenberg reduction. However even at iteration 4 the diagonal in Jacobi's method bears some approximation to the eigenvalues:

$$\begin{pmatrix} -7\cdot0861 & 0\cdot0435 & 0\cdot2085 & 0\cdot3790 \\ 0\cdot0435 & -1\cdot8956 & 0 & 0\cdot4505 \\ 0\cdot2085 & 0 & 6\cdot8927 & 0\cdot9260 \\ 0\cdot3790 & 0\cdot4505 & 0\cdot9260 & 1\cdot0889 \end{pmatrix}$$

The eigenvalues of the 2×2 diagonal submatrices are even better approximations at $\{-1\cdot895, -7\cdot086\}$, $\{6\cdot893, -1\cdot896\}$, $\{7\cdot037, 0\cdot945\}$. (The true values given in §3.2 are $7\cdot042$, $1\cdot028$, $-1\cdot964$, $-7\cdot106$.) In contrast the eigenvalues of the corresponding diagonal submatrices in the Hessenberg form bear no relation to the eigenvalues: they are $\{3\cdot564, -3\cdot291\}$, $\{6\cdot083, -6\cdot203\}$, $\{1\cdot108, -2\cdot380\}$. The problem now is clearly 'How do we determine the eigenvalues of a tri-diagonal or Hessenberg form?'.

10 John Francis' Algorithm

According to David Watkins of Washington State University this is the most important algorithm for computing eigenvalues. It was named as one of the top ten algorithms of the 20th century in a review by Barry Cipra[8] alongside the Monte Carlo method and the fast Fourier transform. It was devised by John Francis, a British computer engineer, about 1960 and published in 1961 in the Computer Journal, volume 4. 'Francis' Algorithm' is the name Watkins wants it to be known by, though for many years it was called the 'implicitly shifted QR algorithm' because of similarity with the explictly shifted QR-Schur algorithm of §7. However Watkins argues in the third edition of his book that the method is sufficiently distinct.

There are in fact two versions of the algorithm in Francis' Part 2 paper; the first works for real matrices with only real eigenvalues, and the second works where the matrix elements are real but there are complex eigenvalues pairs. They are known as the 'single shift' and 'double shift' methods respectively and both have these key features:

1. To start, the given matrix is converted to Hessenberg form. I will write the matrix being solved as $\mathbf{H}$ to emphasise that it is the Hessenberg version of the given, starting matrix $\mathbf{A}$.

2. The two versions in effect implement the QR algorithm of §8.1, factorising the $n \times n$ matrix at each iteration into an orthonormal matrix $\mathbf{Q}$ and an upper triangular matrix $\mathbf{R}$, and then reverse multiplying $\mathbf{QR}$ into $\mathbf{RQ}$. However, the iteration $\mathbf{H_{k+1}} = \mathbf{Q_k}^{-1}\mathbf{H_k}\,\mathbf{Q_k}$ is implemented in these two stages: a) $\mathbf{R_k} = \mathbf{Q_k}^{-1}\mathbf{H_k}$, b) $\mathbf{H_{k+1}} = \mathbf{R_k}\,\mathbf{Q_k}$.

3. The matrix $\mathbf{Q_k}$ is not calculated explicitly by the Gram-Schmitz method, but instead implemented as if by a sequence of $n-1$ elementary matrices, each of which is a rotation about one axis. This subtle part of the algorithm is explained below.

4. A shift to the diagonal is applied at each iteration, the shift being the current closest estimate of the eigenvalue in the last row of the matrix. This accelerates convergence greatly for the reason set out in §8.3. At the end of each iteration the shift is added back. In the single shift version suppose the shift at iteration k is β_k. The algorithm in point 1) above can be written more explicitly as: a) $\mathbf{\hat{R}_k} = \mathbf{\hat{Q}_k^{-1}}(\mathbf{H_k} - \beta_k\mathbf{I})$, b) $\mathbf{H_{k+1}} - \beta_k\mathbf{I} = \mathbf{\hat{R}_k}\,\mathbf{\hat{Q}_k}$ where the $\hat{}$ denotes the appropriate matrix for the shifted diagonal.

5. The single shift algorithm homes in on one eigenvalue at the time, this being the one in the last row. The algorithm has converged to an eigenvalue, to the required precision, when the shifted matrix has become singular, with near-zeros in its bottom row and last column. When the shift is added back, the eigenvalue appears in the bottom right corner. At this stage the matrix is deflated by deleting or ignoring the last row and last column. Subsequent iterations home in on the next eigenvalue which is in the last row of the deflated matrix. In the double shift algorithm two rows are dealt with together, these containing complex conjugate eigenvalues.

Francis recognised that working with the Hessenberg form not only requires far fewer multiplication and addition operations than with general matrices, but enables some cunning simplifications. He states that the QR factorisation can be done with order n^2 operations instead of n^3.

10.1 Francis' single shift algorithm for real eigenvalues

In this account I assume that the matrix diagonal is already shifted, so drop the $\hat{}$ accents on the matrices $\mathbf{Q_k}$ and $\mathbf{R_k}$.

[8] SIAM News, Vol. 33 No 4, May 2000

Perhaps the most important ingredient in Francis's algorithm is the way the Hessenberg matrix is factorised and multiplied. Recall that in the Jacobi method, §7, off-diagonal elements are eliminated (set to zero) one at a time by pre- and post-multiplication by a rotation matrix $\mathbf{P}$ as at Eqs 27 and 28. When only pre-multiplication is used, the rotation matrix is called a 'Given's rotator' after Wallace Givens. Consider the elementary step

$$\begin{pmatrix} \mathbb{c} & \mathbb{s} \\ -\mathbb{s} & \mathbb{c} \end{pmatrix}\begin{pmatrix} f & g \\ d & h \end{pmatrix} = \begin{pmatrix} f\mathbb{c}+d\mathbb{s} & g\mathbb{c}+h\mathbb{s} \\ d\mathbb{c}-f\mathbb{s} & h\mathbb{c}-g\mathbb{s} \end{pmatrix} \tag{39a}$$

where $\mathbb{c}^2+\mathbb{s}^2=1$, $\mathbb{c}=\cos\theta$, $\mathbb{s}=\sin\theta$. The element $d\mathbb{c}-f\mathbb{s}$ will be zero if

$$\tan\theta=\frac{d}{f},\quad \mathbb{c}=\frac{f}{T_1},\quad \mathbb{s}=\frac{d}{T_1},\quad {T_1}^2=d^2+f^2\,. \tag{39b}$$

With this substitution the product is

$$\begin{pmatrix} T & \frac{1}{T}(fg+dh) \\ 0 & \frac{1}{T}(fh-dg) \end{pmatrix}. \tag{39c}$$

The significance of this is only fully seen when the 2×2 rotation matrix Eq 39a is embedded in a larger identity matrix. For a 5×5 case and with appropriate values of $\mathbb{c}$ and $\mathbb{s}$

$$\mathbf{{q_1}^{-1}H_k} = \begin{pmatrix} \mathbb{c}_1 & \mathbb{s}_1 & 0 & 0 & 0 \\ -\mathbb{s}_1 & \mathbb{c}_1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 \\ b_1 & b_2 & b_3 & b_4 & b_5 \\ 0 & c_2 & c_3 & c_4 & c_5 \\ 0 & 0 & d_3 & d_4 & d_5 \\ 0 & 0 & 0 & e_4 & e_5 \end{pmatrix} = \begin{pmatrix} T_1 & a_2' & a_3' & a_4' & a_5' \\ 0 & b_2' & b_3' & b_4' & b_5' \\ 0 & c_2 & c_3 & c_4 & c_5 \\ 0 & 0 & d_3 & d_4 & d_5 \\ 0 & 0 & 0 & e_4 & e_5 \end{pmatrix} = \mathbf{r_1}$$

$$\text{where}\quad T_1=\sqrt{a_1^2+b_1^2},\quad \mathbb{c}_1=\frac{a_1}{T},\quad \mathbb{s}_1=\frac{b_1}{T},$$

$$a_j'=(a_j\mathbb{c}_1+b_j\mathbb{s}_1)=\frac{1}{T_1}(a_1a_j+b_1b_j),\quad b_j'=-a_j\mathbb{s}_1+b_j\mathbb{c}_1=\frac{1}{T_1}(a_1b_j-b_1a_j)\,.$$

Matrix $\mathbf{{q_1}^{-1}}$ is an elementary rotation and $\mathbf{r_1}$ is the first step in moving towards a triangular matrix. The number of superfix primes $'$ on an element denotes the number of times it has changed so far. Observe that only the first two rows have been changed by this operation. Now a rotation is applied to the second row:

$$\mathbf{{q_2}^{-1}r_1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \mathbb{c}_2 & \mathbb{s}_2 & 0 & 0 \\ 0 & -\mathbb{s}_2 & \mathbb{c}_2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} T_1 & a_2' & a_3' & a_4' & a_5' \\ 0 & b_2' & b_3' & b_4' & b_5' \\ 0 & c_2 & c_3 & c_4 & c_5 \\ 0 & 0 & d_3 & d_4 & d_5 \\ 0 & 0 & 0 & e_4 & e_5 \end{pmatrix} = \begin{pmatrix} T_1 & a_2' & a_3' & a_4' & a_5' \\ 0 & T_2 & b_3'' & b_4'' & b_5'' \\ 0 & 0 & c_3' & c_4' & c_5' \\ 0 & 0 & d_3 & d_4 & d_5 \\ 0 & 0 & 0 & e_4 & e_5 \end{pmatrix} = \mathbf{r_2}$$

$$\text{where}\quad T_2=\sqrt{{b_2'}^2+c_2^2},\quad b_j''=\frac{1}{T_2}(b_2'b_j'+c_2c_j),\quad c_j'=\frac{1}{T_2}(b_2'c_j-c_2b_j')\,.$$

With $n-1$ such rotations the matrix is transformed to the triangular $\mathbf{r_{n-1}}=\mathbf{R}$ and the orthogonal matrix $\mathbf{Q}^{-1}$ for this iteration is the product $\mathbf{{q_{n-1}}^{-1}}....\mathbf{{q_2}^{-1}{q_1}^{-1}}$. That completes the first half of the iteration. In each row one $\surd$ operation plus 2 $\times$ and 2 $\div$ operations are required to calculated T, $\mathbb{c}$ and $\mathbb{s}$, then a further $4(r-1)$ $\times$ operations where r is the index of the row. This accumulates to $n-1$ $\surd$ and $2(n+2)(n-1)$ floating point multiplication or division operations.

