Abstract. When using an iterative method for solving a generalized nonsymmetric eigenvalue problem of the form $Fu = \lambda Mu$, where F and M are real banded matrices, it is often desirable to work with the shifted and inverted operator $B = (F - \sigma M)^{-1}M$ in order to enhance the eigenvalue separation and improve efficiency. Unfortunately, the shift σ is generally complex and so is the matrix B. The question then is whether it is possible to avoid complex arithmetic while preserving any advantages of bandedness of the pair (F, M). For the classical problem where $M = I$ and F is banded, complex arithmetic can be avoided by using double shifts, i.e., by working with the real matrix $B\bar{B}$ whose bandwidth is double that of F. This satisfactory solution extends to the case where M is diagonal as well. In the generalized case the answer to the above question is negative, in the sense that complex arithmetic can be avoided only at the expense of loosing the advantage of bandedness. One solution is to factor the shifted matrix $F - \sigma M$ in complex arithmetic but employ real arithmetic subsequently in the iterative procedure. This paper examines several approaches and discusses their respective merits under various circumstances.

Complex Shift and Invert Strategies for Real Matrices

Beresford N. Parlett* and Youcef Saad

Research Report YALEU/DCS/RR-424
July 1986

* University of California at Berkeley, Mathematics Departement.

Revised version. The work of B.N. Parlett was supported in part by the Office of Naval Research under contract N00014-76-C-0013. The work of Y. Saad was supported in part by the Department of Energy under contract DE-AC02-81ER10996 and in part by the Army Research Office under contract DAAG-83-0177.

1. Introduction

This communication is concerned with the eigenvalue problem: solve

$$(F - \lambda M)z = 0, \tag{1.1}$$

for $\lambda \in \mathbf{C}$ and $z \in \mathbf{C}^N$, when F, M are real banded N x N matrices. Throughout the discussion, M will be symmetric and positive definite. In a number of applications, $M = I$, the identity matrix. There will be no restriction on F except that the pair (F, M) have complex eigenvalues. More precisely, we suppose that only a few of the eigenvalues λ are wanted, namely those in the vicinity of a given complex number σ. What makes this task challenging is

(I) the desire to keep computation in $\mathbf{R}$ rather than in $\mathbf{C}$;

(II) the desire to exploit any narrow band structure enjoyed by F and M.

These two desires can be in conflict. We shall assume for convenience that F and M have the same bandwidth $2\beta + 1$; the (i, j) elements vanish whenever $|i - j| > \beta$. Moreover, we shall assume that the band is narrow, i.e., that $\beta << N$. Note that the goal (II) can be generalized into that of exploiting any particular sparse structure, not just bandedness.

Every reasonable approach known to us requires an iterative process at each stage of which a system of equations must be solved. The simplest of these is

$$(F - \sigma M)y = Mb \tag{1.2}$$

where b is given and y is to be computed. Our task reduces to an attempt to reconcile the two aims (I) and (II), when solving the above system, or rival ones similar to it. Notice that a more general equation than (1.2) is $(F - \sigma M)y = (\gamma F - \delta M)b$, but as can be readily seen the eigenvalues of the corresponding transformed eigenvalue problem are related to those of (1.2) in a linear way and therefore this adds no generality.

In the body of this paper we present all the alternatives that have occured to us and analyze them. In particular, we show a surprising connection between two of them. Unfortunately, our analysis leads to no "best" method, but we give operation counts and storage costs for the better techniques.

Before proceeding with the algebra, we should say something about complex arithmetic. We can imagine an arithmetic engine that would employ 4 real arithmetic processors in parallel to compute the product of two complex numbers in almost the same time as required for a real multiplication. We know of no such computer at present. In some systems we know of, (VAX 780), the ratio complex:real arithmetic is nearly four but, in others, the cost of accessing the arguments has become sufficiently large to reduce the ratio to nearly two. The storage penalty remains at 2:1.