The second stage in each iteration is $\mathbf{R_k\,Q_k = H_{k+1}}$. The inverse of $\mathbf{Q_k}^{-1}$ is simply its transpose in the case of real matrix elements, or conjugate transpose in the case of complex ones. It reconverts the triangular form of $\mathbf{R}$ back into the Hessenberg form $\mathbf{H_{k+1}}$

$$\mathbf{Rq_1} = \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 \\ 0 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ 0 & 0 & \gamma_3 & \gamma_4 & \gamma_5 \\ 0 & 0 & 0 & \delta_4 & \delta_5 \\ 0 & 0 & 0 & 0 & \epsilon_5 \end{pmatrix} \begin{pmatrix} \mathbb{c}_1 & -\mathbb{s}_1 & 0 & 0 & 0 \\ \mathbb{s}_1 & \mathbb{c}_1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \alpha'_1 & \alpha'_2 & \alpha_3 & \alpha_4 & \alpha_5 \\ \beta'_1 & \beta'_2 & \beta_3 & \beta_4 & \beta_5 \\ 0 & 0 & \gamma_3 & \gamma_4 & \gamma_5 \\ 0 & 0 & 0 & \delta_4 & \delta_5 \\ 0 & 0 & 0 & 0 & \epsilon_5 \end{pmatrix} = \mathbf{s_1}$$

$$\text{where} \quad \alpha'_1 = \alpha_1\mathbb{c}_1 + \alpha_2\mathbb{s}_1, \quad \alpha'_2 = -\alpha_1\mathbb{s}_1 + \alpha_2\mathbb{c}_1, \quad \beta'_1 = \beta_2\mathbb{s}_1, \quad \beta'_2 = \beta_2\mathbb{c}_1\,.$$

$\mathbf{s_1}$ is the first step in moving towards the Hessenberg matrix. Observe that only the four top left elements have been changed. Next a rotation is applied to the second row and it changes elements only in columns 2 and 3:

$$\mathbf{s_1 q_2} = \begin{pmatrix} \alpha'_1 & \alpha'_2 & \alpha_3 & \alpha_4 & \alpha_5 \\ \beta'_1 & \beta'_2 & \beta_3 & \beta_4 & \beta_5 \\ 0 & 0 & \gamma_3 & \gamma_4 & \gamma_5 \\ 0 & 0 & 0 & \delta_4 & \delta_5 \\ 0 & 0 & 0 & 0 & \epsilon_5 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \mathbb{c}_2 & -\mathbb{s}_2 & 0 & 0 \\ 0 & \mathbb{s}_2 & \mathbb{c}_2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \alpha'_1 & \alpha''_2 & \alpha'_3 & \alpha_4 & \alpha_5 \\ \beta'_1 & \beta''_2 & \beta'_3 & \beta_4 & \beta_5 \\ 0 & \gamma'_2 & \gamma'_3 & \gamma_4 & \gamma_5 \\ 0 & 0 & 0 & \delta_4 & \delta_5 \\ 0 & 0 & 0 & 0 & \epsilon_5 \end{pmatrix} = \mathbf{s_2}$$

$$\text{Here} \quad \alpha''_2 = \alpha'_2\mathbb{c}_2 + \alpha_3\mathbb{s}_2\,, \quad \alpha'_3 = -\alpha'_2\mathbb{s}_2 + \alpha_3\mathbb{c}_2\,,$$

$$\beta''_2 = \beta'_2\mathbb{c}_2 + \beta_3\mathbb{s}_2\,, \quad \beta'_3 = \beta'_2\mathbb{s}_2 + \beta_3\mathbb{c}_2\,, \quad \gamma'_2 = \gamma_3\mathbb{s}_2\,, \quad \gamma'_3 = \gamma_3\mathbb{c}_2\,.$$

In $n-1$ such operations $\mathbf{H_{k+1}}$ has been calculated. That completes the second and final half of the iteration. I calculate that $2(k+1)$ elements are changed in multiplying by $\mathbf{q_k}$ and that $2n^2-2$ floating point × operations are used in this second part. Therefore the total number of multiplication or division operations in a full iteration is $2(2n+3)(n-1) \approx 4n^2+2n$.

The reader will see that the rotation matrices $\mathbf{q_k}$ and their inverses do not need to be formed explicitly. A great simplification for computer storage and CPU time comes about because the matrix multiplication is rendered through × operations on a few matrix elements at a time.

I have not yet explained why the algorithm is called 'single shift'. The reason is simply that at each iteration the best current estimate, μ, of the real eigenvalue being sought is subtracted from the diagonal of the matrix. This was mentioned in item 4 on the first page of §10.

The algorithm is primarily a very efficient method for finding eigenvalues but not eigenvectors. In principle the eigenvectors of the given matrix can be found from those of the deflated near-triangular matrices when the algorithm has converged to an eigenvalue. This would have three stages:

1. find the eigenvector of the resulting near-triangular matrix [9],
2. convert it to the corresponding eigenvector of the Hessenberg matrix using the product of all elementary matrices $\mathbf{q}$,

[9] The final two eigenvalues are found by solving the 2×2 sub-matrix on the diagonal in rows 1 and 2 rather than by reducing the matrix fully to triangular form. The eigenvectors would be found by solving two simultaneous equations derived from these two rows.

3. convert it to the corresponding eigenvector of the original matrix using the transformation matrix $\mathbf{P}$ as at Eq 38.

This would involve considerable matrix multiplication. It may be simpler, therefore, just to find the eigenvectors as a separate and subsequent operation using, for instance, the LU decomposition method of §5.1.

10.2 Numerical example

Staying with a 5×5 matrix, I will used $\mathbf{C}$ of §9.2, Eq 35 since we already have its Hessenberg form Eq 36. The first stage is to find the eigenvalue 5 associated with the bottom row. Start with no diagonal shift and apply the first rotation matrix to eliminate the element at (2,1). This matrix $\mathbf{q}$ and the resulting $\mathbf{qC_H}$ are respectively

$$\begin{pmatrix} 0\cdot7625 & 0\cdot6470 & 0 & 0 & 0 \\ -0\cdot6470 & 0\cdot7625 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 6\cdot557 & -1\cdot689 & 1\cdot924 & -0\cdot986 & -2\cdot507 \\ 0 & -7\cdot091 & 4\cdot735 & -2\cdot835 & -3\cdot923 \\ 0 & 2\cdot363 & 2\cdot863 & 1\cdot334 & -0\cdot848 \\ 0 & 0 & 0\cdot894 & -4\cdot620 & -1\cdot575 \\ 0 & 0 & 0 & 3\cdot880 & 2\cdot257 \end{pmatrix}.$$

Notice how element (2,1) is now zero. Proceeding to apply similar rotation matrices to rows 2 and 3, then 3 and 4, their accumulated product is a matrix $\mathbf{Q}$. This and the resulting triangular matrix $\mathbf{QC_H}$ are

$$\begin{pmatrix} 0\cdot762 & 0\cdot647 & 0 & 0 & 0 \\ 0\cdot614 & -0\cdot723 & 0\cdot316 & 0 & 0 \\ -0\cdot200 & 0\cdot236 & 0\cdot928 & 0\cdot208 & 0 \\ -0\cdot032 & 0\cdot038 & 0\cdot151 & -0\cdot748 & 0\cdot645 \\ 0\cdot0274 & -0\cdot032 & -0\cdot127 & 0\cdot631 & 0\cdot764 \end{pmatrix}, \quad \begin{pmatrix} 6\cdot557 & -1\cdot689 & 1\cdot924 & -0\cdot987 & -2\cdot507 \\ 0 & 7\cdot475 & -3\cdot587 & 3\cdot112 & 3\cdot454 \\ 0 & 0 & 4\cdot307 & -0\cdot598 & -2\cdot327 \\ 0 & 0 & 0 & 6\cdot015 & 2\cdot309 \\ 0 & 0 & 0 & 0 & 1\cdot005 \end{pmatrix}.$$

The second stage of this first iteration is to right multiply in turn by the inverse row rotation matrices $\mathbf{q}^{-1}$, the net effect being $(\mathbf{QC_H})\mathbf{Q}^{-1}$ which is

$$\begin{pmatrix} 3\cdot907 & 5\cdot855 & -0\cdot129 & -0\cdot867 & -2\cdot548 \\ 4\cdot836 & -6\cdot541 & -0\cdot920 & -0\cdot352 & 4\cdot817 \\ 0 & 1\cdot362 & 3\cdot873 & -0\cdot406 & -2\cdot703 \\ 0 & 0 & 1\cdot249 & -3\cdot007 & 5\cdot559 \\ 0 & 0 & 0 & 0\cdot648 & 0\cdot768 \end{pmatrix}.$$

Thus we arrive back at a Hessenberg form. There is no shift to add back.

We start the second iteration by applying a shift. The bottom right 2×2 matrix $\left(\begin{smallmatrix} -3\cdot007 & 5\cdot559 \\ 0\cdot648 & 0\cdot768 \end{smallmatrix}\right)$ has eigenvalues $1\cdot5574$ and $-3\cdot7965$, so we choose the one closest to the element $0\cdot768$ to be the shift value, *viz.* $1\cdot5574$. In this iteration stage one produces a triangular matrix and stage two another Hessenberg one. These are respectively

$$\begin{pmatrix} 5\cdot377 & -4\cdot725 & -0\cdot884 & -0\cdot696 & 3\cdot219 \\ 0 & 8\cdot910 & 0\cdot636 & -0\cdot680 & -4\cdot759 \\ 0 & 0 & 2\cdot569 & -2\cdot487 & 0\cdot956 \\ 0 & 0 & 0 & 3\cdot895 & -5\cdot880 \\ 0 & 0 & 0 & 0 & 0\cdot192 \end{pmatrix}, \quad \begin{pmatrix} -1\cdot901 & 6\cdot685 & -2\cdot024 & 0\cdot2111 & 3\cdot229 \\ 8\cdot015 & -3\cdot751 & 0\cdot739 & 0\cdot380 & -4\cdot890 \\ 0 & 0\cdot393 & 1\cdot010 & 3\cdot518 & 0\cdot375 \\ 0 & 0 & 1\cdot893 & -4\cdot335 & -5\cdot232 \\ 0 & 0 & 0 & 0\cdot032 & 0\cdot189 \end{pmatrix}$$

and at this point the shift of $1\cdot5574$ is added back to the diagonal.

On starting iteration 3 the shift derived from the bottom right 2×2 matrix is $1 \cdot 7094$. At iteration 4 the shift is $1 \cdot 73058$, then $1 \cdot 73064167$, and at iteration 6 is $1 \cdot 730641647064$. The bottom row of the Hessenberg matrix before the diagonal shift is added back is now

$$1 \cdot 2E_{-19} \quad 1 \cdot 8E_{-19} \quad 8 \cdot 6E_{-21} \quad -1 \cdot 77E_{-21} \quad -1 \cdot 8E_{-17}\,;$$

that is the bottom right element is essentially zero. Adding back the shift gives the Hessenberg matrix

$$\begin{pmatrix} -8 \cdot 733 & -0 \cdot 926 & -0 \cdot 682 & -1 \cdot 247 & 5 \cdot 864 \\ 0 \cdot 497 & 6 \cdot 060 & -1 \cdot 007 & -0 \cdot 917 & -0 \cdot 933 \\ 0 & 0 \cdot 207 & -3 \cdot 690 & 1 \cdot 961 & 4 \cdot 574 \\ 0 & 0 & 0 \cdot 390 & 3 \cdot 633 & -2 \cdot 448 \\ 0 & 0 & 0 & 0 & 1 \cdot 7306... \end{pmatrix}$$

and now the bottom row is zero except for the last element, which must be the eigenvalue $1 \cdot 7306416470644$.

The algorithm next deletes (or ignores) the last row and last column and continues with the residual 4×4 matrix. The bottom 2×2 matrix of this is $\left(\begin{smallmatrix} -3 \cdot 690 & 1 \cdot 961 \\ 0 \cdot 390 & 3 \cdot 633 \end{smallmatrix}\right)$ and from this we take the eigenvalue $3 \cdot 7356$, this being the closer to element $3 \cdot 633$. Continuing in this way, the staring Hessenberg matrix is deflated until it is only the 2×2 $\left(\begin{smallmatrix} -8 \cdot 703 & 1 \cdot 440 \\ 5 \cdot 6E_{-5} & 6 \cdot 003 \end{smallmatrix}\right)$ with eigenvalues $-8 \cdot 703099780845$ and $6 \cdot 002887274905$, correct to 12 decimal places.

The shift and deflate strategies lead to rapid convergence. The convergence criterion was that the change between iterations be less than 10^{-8}. 6 iterations were required for the bottom row's eigenvalue, 6 for the next, 5 for the next, then 2, and the last two are solve for analytically.

It is easy to find the eigenvectors of the final Hessenberg matrix by back substitution, and I have obtained these. However, they are of limited use in themselves. In order to relate them to the eigenvectors of the staring Hessenberg matrix and hence to the original given matrix, it would be necessary to trace back through the sequence of rotation matrices $\mathbf{Q}$ for all the iteration for all eigenvalues. Francis does not discuss obtaining the eigenvectors. I have calculated them by the traditional direct method, outlined in §1, of substituting the eigenvalue into the original matrix and solving the resulting set of simultaneous equations for the components of the eigenvector. For example, if the eigenvalue $-8 \cdot 7031$ is subtracted from the diagonal of $\mathbf{C}$, we obtain the following row-equivalent triangular matrix by using elementary row operations, much as in LU decomposition:

$$\begin{pmatrix} 13 \cdot 703 & 1 & 2 & -3 & 1 \\ 3 & 9 \cdot 703 & -4 & 5 & 2 \\ 2 & -3 & 7 \cdot 703 & 4 & -1 \\ -2 & 1 & 2 & 4 \cdot 703 & 1 \\ 1 & 2 & 2 & 7 & 6 \cdot 703 \end{pmatrix} \quad \rightarrow \quad \begin{pmatrix} 13 \cdot 703 & 1 & 2 & -3 & 1 \\ 0 & 9 \cdot 484 & -4 \cdot 438 & 5 \cdot 657 & 1 \cdot 781 \\ 0 & 0 & 5 \cdot 939 & 6 \cdot 314 & -0 \cdot 555 \\ 0 & 0 & 0 & 0 \cdot 575 & 1 \cdot 195 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

When multiplied by the eigenvector, this equals the zero vector. We may therefore take the bottom component of the eigenvector to be 1 and obtain the others by back substitution. The result is

$$\begin{pmatrix} -1 \cdot 01957 \\ 2 \cdot 12960 \\ 2 \cdot 30319 \\ -2 \cdot 07844 \\ 1 \end{pmatrix}.$$

This article has concentrated on finding the real eigen pairs of real matrices, touching on complex eigenvalues of real matrices only lightly. However, it would be missing a great deal if I were not to outline John Francis' second algorithm, previously called the 'implicit double shift QR method' which solves real matrices for real and complex pairs of eigenvalues. It has been described by Prof. Nick Higham as the 'jewel in the crown' of eigenvalue algorithms.

10.3 Dealing with complex conjugate eigenvalues

If a matrix has entirely real elements, its eigenvalues must all be real or else involve one or more pairs of complex conjugates. If n is odd, an $n \times n$ matrix must have at least one real eigenvalue. The high rate of convergence in Francis' the single shift version is attained by subtracting, at each iteration, the latest best estimate of the eigenvalue from the Hessenberg matrix's diagonal. Clearly, if the eigenvalue is complex, a complex shift would be required and we have to conduct the algorithm in complex numbers. In fact Francis found a clever way of avoiding complex arithmetic, using his double shift algorithm as described in the next subsection. At this stage, however, let us suppose that, for whatever reason, we are interested only in the real eigenvalues. Can the single shift algorithm still be used to finds all of these while perhaps ignoring the complex conjugate pairs? I offer here my own thoughts on this, with two numerical examples.

At each iteration the diagonal shifts have been chosen by solving the quadratic equations which yield the eigenvalues of the 2×2 submatrices down the diagonal of the Hessenberg matrix. For an $n \times n$ matrix there will be $n-1$ such matrices, each with two eigenvalues. Call these eigenvalues $r_j,\ j = 1 \to 2(n-1)$. Suppose we calculate all these – not a demanding task – and sort the real values by absolute size, ignoring the complex ones. It is likely that the largest will be rough approximation to the largest real eigenvalue of the whole matrix, and possibly similarly for the smallest. Either of these can be used as a trial shift to see if the iterations converge. I have written a computer program to see how this works in practice. The program uses the following procedure to select the diagonal shift at each iteration:

- for the first iteration for each new eigenvalue, the shift is the largest (in absolute value) of the real r_j,
- for subsequent iterations the shift is set to the nearest (in absolute value) of the current set of r_j to the last shift used,
- a limit on the number of iterations is set to a small number, say 10. If there has not been convergence to an eigenvalue by then, another real r_j may be chosen as starting point and we try again
- if it appears that a real eigenvalue has been found, the latest Hessenberg matrix is checked to confirm that it has only one non-zero element, in the last row, last column. More than one non-zero element is incompatible with a real eigenvalue.

In limited trials I have had some success with this rough and ready approach. The first example is the 6×6 matrix

$$\mathbf{M} = \begin{pmatrix} -5 & 1 & 2 & -3 & 1 & 3 \\ 3 & 1 & -4 & 5 & 2 & -4 \\ 2 & -3 & -1 & -4 & -1 & 0 \\ 2 & 1 & 2 & 0 & 4 & 1 \\ 1 & 2 & 2 & -7 & -2 & 5 \\ 4 & -3 & -3 & 0 & 7 & 1 \end{pmatrix}. \tag{40}$$

This has only two real eigenvalues $-9\cdot97116$ and $-4\cdot41896$. The complex conjugate pairs are $4\cdot128837\pm i0\cdot251512$ and $0\cdot066222\pm i4\cdot057590$. To three decimal places the Hessenberg form is

$$\begin{pmatrix} -5 & 2\cdot401 & -2\cdot570 & -1\cdot083 & 0\cdot886 & 3\cdot110 \\ 5\cdot831 & -1\cdot147 & 5\cdot747 & -2\cdot942 & 2\cdot333 & 3\cdot063 \\ 0 & 5\cdot068 & -2\cdot993 & -1\cdot578 & 1\cdot671 & 2\cdot225 \\ 0 & 0 & 8\cdot801 & -2\cdot018 & 0\cdot383 & -6\cdot286 \\ 0 & 0 & 0 & 3\cdot539 & 1\cdot044 & -3\cdot243 \\ 0 & 0 & 0 & 0 & -0\cdot224 & 4\cdot114 \end{pmatrix}.$$

The r_j values for successive 2×2 submatrices down the diagonal are

$$\begin{array}{cc} 1\cdot13495, & -7\cdot28201 \\ 3\cdot40492, & -7\cdot54547 \\ \text{Complex conjugates} & \textit{ignored} \\ 1\cdot43637, & -2\cdot41013 \\ 4\cdot33509, & 0\cdot82332. \end{array}$$

The largest absolute one $-7\cdot54547$ and this is taken as the shift for the first iteration. At the start of the second iteration new r_j values are found from successive 2×2 submatrices down the diagonal of the new Hessenberg matrix. They are

$$\begin{array}{cc} 4.07483, & -5.36422 \\ \text{Complex conjugates} & \textit{ignored} \\ 1\cdot02110, & -5\cdot08356 \\ 1\cdot17397, & -6\cdot50568 \\ 4\cdot12759, & -4\cdot77574 \end{array}$$

The nearest of these to the previous shift is $-6\cdot50568$ so this is the new shift. It happens also to be the current largest. However at the start of iteration 3 the nearest is no longer the largest; the r_j are

$$\begin{array}{cc} \text{Complex conjugates} & \textit{ignored} \\ 2\cdot03813, & -5\cdot94380 \\ 1\cdot16675, & -8\cdot24092 \\ \text{Complex conjugates} & \textit{ignored} \\ 4\cdot03213, & -4\cdot91037 \end{array}$$

The shift used here is $-5\cdot94380$. At iterations 4 to 7 the shifts are $-4\cdot66575$, $-4\cdot419081$, $-4\cdot41895887$ and $-4\cdot418958763$ and here the process has converged on the eigenvalue $-4\cdot4189587629587$ correct to 13 decimal places. The Hessenberg matrix at this stage is

$$\begin{pmatrix} 3\cdot231 & 2\cdot309 & 3\cdot614 & 5\cdot911 & -0\cdot921 & -1\cdot023 \\ 0\cdot850 & -2\cdot788 & -2\cdot943 & -5\cdot532 & 3\cdot166 & 3\cdot109 \\ 0 & 5\cdot103 & 0\cdot243 & -0\cdot549 & -6\cdot382 & 0\cdot814 \\ 0 & 0 & 2\cdot647 & -1\cdot859 & 6\cdot121 & -1\cdot388 \\ 0 & 0 & 0 & 8\cdot789 & -0\cdot408 & 0\cdot536 \\ 0 & 0 & 0 & 0 & 0 & -4\cdot4189... \end{pmatrix}$$

whereupon it is deflated by ignoring the last row and column. Notice how the eigenvalue has been pushed to the bottom. The search for the next eigenvalue is in the 5×5 matrix and starts again with the largest r_j which is $-8\cdot504...$. At successive iterations the nearest r_j happens in this case to

be the largest and at the end of iteration 5 there is convergence on the eigenvalue $-9\cdot971160$. The matrix is deflated to 4×4:

$$\begin{pmatrix} 3\cdot943 & 1\cdot312 & -2\cdot357 & 3\cdot995 \\ 0\cdot310 & 3\cdot952 & -1\cdot885 & 5\cdot458 \\ 0 & 8\cdot022 & 1\cdot456 & -0\cdot930 \\ 0 & 0 & 2\cdot006 & -0\cdot961 \end{pmatrix}$$

and this in fact has only complex eigenvalues. Of course the algorithm does not 'know' this so evaluates the r_j values. Only the first (top left) 2×2 submatrix has real eigenvalues, so $r_1 = 4{\cdot}5851795$ and $r_2 = 3\cdot30976$. The larger of these would be applied as the shift. At iteration 2 the r_j values are

$$\begin{array}{cc} 4\cdot70729 & 0\cdot05877 \\ \text{Complex conjugates} & \textit{ignored} \\ 4\cdot5851795 & -0\cdot96113 \end{array}$$

and it appears that it has converged on $4\cdot5851795$ as an eigenvalue since this is not just near to the previous shift but identical. However, it is easily seen that this would be a false conclusion because the Hessenberg matrix at this juncture is

$$\begin{pmatrix} 3\cdot3098 & 1\cdot3040 & 0\cdot0782 & 1\cdot5836 \\ 3\cdot4841 & 1\cdot4563 & -6\cdot368 & -3\cdot5398 \\ 0 & 3\cdot3810 & -0\cdot5446 & 6\cdot9581 \\ 0 & 0 & 0\cdot3071 & 4\cdot1686 \end{pmatrix}$$

that is, it has two non-zero entries in the bottom row, not just one. This is inconsistent with a real eigenvalue so the process should stop and conclude, perhaps tentatively, that there are only two real eigenvalues – the two already found. Another clue comes simply in observing the difference between successive iterations to see whether they are converging or varying in a seemingly random manner.