2. Inverse Iteration

Throughout the paper we write the shift σ as

$$\sigma = \rho + i\ \theta,$$

with $\theta > 0$ and $i^2 = -1$. In our context, inverse iteration is defined as follows, [9]:

1. Choose $x^{(1)} \neq 0$.
2. For $k = 1, 2, \ldots$, until convergence do
   2.1. Solve
   $$(F - \sigma M)y^{(k)} = Mx^{(k)}, \tag{2.1}$$
   2.2. Normalize: $x^{(k+1)} = y^{(k)}/\xi^{(k)}$, where $\xi^{(k)}$ is the component of $y^{(k)}$ of largest modulus.
   2.3. Check for convergence.

In the generic case the sequence $\{x^{(k)}\}$ converges to z where $Fz = \lambda Mz$ and λ is the eigenvalue closest to σ. One can approximate λ by $\sigma + x^{(k)}(j)/y^{(k)}(j)$ where $y^{(k)}(j)$ is an above average element of $y^{(k)}$. A better but more expensive approximation is the Rayleigh Quotient,

$$\rho(x^{(k)}) = \frac{(Fx^{(k)}, x^{(k)})}{(Mx^{(k)}, x^{(k)})}.$$

Of course we may seek several eigenvalues close to σ, not just one. Consequently, more elaborate iterations are needed. Examples are Arnoldi [1, 8], Lanczos [6, 2], or simultaneous iteration [5] (also known as subspace iteration). The differences between these methods are not important here because they may all be used with the same operator, namely

$$B = (F - \sigma M)^{-1}M. \tag{2.2}$$

Note that the sequence $\{x^{(k)}\}$ may be thought of as generated by multiplying each term by B and then normalizing in order to get the next one. The matrix B is not formed explicitly. The dominant part of each step in any of the iterative methods is the solution step 2.1 of the algorithm. One way to carry this out is

Method 1

Compute the triangular factorization,

$$F - \sigma M = LU,$$

once and for all (in complex arithmetic). Then system (2.1) is solved by the two triangular solves (in complex arithmetic):

$$Lw^{(k)} = Mx^{(k)}, \quad Uy^{(k)} = w^{(k)}.$$

Recalling the two goals in the introdution, we see that by abandoning (I), real arithmetic, we can exploit (II), band structure. The costs of Method 1 are as follows.

- Arithmetic. Factorization : $\beta^2 N$ complex multiplications. Forward and backward solutions: $2(\beta + 1)N$ complex multiplications. Normalization: N comparisons and N complex multiplications in step 2.2.
- Storage. $2(2\beta+1)N$ real locations for F and M; $(2\beta+1)N$ complex locations for $F-\sigma M = LU$; plus two complex vectors for storing $x^{(k)}$ and $y^{(k)}$.

Now consider the implementation of step 2.1 in real arithmetic. We write $y = y_r + i\ y_i$, for any vector $y \in \mathbf{C}^N$. In the standard way we equate real and imaginary parts to get

$$(F - \rho M)y_r + \theta M y_i = M x_r \tag{2.3}$$

$$-\theta M y_r + (F - \rho M)y_i = M x_i \tag{2.4}$$

or, in matrix form

$$\begin{pmatrix} F - \rho M & \theta M \\ -\theta M & F - \rho M \end{pmatrix} \begin{pmatrix} y_r \\ y_i \end{pmatrix} = \begin{pmatrix} M & O \\ O & M \end{pmatrix} \begin{pmatrix} x_r \\ x_i \end{pmatrix}. \tag{2.5}$$

This system of order $2N$ has lost the band structure of (2.1). It achieves (I) at the expense of (II). For future reference it is important to observe that the iteration associated with (2.5) attempts to compute the eigenpairs of the $(2N)$ x $(2N)$ real matrix

$$\begin{pmatrix} M & O \\ O & M \end{pmatrix}^{-1} \begin{pmatrix} F - \rho M & \theta M \\ -\theta M & F - \rho M \end{pmatrix} = \begin{pmatrix} M^{-1}F - \rho I & \theta I \\ -\theta I & M^{-1}F - \rho I \end{pmatrix}.$$