My second example is the 7×7 matrix

$$\begin{pmatrix} -5 & 1 & 2 & -3 & 1 & 3 & -4 \\ 3 & 1 & -4 & 5 & 2 & -4 & 0 \\ 2 & -3 & -1 & -4 & -1 & 0 & 8 \\ 2 & 1 & 2 & 0 & 4 & 1 & -9 \\ 1 & 2 & 2 & -7 & -2 & 5 & -1 \\ 4 & -3 & -3 & 0 & 7 & 10 & 1 \\ 0 & 1 & 4 & -3 & 6 & -2 & 1 \end{pmatrix}.$$

This has 5 real eigenvalues and one conjugate pair. The algorithm works smoothly and solves for all real eigenvalues in the order listed below. I have listed also the shift applied at the first iteration for that eigenvalue, and the number of iterations to converge within the 10^{-8} criterion.

1st shift	iterations	eigenvalue
$10\cdot22453$	5	$9\cdot735439$
$12\cdot58435$	3	$12\cdot578527$
$3\cdot31636$	3	$3\cdot318528$
$-6\cdot15014$	5	$-6\cdot557620$
$-3\cdot53107$	4	$-2\cdot932474$

Note how good the initial shifts are! The final 2×2 matrix has only complex eigenvalues, and as a bonus the program finds these to be $-6\cdot0712 \pm i\,5\cdot8022$.

10.4 Francis' double shift algorithm for complex eigenvalues

Francis was working in the early days of computers when memory and storage imposed severe constraints. He wanted to find a way of dealing with complex conjugate eigenvalues without resorting to complex arithmetic. He therefore exploited the fact that a shift by one complex conjugate and then by the other would result in the equal and opposite imaginary components cancelling each other, leaving a purely real number. The double shift algorithm, therefore, involves shifting by the two complex conjugates in succession with the arithmetic worked out as if this were a single iteration. The numbers remain real throughout.

We need to prove that carrying out two single shift iterations in succession is equivalent to factorising the product of two shifted version of the same matrix. Suppose that $\mathbf{H_0}$ is the Hessenberg form of the given matrix $\mathbf{A}$. I will write out the shift explicitly. As a point 4 at the opening of §10, $\mathbf{H_0} - \beta_0\mathbf{I}$ is first factorised into $\mathbf{Q_0}\,\mathbf{R_0}$, then $\mathbf{H_1} - \beta_0\mathbf{I}$ found by reverse multiplying. We thus have

$$\mathbf{H_0} - \beta_0\mathbf{I} \;=\; \mathbf{Q_0}\,\mathbf{R_0}\,, \qquad \mathbf{H_1} - \beta_0\mathbf{I} \;=\; \mathbf{R_0}\,\mathbf{Q_0}\,, \tag{41a}$$

$$\mathbf{H_1} - \beta_1\mathbf{I} \;=\; \mathbf{Q_1}\,\mathbf{R_1}\,, \qquad \mathbf{H_2} - \beta_1\mathbf{I} \;=\; \mathbf{R_1}\,\mathbf{Q_1}\,. \tag{41b}$$

These are the two shifts of the double shift algorithm. Now consider what matrix would have the factorisation $(\mathbf{Q_0}\,\mathbf{Q_1})(\mathbf{R_1}\,\mathbf{R_0})$? The answer is found by substituting from Eq 41a, b:

$$\begin{aligned}(\mathbf{Q_0}\,\mathbf{Q_1})(\mathbf{R_1}\,\mathbf{R_0}) \;&=\; \mathbf{Q_0}(\mathbf{Q_1}\,\mathbf{R_1})\mathbf{R_0} \;=\; \mathbf{Q_0}(\mathbf{H_1} - \beta_1\mathbf{I})\mathbf{R_0}\\ &= \mathbf{Q_0}(\mathbf{R_0}\,\mathbf{Q_0} + \beta_0\mathbf{I} - \beta_1\mathbf{I})\mathbf{R_0}\\ &= (\mathbf{Q_0}\,\mathbf{R_0})\,(\mathbf{Q_0}\,\mathbf{R_0}) \;+\; (\beta_0 - \beta_1)\mathbf{Q_0}\,\mathbf{R_0}\\ &= (\mathbf{H_0} - \beta_0\mathbf{I})^2 \;+\; (\beta_0 - \beta_1)(\mathbf{H_0} - \beta_0\mathbf{I})\\ &= \mathbf{H_0}^2 - 2\beta_0\mathbf{H_0} + \beta_0^2\mathbf{I} + \beta_0\mathbf{H_0} - \beta_1\mathbf{H_0} - \beta_0^2\mathbf{I} + \beta_0\beta_1\mathbf{I}\\ &= (\mathbf{H_0} - \beta_0\mathbf{I})(\mathbf{H_0} - \beta_1\mathbf{I})\,.\end{aligned} \tag{41c}$$

Thus the product matrix $\mathbf{G} = (\mathbf{H_0} - \beta_0\mathbf{I})(\mathbf{H_0} - \beta_1\mathbf{I})$ will factorise into the orthogonal matrix $\mathbf{Q_0}\,\mathbf{Q_1}$ and the upper triangular matrix $\mathbf{R_1}\,\mathbf{R_0}$ (since the product of two triangular matrices is itself triangular). So two shifted iterations can be done as one by operating on the product $\mathbf{G}$. Moreover, if β_0 and β_1 are complex conjugates, $\beta_r \pm i\beta_i$,

$$\mathbf{G} \;\rightarrow\; \mathbf{H_0}^2 - 2\beta_r\mathbf{H_0} + (\beta_r^2 + \beta_i^2)\mathbf{I} \tag{42}$$

which is entirely real. This is just what Francis was looking for – the problem solvable using only real arithmetic.

Again the properties of Hessenberg matrices and their manipulation by Givens rotators and Householder reflectors save us from having to calculate $\mathbf{G}$ explicitly before factorising it. As with the single shift algorithm, we operate column by column in a process known in the trade as 'bulge chasing'. The details, which are quite involved, are given in the Part 2 paper by Francis and explained in the books by David Watkins referenced in §1. I refer the interested reader to these.

John Coffey, August 2016

11 Appendix 1 : Solving the sum of several geometric series

11.1 How the sum arises in the Power Method

This Appendix extends §3.1 and §3.2 on the basic power method. Sequences of differences between successive iterates of the power method (obtained by multiplying by the given matrix $mathbfE$) have the structure of the sum of several geometric series. At Eq 8 we saw the general term of the most dominant series whose common ratio is $r = \lambda_2/\lambda_1$. Eq 9 shows how series can be summed to infinity and so give a much closer estimate of the λ_1 and its eigenvector $\mathbf{p}_1$. Here I evaluate the next few terms in the difference $\delta_{m+1} = \mathbf{v}^{(m+1)} - \mathbf{v}^{(m)}$ and show that each of these is the general term of a further geometric series. Each series is readily summed to infinity by the formula '$a/(1-r)$' that we learned at school.

The starting point is the approximate eigenvector given at Eqs 2 and 3. I simplify the problem by assuming that only three eigenvalues contribute; that is the eigenvector estimate at the m^{th} iteration is

$$\mathbf{v}_m \;\approx\; (\mathbf{p}_1 + Cr^m\mathbf{p}_2 + Ds^m\mathbf{p}_3)(1 - Cr^m - Ds^m + \ldots)\,. \qquad (A1.1)$$

and neglect higher terms. I have also simplified the notation from §3.1 as follows

$$\mathbf{v}_m \equiv \mathbf{v}^{(m)}\,, \quad C = \frac{c_2}{c_1}\,, \quad D = \frac{c_3}{c_1}\,, \quad r = \frac{\lambda_2}{\lambda_1}\,, \quad s = \frac{\lambda_3}{\lambda_1}\,, \quad \Sigma_m \equiv Cr^m + Ds^m\,.$$

Eq A1.1 follows after the Taylor expansion has been taken of the denominator by which $\mathbf{v}_m$ is normalised. The difference between iterations $\delta_m = \mathbf{v}_{m+1} - \mathbf{v}_m$ is

$$Cr^m(1-r)(\mathbf{p_1} - \mathbf{p_2}) \;+\; Ds^m(1-s)(\mathbf{p_1} - \mathbf{p_3})$$
$$+\; C^2r^{2m}(1-r^2)\mathbf{p_2} \;+\; CDr^ms^m(1-rs)(\mathbf{p_2} + \mathbf{p_3}) \;+\; D^2s^{2m}(1-s^2)\mathbf{p_3}\,. \qquad (A1.2)$$

The geometric series given at Eq 9 in §3.2 is the first of these terms, with common ratio r. Clearly there are four other series with ratios s, r^2, rs and s^2. Moreover, the Taylor series expansion of the denominator $(1 + Cr^m + Ds^m)^{-1}$ was carried only to the first order of small quantities. Taking it to second order will add terms in C^2r^{2m} *etc.* arising from $(Cr^m + Ds^m)^2$. With many matrices the series will decrease is numerical significance in the order listed above so we might judge it worthwhile retaining only the first two, in Cr and Ds, or perhaps including the next term in C^2r^{2m}. I ignore terms in CDr^ms^m and D^2s^2. The sums to infinity of the three leading series are respectively

$$C(\mathbf{p_1} - \mathbf{p_2})\,, \qquad D(\mathbf{p_1} - \mathbf{p_3})\,, \qquad C^2\mathbf{p_2}\,, \qquad (A1.3)$$

the $1-r$, $1-s$, $1-r^2$ factors in Eq A1.2 cancelling. The vectors $\mathbf{p}_1$, $\mathbf{p}_2$ are treated as sets of separate vector components each with its own fitted geometric series. In the numerical procedure the total of these sums to infinity will be added to the first iterate to give projected values of $\mathbf{p_1}$ and λ_1. Bear in mind that a running set of values $\mathbf{v}_m$ is used, consisting of the most recent run of 5 or 6 power method iterates.