The expression of the inverse of the above matrix is of great help when establishing relationships between the various approaches taken in later sections. Letting

$$A = M^{-1}F$$

we have

$$\begin{pmatrix} A - \rho I & \theta I \\ -\theta I & A - \rho I \end{pmatrix}^{-1} = \begin{pmatrix} X^{-1} & -\theta Y^{-1} \\ \theta Y^{-1} & X^{-1} \end{pmatrix}, \tag{2.6}$$

in which,

$$X = (A - \rho I) + \theta^2 (A - \rho I)^{-1}, \quad Y = X(A - \rho I) = (A - \rho I)^2 + \theta^2 I \tag{2.7}$$

In particular it can be seen from above that y_r, y_i can be obtained by solving for y_r first, via

$$\left[(F - \rho M) + \theta^2 M (F - \rho M)^{-1} M\right] y_r = M x_r - \theta M (F - \rho M)^{-1} M x_i \tag{2.8}$$

and then getting y_i by substitution in the equation (2.3),

$$\theta M y_i = M x_r - (F - \rho M) y_r,$$

which gives

$$y_i = \frac{1}{\theta} \left[x_r - (M^{-1}F - \rho I) y_r\right]. \tag{2.9}$$

Note that one can also compute y_i first from an equation similar to (2.8) and then substitute in (2.4) to get y_r.

When $M = I$ a simplification is possible, by multiplying both sides of (2.8) by $(F - \rho I)$:

Method 2 (For the case $M = I$) .

To solve the system (2.1) in inverse iteration algorithm, compute the real part y_r of $y^{(k)}$ by

$$\left[(F - \rho I)^2 + \theta^2 I\right] y_r = (F - \rho I) x_r - \theta x_i, \tag{2.10}$$

and its imaginary part y_i by (2.9).

The bandwidth has been doubled but not ruined. The matrix F^2 may be stored once and for all, allowing for changes in σ. The costs of this method are as follows.

- Arithmetic. Factorization of $(F-\rho I)^2+\theta^2 I$ (done once) : $4\beta^2 N$ real multiplications. Forward and backward solutions: $8(\beta+1)N$ real multiplications. Normalization: $2N$ comparisons and $2N$ real multiplications in step 2.2.
- Storage. $(4\beta+1)N$ real locations for $(F-\rho I)^2+\theta^2 I$, and $(4\beta+1)N$ real locations for its LU factorization. Four real vectors for storing $x^{(k)}$ and $y^{(k)}$.

3. The general case.

Consider again (2.8). If M is diagonal then Method 2 extends readily and we will not consider this further. We now take up the general case when M has the same band structure as F. Premultiply (2.8) by $(F-\rho M)M^{-1}$ to get the analogue of (2.10):

$$\left[(F-\rho M)M^{-1}(F-\rho M)+\theta^2 M\right] y_r = (F-\rho M)x_r - \theta M x_i \tag{3.1}$$

We will define

$$G \equiv (F-\rho M)M^{-1}(F-\rho M)+\theta^2 M = FM^{-1}F - 2\rho F + |\sigma|^2 M. \tag{3.2}$$

Unless M is diagonal, the presence of M^{-1} ruins the bandedness of the matrix G which will be full in general. Note that

$$M^{-1}G = Y \tag{3.3}$$

where Y is defined in (2.7).

Nevertheless, band structure may still be exploited, especially in a number of applications where F and M are generated by the finite element method. In those situations it is common to replace the consistant mass matrix M by a diagonal lumped mass matrix D. The diagonal elements of D are the elements of the vector Me where $e=(1,1,\ldots,1)^T$. The real matrix

$$FD^{-1}F - 2\rho F + |\sigma|^2 M$$

has twice the bandwith of F and M and may be factored efficiently into LU. This matrix is used as a preconditioner for the proper matrix in a conjugate gradient type method [3]. The inner (conjugate gradient) iteration should converge in a very small number of steps. The simplest alternative to conjugate gradients is the following iteration on the residuals to solve the system $Gy_r = r \equiv (F-\rho M)x_r - \theta M x_i$

Method 3

Until convergence do:

(i) Solve $LUd = r$

(ii) compute $r \leftarrow r - Gd, \quad y_r \leftarrow y_r + d.$

(iii) If $\|r\|$ too large then repeat (i) and (ii).

(iv) else get y_i by (2.9) and return.

We omit the details on the costs of the above method, because the process is iterative and is not comparable to previous techniques.

4. The double shift approach

A problem similar to ours, but in the context of the QR algorithm, was solved by J.G.F. Francis in 1961/62 [4]. If $A \in \mathbf{R}^{N,N}$ and $\sigma \in \mathbf{C}$ then $Y \equiv (A-\sigma I)(A-\bar{\sigma} I) = [(A-\rho I)^2 + \theta^2 I] \in \mathbf{R}^{N,N}$. This matrix, which is real, is a quadratic polynomial in A and shares A's eigenvectors. When A is replaced by $M^{-1}F$ then this matrix coincides with the matrix Y defined by (2.7). By (3.3), the eigenvalues of Y are those of the generalized eigenvalue problem $Gz = \nu Mz$ or:

$$\left[FM^{-1}F - 2\rho F + |\sigma|^2 M\right] z = \nu M z \tag{4.1}$$

where $\nu = \lambda^2 - 2\rho\lambda + |\sigma|^2$. Unless M is diagonal, or block diagonal, the matrix G is real but full. Even the actual formation of G is unatractive in practice.

In the present context, inverse iteration is defined as in section 2 except that the system (2.1) is replaced by

$$Gy^{(k)} = Mx^{(k)}. \tag{4.2}$$

We can solve the above system iteratively as indicated in the previous section. However, there are alternative approaches that fully exploit band structure, provided we relax our constraint on working entirely in real arithmetic. Let LU be the (complex) factorization of $(F-\sigma M)$. The matrices L and U will inherit the band structure of F and M. In order to solve $Gy = Mx$ where $y \in \mathbf{R}^N$, $x \in \mathbf{R}^N$, note that

$$Gy = LUM^{-1}\bar{L}\bar{U}y = Mx.$$

An algorithm for computing y is

Method 4

(i) Solve $La = Mx$ for $a \in \mathbf{C}^N$
(ii) Solve $Ub = a$ for $b \in \mathbf{C}^N$
(iii) Form $c = Mb$
(iv) Solve $\bar{L}d = c$ for $d \in \mathbf{C}^N$
(v) Solve $\bar{U}e = d$ for $e \in \mathbf{C}^N$
(vi) Set $y = Re\,(e)$.

The complex arithmetic is hidden in the above subroutine that maps x into y. The iteration that is used to compute one and two dimensional eigenspaces of Y can confine itself to real arithmetic. We shall have more to say about the matrix Y in the next section. Now we resume the quest for an operator that requires no complex arithmetic and yet takes advantage of narrow bandwidth.

5. Real and Imaginary Part approaches

Inverse iteration with shift σ is equivalent to direct iteration (i.e., the power method) using the operator $(F-\sigma M)^{-1}M$ on $\mathbf{C}^N$. To obtain related operators on $\mathbf{R}^N$ we can take the real and imaginary parts

$$B_+ = \frac{1}{2}\left[(F-\sigma M)^{-1} + (F-\bar{\sigma}M)^{-1}\right]M = Re\left[(F-\sigma M)^{-1}M\right], \tag{5.1}$$

$$B_- = \frac{1}{2i}\left[(F-\sigma M)^{-1} - (F-\bar{\sigma}M)^{-1}\right]M = Im\left[(F-\sigma M)^{-1}M\right]. \tag{5.2}$$

If $Fz = \lambda Mz$ then

$$B_+ z = \frac{1}{2}\left[\frac{1}{\lambda-\sigma} + \frac{1}{\lambda-\bar{\sigma}}\right] z \equiv \mu_+ z \quad , \quad B_- z = \frac{1}{2i}\left[\frac{1}{\lambda-\sigma} - \frac{1}{\lambda-\bar{\sigma}}\right] z \equiv \mu_- z, \tag{5.3}$$

defining μ_+ and μ_- the eigenvalues of B_+ and B_- associated with the eigenvector z of A. It is readily verified that as $\lambda \to \sigma$,

$$\mu_+ \approx \frac{1}{2(\lambda-\sigma)}, \quad \mu_- \approx \frac{1}{2i(\lambda-\sigma)}.$$