11.2 Multi-variable Newton's method

The aim is to recover the constant terms and common ratio of each geometric series given a sequence of 5 or 6 consecutive values of δ for a chosen component of the vector. This can be done using Newton's method provided the starting values are sufficiently close. To keep the notation uncluttered write for any chosen vector component

$$C(1-r)(p_1 - p_2) = \mathcal{C}, \qquad D(1-s)(p_1 - p_3) = \mathcal{D}, \qquad C^2(1-r^2)p_2 = \mathcal{E}\,. \qquad (A1.4)$$

The same will apply to the eigenvalue because it is essentially just the bottom/last component of the eigenvector before it is normalised to 1. The problem is to find $\mathcal{C}$, $\mathcal{D}$, $\mathcal{E}$, r and s given successive values of δ_j which satisfy the equations

$$\begin{aligned}
\delta_2 &\equiv \lambda^{(2)} - \lambda^{(1)} &&= \mathcal{C} + \mathcal{D} + \mathcal{E}\,, \\
\delta_3 &\equiv \lambda^{(3)} - \lambda^{(2)} &&= \mathcal{C}r + \mathcal{D}s + \mathcal{E}r^2\,, \\
\delta_4 &\equiv \lambda^{(4)} - \lambda^{(3)} &&= \mathcal{C}r^2 + \mathcal{D}s^2 + \mathcal{E}r^4\,, \\
\delta_5 &\equiv \lambda^{(5)} - \lambda^{(4)} &&= \mathcal{C}r^3 + \mathcal{D}s^3 + \mathcal{E}r^6\,, \\
\delta_6 &\equiv \lambda^{(6)} - \lambda^{(5)} &&= \mathcal{C}r^4 + \mathcal{D}s^4 + \mathcal{E}r^8\,.
\end{aligned} \qquad (A1.5)$$

These are the triple series formulae. In the double series $\mathcal{E} = 0$ and δ_6 is not needed.

We are familiar from school maths with Newton's formula for finding a better approximation to the root of an equation $f(x) = 0$ in the single variable x given a starting estimate:

$$x_{m+1} \;=\; x_m - \frac{1}{f'(x_m)} f(x_m)\,. \qquad (A1.6)$$

Where there are N variables, there must be at least N independent equations for a solution, so both x and f are replaced by N-vectors $\mathbf{x}$ and $\mathbf{f}$. In $\mathbf{f}$ each element is an expression to be set to zero. The derivative $f'(x)$ is replaced by the matrix of partial derivatives with respect to the variables – the so-called Jacobian matrix, $\mathbf{J}$. The reciprocal $1/f'$ is replaced by the matrix inverse $\mathbf{J}^{-1}$. The generalised Newton's method is therefore

$$\mathbf{x}_{m+1} \;=\; \mathbf{x}_m - \mathbf{J}_m^{-1}\,\mathbf{f}_m\,. \qquad (A1.7)$$

11.3 Sum of two geometric series

Solving one geometric series is trivial and certain – the ratio is common to every pair of consecutive terms. We might expect that solving the sum of two series will require a fairly good initial guess, and that solving the sum of three series will be even more sensitive to the starting conditions. The eigenvalue ratios ensure that after a few iterations $|\mathcal{C}| > |\mathcal{D}|$ since $|r| > |s|$. A starting value for r can be taken from δ_5/δ_4 or, better, δ_6/δ_5 if δ_6 has been evaluated.

With $\mathcal{D}$ set to $\delta_2 - \mathcal{C}$ there are only three variables. The matrices $\mathbf{x}$ and $\mathbf{f}$, and the Jacobian and its inverse, become

$$\mathbf{x} \;=\; \begin{pmatrix} \mathcal{C} \\ r \\ s \end{pmatrix}, \qquad \mathbf{f} \;=\; \begin{pmatrix} r\,\mathcal{C} + (\delta_2 - \mathcal{C})s - \delta_3 \\ r^2\mathcal{C} + (\delta_2 - \mathcal{C})s^2 - \delta_4 \\ r^3\mathcal{C} + (\delta_2 - \mathcal{C})s^3 - \delta_5 \end{pmatrix}. \qquad \mathbf{J} \;=\; \begin{pmatrix} r - s & \mathcal{C} & \delta_2 - \mathcal{C} \\ r^2 - s^2 & 2\,\mathcal{C}r & 2\,(\delta_2 - \mathcal{C})s \\ r^3 - s^3 & 3\,\mathcal{C}r^2 & 3\,(\delta_2 - \mathcal{C})s^2 \end{pmatrix},$$

$$\mathbf{J}^{-1} \;=\; \frac{1}{(r-s)^3} \begin{pmatrix} -6rs & 3(r+s) & -2 \\ \frac{(2r+s)(r-s)s}{\mathcal{C}} & -\frac{(r+2s)(r-s)}{\mathcal{C}} & \frac{r-s}{\mathcal{C}} \\ \frac{(r+2s)(r-s)r}{\delta_2-\mathcal{C}} & -\frac{(2r+s)(r-s)}{\delta_2-\mathcal{C}} & \frac{r-s}{\delta_2-\mathcal{C}} \end{pmatrix}. \qquad (A1.8)$$

Here is a demonstration that the method works for one made-up example series. Table 5 lists the first terms, common ratios and first few terms. The first approximation is $\mathcal{C} < 13$, say 12,

first	ratio	2	3	4	5	6	7	8
10	0·7	10	7	4·90	3·430	2·4010	1·6807 0	1·176490
3	0·4	3	1·2	0·48	0·192	0·0768	0·03072	0·012288
	sum	13	8·2	5·38	3·622	2·4778	1·71142	1·188778

Table 7: Test sum of two geometric series. The top line gives the indices j of δ_j.

and $r \approx 8 \cdot 2/12 = 0 \cdot 68$. Take s to be something smaller, say $0 \cdot 3$. $\delta_2 = 13$, $\delta_3 = 8 \cdot 2$, etc. from Table 2. $\mathbf{f}$ is $(0 \cdot 26, 0 \cdot 2588, 0 \cdot 1782)$ and the iterated matrix $\mathbf{x}$ develops as follows

$$\begin{pmatrix} 12 \\ 0 \cdot 68 \\ 0 \cdot 3 \end{pmatrix}, \quad \begin{pmatrix} 10 \cdot 428 \\ 0 \cdot 694 \\ 0 \cdot 474 \end{pmatrix}, \quad \begin{pmatrix} 11 \cdot 2236 \\ 0 \cdot 6905 \\ 0 \cdot 3206 \end{pmatrix}, \quad \begin{pmatrix} 10 \cdot 3355 \\ 0 \cdot 6958 \\ 0 \cdot 4047 \end{pmatrix}, \quad \begin{pmatrix} 9 \cdot 99614 \\ 0 \cdot 69996 \\ 0 \cdot 39947 \end{pmatrix}.$$

This is very close the the exact values of $\mathcal{C} = 10$, $r = 0 \cdot 7$ and $s = 0 \cdot 4$. The second starting value $\mathcal{D}$ is found from $13 - 9 \cdot 996$. In §9.4 below I present a more precise scheme for obtaining initial values.

To check sensitivity to starting values I evaluated the following sequence

$$\begin{pmatrix} 6 \\ 0 \cdot 9 \\ 0 \cdot 5 \end{pmatrix}, \quad \begin{pmatrix} 5 \cdot 656 \\ 0 \cdot 801 \\ 0 \cdot 505 \end{pmatrix}, \quad \begin{pmatrix} 8 \cdot 688 \\ 0 \cdot 697 \\ 0 \cdot 457 \end{pmatrix}, \quad \begin{pmatrix} 10 \cdot 5925 \\ 0 \cdot 6947 \\ 0 \cdot 3962 \end{pmatrix}, \quad \begin{pmatrix} 9 \cdot 99618 \\ 0 \cdot 69987 \\ 0 \cdot 40122 \end{pmatrix}.$$

This is not computationally involved and the results are encouraging.

11.4 Sum of three series

Moving to three geometric series, the computational effort is in inverting the 4×4 matrix. However, with symbolic algebra software this is readily obtained and the formulae can be cut and pasted into computer code. My limited experience with three series is that the biggest practical difficulty is in obtaining a sufficiently close starting estimate for Newton's method to converge. Any computer program should continue with triple series only if there is convergence. Note that this is not the most general 3-series case because the ratio of the third series is r^2.

We need a vector $\mathbf{f}$ with 4 expressions plus $\mathcal{E} = \delta_2 - \mathcal{C} - \mathcal{D}$.

$$\mathbf{x} = \begin{pmatrix} \mathcal{C} \\ \mathcal{D} \\ r \\ s \end{pmatrix}, \qquad \mathbf{f} = \begin{pmatrix} s\,\mathcal{D} + r^2\,(\delta_2 - \mathcal{C} - \mathcal{D}) + r\,\mathcal{C} - \delta_3 \\ s^2\,\mathcal{D} + r^4\,(\delta_2 - \mathcal{C} - \mathcal{D}) + r^2\,\mathcal{C} - \delta_4 \\ s^3\,\mathcal{D} + r^6\,(\delta_2 - \mathcal{C} - \mathcal{D}) + r^3\,\mathcal{C} - \delta_5 \\ s^4\,\mathcal{D} + r^8\,(\delta_2 - \mathcal{C} - \mathcal{D}) + r^4\,\mathcal{C} - \delta_6 \end{pmatrix}.$$

$$\mathbf{J} = \begin{pmatrix} r - r^2 & s - r^2 & 2\,r\,(\delta_2 - \mathcal{C} - \mathcal{D}) + \mathcal{C} & \mathcal{D} \\ r^2 - r^4 & s^2 - r^4 & 4\,r^3\,(\delta_2 - \mathcal{C} - \mathcal{D}) + 2\,r\,\mathcal{C} & 2\,s\,\mathcal{D} \\ r^3 - r^6 & s^3 - r^6 & 6\,r^5\,(\delta_2 - \mathcal{C} - \mathcal{D}) + 3\,r^2\,\mathcal{C} & 3\,s^2\,\mathcal{D} \\ r^4 - r^8 & s^4 - r^8 & 8\,r^7\,(\delta_2 - \mathcal{C} - \mathcal{D}) + 4\,r^3\,\mathcal{C} & 4\,s^3\,\mathcal{D} \end{pmatrix}. \qquad (A1.9)$$

The terms in r^6, s^6, r^8, s^8 should make us concerned that the route to a solution in likely to be sensitive to the input values.