Thus B_+ and B_- give the same enhancement to eigenvalues close to σ. In contrast, as $\lambda \to \infty$, B_- dampens the eigenvalues more strongly than does B_+ since,

$$\mu_+ = \frac{\lambda-\rho}{(\lambda-\sigma)(\lambda-\bar{\sigma})}, \quad \mu_- = \frac{\theta}{(\lambda-\sigma)(\lambda-\bar{\sigma})}. \tag{5.4}$$

The question now is whether it is possible to reconcile the two aims (I) and (II) set out in the introduction, when computing B_+v and B_-v for any $v \in \mathbf{R}^N$. Reference back to (2.8) shows that

$$B_+v = \left[(F-\rho M) + \theta^2 M(F-\rho M)^{-1}M\right]^{-1} Mv = X^{-1}v.$$

If real arithmetic is mandatory (aim (I)) then the presence of the full matrix $(F-\rho M)^{-1}$ in X precludes the exploitation of bandedness (aim (II)) in the triangular factorization of X. This leaves two possibilities:

1. Solve $Xu = v$ for u iteratively, in real arithmetic, exploiting band structure as described at the end of Section 3.
2. Ignore the structure of B_+ and evaluate B_+v by solving $(F-\sigma M)u = Mv$ in complex arithmetic and then returning the real part (respectively the imaginary part for a method using B_-).

The second approach yields Method 5.

Method 5

1. Solve $(F-\sigma M)w = Mv$ (complex arithmetic).
2. Set $B_+v = Re(w)$ (respecively $B_-v = Im(w)$).

The cost of the above method is as follows.

- Arithmetic. Factorization (done only once) : $\beta^2 N$ complex multiplications, Forward and backward solutions: $2(\beta+1)N$ complex multiplications. Normalization: N comparisons and N real multiplications in step 2.2.
- Storage. $2(2\beta+1)N$ real locations for F and M; $(2\beta+1)N$ complex locations for $F-\sigma M = LU$; plus two complex vectors for storing $x^{(k)}$ and $y^{(k)}$.

Method 5 is a compromise. What must be emphasized here is that from the point of view of the iterative methods, such as Arnoldi, Lanczos, or subspace iteration, that will be making use of B_+ there is no compromise. Goal (I) is realized. These iterations will use real arithmetic exclusively. Goal (II) is achieved by using complex arithmetic in the lower level subroutine that evaluates B_+v.

Note that the cost of Method 5 is lower than that of Method 4 of the previous section. In fact the extra work in Method 4 brings no further benefit in the light of the following surprising result. Recall that $G = (F-\sigma M)M^{-1}(F-\sigma M) + \theta^2 M$.

Theorem 5.1. *The matrices B_-, G and M are related by*

$$B_- = \theta G^{-1} M. \tag{5.5}$$

Proof. We have, by definition,

$$\begin{aligned} B_- &= \frac{1}{2i}\left[(F-\sigma M)^{-1} - (F-\bar{\sigma}M)^{-1}\right]M \\ &= \frac{1}{2i}(F-\sigma M)^{-1}\left[(F-\bar{\sigma}M) - (F-\sigma M)\right](F-\bar{\sigma}M)^{-1}M \\ &= (F-\sigma M)^{-1}\theta M(F-\bar{\sigma}M)^{-1}M \\ &= \theta G^{-1}M \end{aligned}$$

∎

By the theorem the solution $y^{(k)}$ of (4.2) is identical with $B_-x^{(k)}$, apart from the multiplicative scalar θ. Note also that the right-hand side of (5.5) is nothing but θY^{-1}, i.e., the block in position (2,1) of the matrix in (2.6).

6. Numerical experiments

All numerical tests have been performed on a Vax-785 using double precision, i.e., the unit roundoff is $2^{-56} \approx 1.3877 \text{ x } 10^{-17}$. Our test example, taken from [7], models concentration waves in reaction and transport interaction of some chemical solutions in a tubular reactor. The concentrations of the two reacting and diffusing components are $x(\tau, z)$ and $y(\tau, z)$ where τ is the time and z is the distance down the tube. The system is modeled by [7]:

$$\frac{\partial x}{\partial \tau} = \frac{D_x}{L^2}\frac{\partial^2 x}{\partial z^2} + f(x,y), \tag{6.1}$$

$$\frac{\partial y}{\partial \tau} = \frac{D_y}{L^2}\frac{\partial^2 y}{\partial z^2} + g(x,y), \tag{6.2}$$

with the initial condition

$$x(0,z) = x_0(z), \quad y(0,z) = y_0(z), \quad \forall\, z \in [0,1],$$

and the Dirichlet boundary conditions:

$$x(0,\tau) = x(1,\tau) = x^* \ , \quad y(0,\tau) = y(1,\tau) = y^*.$$

We consider in particular the so-called Brusselator wave model [7] in which

$$f(x,y) = A - (B+1)x + x^2 y \ , \quad g(x,y) = Bx - x^2 y. \tag{6.3}$$

Then, the above system admits the trivial stationary solution $x^* = A$, $y^* = B/A$.

In this problem one is primarily interested in the existence of stable periodic solutions to the system as the bifurcation parameter L varies. This occurs when the eigenvalues of largest real parts of the Jacobian of the right hand side of (6.1) - (6.2), evaluated at the steady state solution, are purely imaginary. For the purpose of verifying this fact numerically, one first needs to discretize the equations with respect to the variable z and compute the eigenvalues with largest real parts of the resulting discrete Jacobian.

The exact eigenvalues are known and this problem is analytically solvable. The article [7] considers the following set of parameters

$$D_x = 0.008, \quad D_y = \frac{1}{2}D_x = 0.004, \quad A = 2, \quad B = 5.45.$$

For small L the Jacobian has only eigenvalues with negative real parts. At $L \approx 0.51302$ a purely imaginary eigenvalue appears.

We discretize the interval $[0,1]$ using n interior points, and define the mesh size $h \equiv 1/(n+1)$. The discrete vector is of the form $\begin{pmatrix} x \\ y \end{pmatrix}$ where x and y are n-dimensional vectors. We denote by f_h and g_h the corresponding discretized functions f and g, the Jacobian is a 2 x 2 block matrix in which the diagonal blocks $(1,1)$ and $(2,2)$ are the matrices

$$\frac{1}{h^2}\frac{D_x}{L^2}\ Tridiag\{1,-2,1\} + \frac{\partial f_h(x,y)}{\partial x}$$

and

$$\frac{1}{h^2}\frac{D_y}{L^2}\ Tridiag\{1,-2,1\} + \frac{\partial g_h(x,y)}{\partial y}$$

respectively, while the blocks $(1,2)$ and $(2,1)$ are

$$\frac{\partial f_h(x,y)}{\partial y} \quad \text{and} \quad \frac{\partial g_h(x,y)}{\partial x}$$

respectively. Note that since the two functions f and g do not depend on the variable z, the Jacobians of either f_h or g_h with respect to either x or y are scaled identity matrices. We denote by A the resulting $2n$ x $2n$ Jacobian matrix. In the following tests we took $n = 100$, which yields a matrix A of size 200. We point out that the exact eigenvalues of A are readily computable, since there exists a quadratic relation between the eigenvalues of the matrix A and those of the classical difference matrix $Tridiag\{1,-2,1\}$. In fact part of the spectrum (the 32 rightmost eigenvalues) of the matrix A is shown in Figure 4. We have not shown the rest of the spectrum of A consisting of 168 *real* eigenvalues that are almost uniformely distributed in the interval $[-1235.5, -51.912]$. The rightmost eigenvalues, determined with maximum accuracy, i.e., approximately 16 digits are

$$\lambda_{1,2} = 1.8199876787305946 \text{ x } 10^{-5} \pm i\ 2.139497522076329$$

As is observed the real part is close to zero, which supports the theory within discretization errors.