As expected, $\mathbf{J}^{-1}$ is complicated. Since we want formula into which numbers will be substituted, I give the inverse of a general 4-by-4 matrix which can be matched to $\mathbf{J}$:

$$\mathbf{M} = \begin{pmatrix} t_1 & t_2 & t_3 & t_4 \\ u_1 & u_2 & u_3 & u_4 \\ v_1 & v_2 & v_3 & v_4 \\ w_1 & w_2 & w_3 & w_4 \end{pmatrix}.$$

The determinant Δ is obtained from the four 3×3 cofactor submatrices:

$$\begin{aligned} \Delta = \quad & t_1 \left[u_2 \left(v_3 w_4 - v_4 w_3 \right) - u_3 \left(v_2 w_4 - v_4 w_2 \right) + u_4 \left(v_2 w_3 - v_3 w_2 \right) \right] \\ & -t_2 \left[u_1 \left(v_3\, w_4 - v_4\, w_3 \right) - u_3 \left(v_1\, w_4 - v_4\, w_1 \right) + u_4 \left(v_1\, w_3 - v_3\, w_1 \right) \right] \\ & +t_3 \left[u_1 \left(v_2\, w_4 - v_4\, w_2 \right) - u_2 \left(v_1\, w_4 - v_4\, w_1 \right) + u_4 \left(v_1\, w_2 - v_2\, w_1 \right) \right] \\ & -t_4 \left[u_1 \left(v_2\, w_3 - v_3\, w_2 \right) - u_2 \left(v_1\, w_3 - v_3\, w_1 \right) + u_3 \left(v_1\, w_2 - v_2\, w_1 \right) \right]. \end{aligned}$$

The notation can be condensed by observing that each 2×2 determinant is identified by its first product. For example $(v_3 w_4 - v_4 w_3)$ can be abbreviated to $\{v_3 w_4\}$ without confusion. Using this shorthand the determinant is

$$\begin{aligned} \Delta = \quad & t_1 \left[u_2 \left\{ v_3 w_4 \right\} - u_3 \left\{ v_2 w_4 \right\} + u_4 \left\{ v_2 w_3 \right\} \right] \\ & -t_2 \left[u_1 \left\{ v_3\, w_4 \right\} - u_3 \left\{ v_1\, w_4 \right\} + u_4 \left\{ v_1\, w_3 \right\} \right] \\ & +t_3 \left[u_1 \left\{ v_2\, w_4 \right\} - u_2 \left\{ v_1\, w_4 \right\} + u_4 \left\{ v_1\, w_2 \right\} \right] \\ & -t_4 \left[u_1 \left\{ v_2\, w_3 \right\} - u_2 \left\{ v_1\, w_3 \right\} + u_3 \left\{ v_1\, w_2 \right\} \right]. \end{aligned}$$

In the inverse of $\mathbf{J}$ each element is the sum of three terms. $\mathbf{J} = 1/\Delta$ multiplied by the following matrix. Columns 1 and 2 are

$$\begin{array}{ll} \{u_2\, v_3\}\, w_4 + \{u_4\, v_2\}\, w_3 + \{u_3\, v_4\}\, w_2 & \{t_3\, v_2\}\, w_4 + \{t_2\, v_4\}\, w_3 + \{t_4\, v_3\}\, w_2 \\ \{u_3\, v_1\}\, w_4 + \{u_1\, v_4\}\, w_3 + \{u_4\, v_3\}\, w_1 & \{t_1\, v_3\}\, w_4 + \{t_4\, v_1\}\, w_3 + \{t_3\, v_4\}\, w_1 \\ \{u_1\, v_2\}\, w_4 + \{u_4\, v_1\}\, w_2 + \{u_2\, v_4\}\, w_1 & \{t_2\, v_1\}\, w_4 + \{t_1\, v_4\}\, w_2 + \{t_4\, v_2\}\, w_1 \\ \{u_2\, v_1\}\, w_3 + \{u_1\, v_3\}\, w_2 + \{u_3\, v_2\}\, w_1 & \{t_1\, v_2\}\, w_3 + \{t_3\, v_1\}\, w_2 + \{t_2\, v_3\}\, w_1 \end{array}$$

and columns 3 and 4 are

$$\begin{array}{ll} \{t_2\, u_3\}\, w_4 + \{t_4\, u_2\}\, w_3 + \{t_3\, u_4\}\, w_2 & \{t_3\, u_2\}\, v_4 + \{t_2\, u_4\}\, v_3 + \{t_4\, u_3\}\, v_2 \\ \{t_3\, u_1\}\, w_4 + \{t_1\, u_4\}\, w_3 + \{t_4\, u_3\}\, w_1 & \{t_1\, u_3\}\, v_4 + \{t_4\, u_1\}\, v_3 + \{t_3\, u_4\}\, v_1 \\ \{t_1\, u_2\}\, w_4 + \{t_4\, u_1\}\, w_2 + \{t_2\, u_4\}\, w_1 & \{t_2\, u_1\}\, v_4 + \{t_1\, u_4\}\, v_2 + \{t_4\, u_2\}\, v_1 \\ \{t_2\, u_1\}\, w_3 + \{t_1\, u_3\}\, w_2 + \{t_3\, u_2\}\, w_1 & \{t_1\, u_2\}\, v_3 + \{t_3\, u_1\}\, v_2 + \{t_2\, u_3\}\, v_1\,. \end{array}$$

A test case is listed in Table 6. What guidance is there as to starting values? The ratios of consecutive terms are $0 \cdot 656$, $0 \cdot 72862$, $0 \cdot 76$, $0 \cdot 78$, trending towards about $0 \cdot 8$. Suppose the guess is $r = 0 \cdot 79$. If the ratio δ_3/δ_2 is expanded as a Taylor series in $\mathcal{D}$ to linear terms the result is

$$\frac{\delta_3}{\delta_2} \equiv d_{32} \approx r - (r - s)\frac{\mathcal{D}}{\mathcal{C}}.$$

$$\text{Similarly} \quad \frac{\delta_4}{\delta_3} \equiv d_{43} \approx r - (r - s)\frac{\mathcal{D}s}{\mathcal{C}r}.$$

Solving these simultaneously gives

$$\frac{\mathcal{D}}{\mathcal{C}} \approx \frac{(r - d_{32})^2}{r(d_{43} - d_{32})}, \qquad s \approx \frac{r(r - d_{43})}{r - d_{32}}. \tag{A1.10}$$

		2	3	4	5	6	7	8
first	ratio							
10	0·8	10	8	6·4	5·12	4·096	3·2768	2·62144
4	0·3	4	1·2	0·36	0·108	0·0324	0·00972	0·00292
1	0·64	1	0·64	0·4096	0·26214	0·16777	0·10737	0·06872
	sum	15	9·84	7·1696	5·490144	4·29617216	3·39389	2·69308

Table 8: Sum of three geometric series with first terms 10, 4, 1 and ratios $0\cdot8$, $0\cdot3$, $0\cdot8^2$.

With $d_{32} = 0\cdot656$, $d_{43} = 0\cdot72862$, $r = 0\cdot79$ we find $\mathcal{D} \approx 0\cdot313\mathcal{C}$ and $s \approx 0\cdot362$. We have to guess a value for $\mathcal{E}$, suspecting it to be smaller than $\mathcal{D}$. Taking $\mathcal{E} = 1\cdot5$, $\mathcal{C} \approx (15 - 1\cdot5)/1\cdot313 = 10\cdot28$ and $\mathcal{D} \approx 3\cdot22$. These are our starting values. The iterations give

$$\begin{pmatrix}10\cdot28\\3\cdot22\\0\cdot79\\0\cdot362\end{pmatrix},\quad \begin{pmatrix}8\cdot704\\3\cdot115\\0\cdot810\\0\cdot238\end{pmatrix},\quad \begin{pmatrix}9\cdot326\\3\cdot692\\0\cdot8056\\0\cdot2917\end{pmatrix},\quad \begin{pmatrix}10\cdot047\\3\cdot981\\0\cdot7995\\0\cdot2999\end{pmatrix},\quad \begin{pmatrix}10\cdot00001\\3\cdot99996\\0\cdot80000\\0\cdot30000\end{pmatrix}.$$

We can be relieved that it has converged to the correct values.

However – and it is a big 'however '– the process can vary wildly for only slightly different starting values. I has found this sequence

$$\begin{pmatrix}11\\4\cdot5\\0\cdot7\\0\cdot35\end{pmatrix},\quad \begin{pmatrix}-0\cdot372\\-10\cdot99\\0\cdot92\\0\cdot076\end{pmatrix},\quad \begin{pmatrix}21\cdot04\\2\cdot82\\0\cdot866\\0\cdot044\end{pmatrix},\quad \begin{pmatrix}-1\cdot37\\2\cdot388\\0\cdot888\\0\cdot153\end{pmatrix},\quad \begin{pmatrix}4\cdot11\\3\cdot216\\0\cdot861\\0\cdot274\end{pmatrix},$$

$$\begin{pmatrix}-0\cdot598\\4\cdot138\\0\cdot896\\0\cdot309\end{pmatrix},\quad \begin{pmatrix}0\cdot853\\4\cdot133\\0\cdot885\\0\cdot304\end{pmatrix},\quad \begin{pmatrix}0\cdot889\\4\cdot117\\0\cdot884\\0\cdot303\end{pmatrix},\quad \begin{pmatrix}0\cdot8861\\4\cdot1173\\0\cdot8838\\0\cdot3033\end{pmatrix},\quad \begin{pmatrix}0\cdot88609\\4\cdot11735\\0\cdot88379\\0\cdot30330\end{pmatrix}.$$

After thrashing around for a few iterations it has settled down and converged, but to a different solution! It has $\mathcal{E} = 9\cdot996557$ and indeed is a solution over the first five terms only

$$15,\quad 9\cdot84,\quad 7\cdot1696,\quad 5\cdot490144,\quad 4\cdot29617216,\quad 3\cdot3945\,,\quad 2.6954$$

to compare with the bottom line in Table 3. This non-uniqueness is not acceptable though it makes only a small numerical difference to the sum to infinity – $59\cdot2$ instead of $58\cdot5$. For this reason and because of the greater complexity of the three-series problem, I judge that it is not to be pursued except in special circumstances. In contrast, on the basis of the little evidence above, the two-series solution seems reliable. I have therefore used it to enhance convergence of the Power Method.

Had the three-series solution worked more readily and reliably, use could have been made of it to estimate the second eigenvector $\mathbf{p}_2$ which would be very useful as a starting point for any iterative process to find it precisely. The basis of this claim is in the simultaneous solution of the sums to infinity at Eq A1.3. Calling these Σ_C, Σ_D, Σ_E respectively,

$$C \equiv \frac{c_2}{c_1} \;=\; \frac{1}{2p_1}\left(\Sigma_C + \sqrt{4p_1\Sigma_E + \Sigma_C^2}\;\right),$$

$$p_2 \;=\; \frac{1}{2\Sigma_E}\left(\Sigma_C^2 \;+\; 2p_1\Sigma_E \;-\; \Sigma_C\sqrt{4p_1\Sigma_E + \Sigma_C^2}\;\right). \qquad (A1.11)$$

p_1 is the component of the eigenvector already obtained. In Tables 3 and 4 of §3.1 I give an example of this to find the second most dominant eigenvalue and eigenvector.

12 Appendix 2 : A straightforward matrix by the Power Method

This Appendix takes a 5×5 matrix with well spaced real eigenvalues and solves it using the direct and inverse power methods. The aim is to illustrate combined use of these methods, together with the signs of the pivots of a row-equivalent triangular matrix, in what should be a straightforward problem.

I invented a matrix with eigenvalues near $9 \cdot 3$, $7 \cdot 6$, $3 \cdot 1$, $0 \cdot 4$ and $-1 \cdot 7$ by starting within a diagonal matrix which had precisely these eigenvalues, then carrying out a similarity transformation with an invertible square matrix whose elements were from a random number generator. Finally I rounded the resulting matrix to 3 places of decimal. Calling this $\mathbf{G}$,

$$\mathbf{G} \;=\; \begin{pmatrix} 8 \cdot 711 & 0 \cdot 229 & 1 \cdot 223 & -1 \cdot 839 & 3 \cdot 102 \\ 15 \cdot 772 & 4 \cdot 614 & 15 \cdot 704 & -0 \cdot 257 & -39 \cdot 179 \\ -4 \cdot 356 & 0 \cdot 036 & 1 \cdot 501 & 1 \cdot 933 & -3 \cdot 754 \\ 15 \cdot 458 & 3 \cdot 824 & 19 \cdot 314 & 3 \cdot 104 & -36 \cdot 508 \\ 1 \cdot 568 & -0 \cdot 112 & 1 \cdot 092 & -0 \cdot 322 & 0 \cdot 770 \end{pmatrix} .$$

The trace is $18 \cdot 7$ and the precise eigenvalues, to compare with the eventual iterated estimates, are

$$9 \cdot 30257881867, \quad 7 \cdot 60429497945, \quad 3 \cdot 0965570746, \quad 0 \cdot 400321504951, \quad -1 \cdot 70375237767.$$

I have found it useful to obtain a rough picture of where the eigenvalues lie by running a preliminary computer program which I wrote to shift the diagonal, convert $\mathbf{G}$ to a row-equivalent triangular matrix and count its positive and negative pivots. This is a fairly quick operation, so I have used a sequence of diagonal shifts at increments of 1 from -2 to $+10$. The results are best shown by marks on the number line:

$$-2 \;..*..-1.....0..*..1.....2.....3..*..4.....5.....6.....7..*..8.....9..*..10$$

Each * denotes that an eigenvalue lies in that interval. In other words, the eigenvalues are within $\pm 0 \cdot 5$ of $-1 \cdot 5$, $0 \cdot 5$, $3 \cdot 5$, $7 \cdot 5$ and $9 \cdot 5$.

We will use the direct power method to find λ_1 near $9 \cdot 5$, choosing a shift of diagonal which will enhance the ratio to the next largest eigenvector. The midpoint of $7 \cdot 5$ and $-1 \cdot 5$ is 3, so try a shift of 3. The three starting vectors were $(1, 1, 0, 0, 1)$, $(-1, 0, 0, 1, 1)$ and $(1, -2, 1, 0, 1)$ and from these the best was chosen after three iterations. Convergence was disappointingly slow. The error criterion, being the required change from one iteration to the next, was $< 10^{-8}$, and this had not been met by the single geometric series projection after 33 iterations and, moreover, the algorithm to fit two geometric series to the differences was failing to converge. I therefore increased the number of allowed iterations by 5 and changed the shift from 3 to 2. The double geometric series now converged and the required precision was achieved at iteration 37, giving $\lambda_1 = 7 \cdot 302578828 + 2 \cdot 0$ with error $< 7 \times 10^{-9}$. The eigenvector is similarly precise, though I will not quote it to save cluttering this Appendix with long numbers. The single geometric series' projected value at iteration 37 was $9 \cdot 30257884$, and the value from the basic power method multiplication was the disappointing $9 \cdot 30266$.

The ratio r of the double series is $0 \cdot 76741$ which means that the next largest eigenvalue is $5 \cdot 6041 + 2 \cdot 0$, in remarkable agreement with the actual value given above. The ratio s predicts the third largest eigenvalue to be $-1 \cdot 13$. In view of the slow convergence I did not attempt the triple geometric series projection, knowing that it is even more sensitive to the starting values provided.

It may be that the shift of 2 proved better than 3 because it favoured one eigenvalue ($7\cdot3$) over the other ($-1\cdot7$); in contrast, a shift of 3 places these two eigenvalues equally distance from the shifted λ_1 and this perhaps 'confuses' the power method.

The strategy I have followed has been now to find the most negative eigenvalue because it can be made the largest in absolute value by a suitable shift. Since the trace is $18\cdot7$ and the most negative λ is about $-1\cdot5$, the average value of the others is about 5. However, with a shift $\beta = 5$ the required precision is not achieved after 38 iterations and the double series algorithm has consistently failed to converge. Suspecting that this may be another case of the power method becoming 'confused' by two eigenvalues almost equally distant from the currently largest, I adjusted the shift to 6. It then completed at iteration 29 with the two-series projection $\lambda = -7\cdot703752384 + 6$, $\varepsilon < 5\cdot4\times10^{-9}$. The ratio r implies another eigenvalue at $-5\cdot59947 + 6 = 0\cdot40053$ and another at $3\cdot5609 + 6$, this latter having already been found as λ_1.

At this stage we have the most positive and most negative eigenvalue and vectors to 8 places of decimal, plus close estimates of two others at $7\cdot6041$ and $0\cdot40053$. Since the trace is $18\cdot7$, the final eigenvalue must be close to $3\cdot09655$, as indeed it is. The best way to refine these three values and to find their eigenvectors is to use the inverse power method with shifts equal to these values. I do not think it worthwhile using the strict Rayleigh quotient formula, Eq 18, because of the need to make LU decompositions at each cycle. With $7\cdot6041$ subtracted from the diagonal the LU decomposition I used was

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 14\cdot249 & 1 & 0 & 0 & 0 \\ -3\cdot9353 & -0\cdot1499 & 1 & 0 & 0 \\ 13\cdot965 & -0\cdot1001 & -1\cdot3319 & 1 & 0 \\ 1\cdot4166 & 0\cdot0698 & 0\cdot3360 & 0\cdot0433 & 1 \end{pmatrix} \begin{pmatrix} 1\cdot1069 & 0\cdot229 & 1\cdot223 & -1\cdot839 & 3\cdot102 \\ 0 & -6\cdot2531 & -1\cdot7223 & 25\cdot947 & -83\cdot379 \\ 0 & 0 & -1\cdot5483 & -1\cdot4153 & -4\cdot0432 \\ 0 & 0 & 0 & 21\cdot894 & -93\cdot560 \\ 0 & 0 & 0 & 0 & -0\cdot0004 \end{pmatrix}.$$

Observe that in the upper triangular matrix three pivots are negative, but the bottom one is almost zero, meaning that one eigenvalue is almost zero, consistent with the shift of $7\cdot6041$ being a close estimate.

By iteration 5 the eigenvalue estimates changed in only the 7th decimal place to $5128\cdot7455055 = 1/0\cdot0001949794543$, making $\lambda = 7\cdot60429497945$, correct in all digits. The eigenvector was simultaneously found. The same inverse process can be applied to refine the two remaining eigenvalues and find their vectors. Apart from the slow convergence when finding the most positive and most negative eigenvalues, the solution of this matrix has been straightforward.

13 Appendix 3 : A more difficult matrix

In order to test some of the algorithms described in the main text, I have contrived a 6×6 matrix with some eigenvalues that are close together, since this condition is known almost to challenge both the Power and QR methods. The matrix has been created by starting with a diagonal matrix $\mathbf{D}$ with prescribed eigenvalues as diagonal elements:

$$7 \cdot 02, \qquad 6 \cdot 9, \quad 4 \cdot 02, \quad 3 \cdot 95, \quad 1 \cdot 0 \quad 0 \cdot 05 \,.$$

This was transformed into a general non-symmetric matrix $\mathbf{A}$ by a similarity transformation $\mathbf{SDS}^{-1}$ using a random square matrix $\mathbf{S}$. The values were truncated at four decimal places, so the eigenvalues are not exactly the values above, but close. The test matrix is

$$\mathbf{A} \;=\; \begin{pmatrix} 4 \cdot 5414 & -1 \cdot 6042 & -2 \cdot 5242 & 0 \cdot 8774 & -0 \cdot 4240 & -0 \cdot 7963 \\ 11 \cdot 1351 & -3 \cdot 1788 & -9 \cdot 6235 & 3 \cdot 7092 & -1 \cdot 3859 & -0 \cdot 5861 \\ -17 \cdot 6652 & 7 \cdot 7166 & 13 \cdot 0719 & -3 \cdot 7933 & 1 \cdot 0398 & -0 \cdot 2998 \\ -24 \cdot 0686 & 10 \cdot 8624 & 12 \cdot 5539 & -1 \cdot 2481 & 1 \cdot 3851 & 0 \cdot 5853 \\ 11 \cdot 6915 & -11 \cdot 2681 & -5 \cdot 5112 & 2 \cdot 6536 & 3 \cdot 7486 & -1 \cdot 2444 \\ -7 \cdot 2994 & 1 \cdot 1925 & 1 \cdot 0205 & -0 \cdot 5831 & -0 \cdot 3855 & 6 \cdot 1250 \end{pmatrix} .$$

13.1 Preliminary assessment

We pretend we have no idea of the eigenvalues. It is not necessary to apply the bound estimates on the eigenvalues of items 14, 15 16 of §2, but they may be a guide to a strategy for solution to get a rough idea of where the eigenvalues lie. The trace of $\mathbf{A}$ is $23 \cdot 06$ and of $\mathbf{E}^2$ is $129 \cdot 868$. The Wolkowicz and Styan parameters (§2, item 16) are $m = 3 \cdot 843333$ and $s = 2 \cdot 621727$, and their criteria place the lowest and highest eigenvalues in these rather generous intervals

$$-2 \cdot 02 < \lambda_6 < 2 \cdot 67 \,, \qquad 5 \cdot 02 < \lambda_1 < 9 \cdot 71 \,.$$

Another way of getting bounds is to transform to a row-equivalent triangular matrix $\mathbf{T}$ and count the signs of the diagonal elements. Diagonal shifts by subtracting $\beta\mathbf{I}$ can be used to move the boundary between positive and negative pivots. With $\beta = 0$ there are 6 positive pivots, meaning that all six eigenvalues are positive.

13.2 QR-Schur iteration

I have applied the QR algorithm with shifts of 0 to 2·5 in steps of 0·5. At each iteration the computer program also evaluated the eigenvalues of all 2×2 submatrices on adjacent rows down the diagonal. I looked for evidence of geometric series in these 2×2 eigenvalues and also in the the diagonal elements of the iterated RQ matrices.

The behaviour with zero diagonal shift is not dissimilar to that with shifts of 1 or 2. This is $\mathbf{RQ}$ after iteration 30, to 4 places of decimal:

$$\begin{pmatrix} 6 \cdot 9589 & -0 \cdot 0442 & -1 \cdot 1689 & 3 \cdot 3572 & -13 \cdot 2610 & -24 \cdot 0762 \\ -0 \cdot 0815 & 6 \cdot 9610 & -2 \cdot 5133 & 1 \cdot 6856 & 4 \cdot 2697 & 7 \cdot 8984 \\ 4 \cdot 12_{E-7} & -5 \cdot 75_{E-6} & 3 \cdot 9635 & -0 \cdot 0118 & 11 \cdot 4650 & 30 \cdot 3903 \\ -2 \cdot 22_{E-8} & -2 \cdot 87_{E-7} & -0 \cdot 0633 & 4 \cdot 0067 & 4 \cdot 5184 & 5 \cdot 6211 \\ 1 \cdot 09_{E-24} & -2 \cdot 98_{E-23} & -1 \cdot 47_{E-17} & 7 \cdot 09_{E-18} & 1 \cdot 1000 & 2 \cdot 496 \\ -1 \cdot 31_{E-60} & -1 \cdot 87_{E-59} & -1 \cdot 12_{E-53} & 4 \cdot 28_{E-54} & -6 \cdot 67_{E-37} & 0 \cdot 0699 \end{pmatrix} .$$

The below-diagonal elements show that there is full convergence to the last two eigenvalues, $1 \cdot 1000354966611$ and 0.069933443993549, which do not change even in the fifteenth decimal place at the next iteration. However the elements in positions (2, 1) and (4, 3) are not vanishingly small, meaning that the upper four diagonal elements cannot be taken as precise estimates of the larger four eigenvalues. Rather, the eigenvalues of the 2×2 matrix on the diagonal in rows 1, 2 and columns 1, 2 are found to be $7 \cdot 01997$ and 6.89994, which are λ_1 and λ_2 correct to 5 decimal places. The equivalent matrix in rows 3, 4 and columns 3, 4 has eigenvalues $4 \cdot 01986$, $3 \cdot 95026$ which are λ_3 and λ_4. Clearly here is another case where the eigenvalues of the diagonal 2×2 submatrices are a much stronger estimate of the eigenvalues of the whole matrix than the bare diagonal elements themselves.

It is not necessary to go as far as iteration 30 to obtain useful estimates of the λ_j, $j = 1, 6$, especially if the plan is to use all the values from QR as input to the Inverse Power Method. The computer program took each 2×2 submatrix down the diagonal of **RQ** and solved its characteristic equation by the usual quadratic formula to give two eigenvalue estimates. There are therefore five pairs of estimates at each iteration. These are plotted in the left panel of Figure 4 which is for zero diagonal shift. The same colours are used for the same pair, with the higher value plotted as a small circle and the lower as a triangle. The plot makes it clear that there are six eigenvalues, two pairs of which are close together, and the least is near zero. The colour key in Figure 4 continues into the right panel, which shows the differences between successive iterates on a logarithmic scale. A straight line here indicates convergence as a geometric series. Two curves (blue circles, green triangles) have clearly not settled to straight lines, but the other eight are almost straight apart from their last point, which may be due to rounding errors (burgundy circles, yellow triangles). We hope for geometric series on six estimates which should produce usefully accurate eigenvalues.

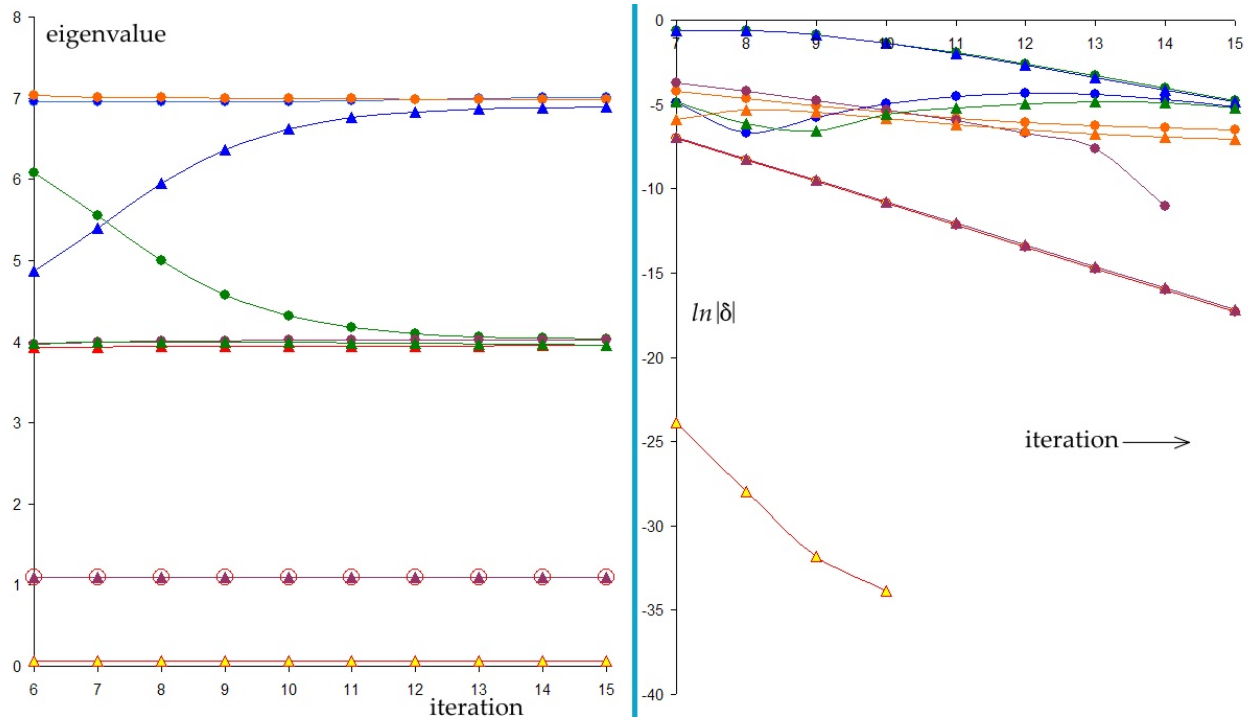

Figure 7: Convergence of eigenvalue estimates from 2×2 diagonal submatrices from iterations 6 to 15. Left: λ_j estimates,. Right: Logarithm of differences $\delta_j = \lambda_j - \lambda_{j-1}$. No diagonal shift of matrix.

The lowest line in the right panel is for the smallest eigenvalue, λ_6. It clearly converges very rapidly, and indeed the last point does not lie on the straight line probably because of rounding errors. The ratio of differences at (iteration 8/iteration 7) is $0 \cdot 0171$ and consequent sum of the geometric series is $7 \cdot 5 \times 10^{-13}$. This gives $\lambda_6 = 0 \cdot 069933443993547$ which is in error by only 2 in the last decimal place. High precision is also attained with the two estimates of λ_5 (burgundy coloured triangles and open circles). These each lie convincingly on a straight line. At iteration 8 the values of λ_5 projected from the corresponding geometric series are $1 \cdot 10003554$ and $1 \cdot 10003551$, to compare with the true value $1 \cdot 10003549666$. By iteration 12 one estimate has become correct to 9 decimal places and the other to 10.

The other plots in the right panel have lower gradient and some curvature which means that the common ratio in the fitted geometric series is large. The two orange curves in particular are converging slowly. Eigenvalues from these solutions of the 2×2 characteristic equations are going to be less precise and less confident. Table 10 lists the values I obtain using the sections of plot in Figure 4, right panel, which appear locally most like a straight line. For reference, the precise eigenvalues are given alongside. These estimates are not good enough in themselves, but the eigenvalues have all been separated and the estimates are good enough to input to the Inverse Power Method, which will also provide the eigenvectors.

blue circles	iteration 14, 15	$\lambda_1 \approx 7 \cdot 02$	$7 \cdot 01997321188480$
blue triangles	iterations 12, 13, 14, 15	$\lambda_2 \approx 6 \cdot 899$	$6 \cdot 89994138219623$
orange circles	iterations 13, 14, 15	$\lambda_2 \approx 6 \cdot 97$	$6 \cdot 89994138219623$
burgundy circles	iterations 12, 13, 14, 15	$\lambda_3 \approx 4 \cdot 020$	$4 \cdot 01985644547121$
green circles	iterations 12, 13, 14, 15	$\lambda_3 \approx 4 \cdot 022$	$4 \cdot 01985644547121$
green triangles	iterations 14, 15	$\lambda_4 \approx 3 \cdot 9$	$3 \cdot 95026001979312$
orange triangles	iterations 13, 14, 15	$\lambda_2 \approx 3 \cdot 955$	$3 \cdot 95026001979312$

Table 9: Eigenvalue estimates from characteristic polynomials of 2×2 diagonal submatrices, estimates by fitting geometric series at selected iterations.

To complete the story, here is the inverse power method with a shift of $6 \cdot 92$, this being an average of the two λ_2 estimates in Table 10, weighted roughly according to their precision. I chose as starting vector $(1, -1, 1, 0, -1, 1)$ and used LU decomposition as in §5.1. I kept the shift constant at $6 \cdot 92$ since to do otherwise (as in Rayleigh iteration) involves a new LU decomposition at each stage. It takes to iteration 10 for the change in the eigenvectors to be in the fifth decimal place, and iteration 13 in the seventh. At iteration 16 the changes in the all the vector components are less than 5×10^{-10}. The (shifted inverse) eigenvalue is $-49 \cdot 8538838 = 1/-0 \cdot 020058618$ which when shifted back is $6 \cdot 8999413822$, correct to 10 decimal places. The eigenvector is

$$\begin{pmatrix} 0 \cdot 599783708 \\ 2 \cdot 999185988 \\ -2 \cdot 6992173670 \\ -2 \cdot 999024143 \\ -6 \cdot 698556985 \\ 1 \end{pmatrix}.$$

13.3 Francis' single shifted QR algorithm

This powerful algorithm uses the convergence-enhancing devices of diagonal shifting optimised at each iteration, deflation and QR - RQ decomposition, so has in-built all the features described

above. The first step is conversion Hessenberg form, the result, written here to 3 decimal places, being

$$\begin{pmatrix} 4\cdot541 & 0\cdot186 & 0\cdot434 & -2\cdot552 & -0\cdot792 & -1\cdot779 \\ 34\cdot718 & 1\cdot923 & -0\cdot849 & -23\cdot318 & -7\cdot991 & -10\cdot568 \\ 0 & 1\cdot994 & 5\cdot188 & 5\cdot040 & -1\cdot210 & -1\cdot04 \\ 0 & 0 & 0\cdot393 & 0\cdot039 & -1\cdot712 & -2\cdot180 \\ 0 & 0 & 0 & 1\cdot164 & 7\cdot458 & 0\cdot769 \\ 0 & 0 & 0 & 0 & -1\cdot173 & 3\cdot910 \end{pmatrix}.$$

The search is firstly for the eigenvalue associated with the sixth row. Initially the diagonal shift is 0, and for subsequent shifts I used the eigenvalue of the bottom diagonal 2×2 matrix which is closest to the bottom right matrix element. These converge to the eigenvalue of the total matrix in the sequence over five iterations, as follows:

$$0,\quad -0\cdot05166,\quad 0\cdot062667,\quad 0\cdot0698927,\quad 0\cdot06993344457,$$

and at this stage the eigenvalue $0\cdot0699334439935$ has been found to better than 10^{-8}. The matrix is now deflated by ignoring the last row and last column, and the search for the eigenvalue associated with the new bottom row starts using a shift of $1\cdot00839$, this being the smaller eigenvalue of the new bottom diagonal 2×2 matrix.

Numbering the eigenvalues by their row, number 6 is found in 5 iterations, number 5 in 4, number 4 in 4, and number 3 in 2 iterations whereupon the last two eigenvalues are found by solving algebraically the residual 2×2 matrix. So only 15 iterations have been needed. A check on the calculated eigenvalues shows that they add to the trace of $23\cdot06$ with a discrepancy of only 3×10^{-16}. No wonder this algorithm has been so highly praised and so widely used.