The purpose of these experiments is to compare the performances of the methods using the three approaches $B = (A-\sigma I)^{-1}$, $B_+ = Re(B)$ and $B_- = Im(B)$, all in conjunction with Arnoldi's method. We have plotted the convergence history for the three methods for three choices of the shift σ, namely $\sigma = 0.1 + 2.1i$, $\sigma = 0.0 + 2.5i$, and $\sigma = 0.5 + 2.1i$. The plots in Figures 1, 2 and 3 show the relative errors

$$\left| \frac{\lambda_1^{(m)} - \lambda_1}{\lambda_1} \right| \tag{6.4}$$

versus the number of Arnoldi steps. As is observed the performances of the two different approaches are not constant.

In Figures 5 and 6 we show the spectra of the corresponding matrices B_+ and B_- for the last two cases, i.e., for $\sigma = 0.5 + 2.1i$ and for $\sigma = 2.5i$. In each case we have circled the eigenvalue of largest modulus. Notice the very good separation properties of the dominant eigenvalue despite a relatively distant shift. Also observe the concentration around the origin of the transformed large eigenvalues of A. The reader should note that the scales are different. For example in the top graph in Figure 6 the x-axis has a total length of 0.08, which means that the spectrum is almost purely imaginary in this case. The spectra of B_+ and B_- bear no particular resemblance and it is hard to predict from looking at the pictures only which method will converge faster.

It is also instructive to compare the two mappings $\mu_+(\lambda)$ and $\mu_-(\lambda)$ as defined by (5.4). As an experiment we plotted the images $\mu_+(\lambda)$ and $\mu_-(\lambda)$ of several circles of small radii, centered at the shift σ. The goal is to compare the two mappings for a similar situation where the eigenvalues of the original matrix A are distributed in circles around the shift. When $\sigma = 0.5 + 2.1i$ and the radii were $0.1, 0.2, \ldots, 0.5$ respectively, the two resulting plots looked very much like five concentric circles and were almost indistinguishable for the cases B_+ and B_-. For this reason we omit to show the resulting figures. Changing σ to $\sigma = 0.5i$ and taking the same radii as above produced the graphs in Figure 7.

Notice again that the outmost curves, those corresponding to the dominant eigenvalues of B_+ and B_-, are slight perturbations of circles. Although not apparent at first glance, the outmost curve for B_+ is almost superposable with that of B_- provided we shifted the whole plot of B_- in the south-west direction by about one unit of the graph. Notice also that the B_- graph is symmetric about the real axis as is expected from the definition of μ_-. Similarly, the μ_+ curves can be seen to be symmetric with respect to the imaginary axis, as is verified in the plots.

In the following discussion we assume that there is one actual eigenvalue of A per circle, i.e, there is only one eigenvalue of B_+ or B_- per curve represented in the two plots of Figure 7 repectively. Assume at first that the two dominant eigenvalues for B_+ are located on the imaginary axis in the lower half plane. These are roughly $\mu_{+,1} \approx -5.45i$ and $\mu_{+,2} \approx -2.9i$ which means that the convergence ratio in inverse iteration would be

$$\left|\frac{\mu_{+,2}}{\mu_{+,1}}\right| \approx 0.532 \ .$$

It is found that the corresponding eigenvalues of the B_- matrix are the two dominant eigenvalues $\mu_{-,1} \approx -4.54$ and $\mu_{-,2} - 2.08$, which leads to the convergence ratio of

$$\left|\frac{\mu_{-,2}}{\mu_{-,1}}\right| \approx 0.462 \ ,$$

much better than that of the B_+ approach.

Assume on the other hand that both the dominant and the subdominant eigenvalues of B_- are located on the real axis on the right half plane: $\mu_{-,1} \approx 5.554$ and $\mu_{-,2} \approx 3.12$. Then the associated convergence ratio for inverse iteration with B_- becomes

$$\left|\frac{\mu_{-,2}}{\mu_{-,1}}\right| \approx 0.562 \ .$$

The corresponding eigenvalues of B_+ are found to be approximately $\mu_{+,1} \approx 4.44$ and $\mu_{+,2} \approx 1.875$ which gives the convergence ratio

$$\left|\frac{\mu_{+,2}}{\mu_{+,1}}\right| \approx 0.44 \ ,$$

for inverse iteration with B_+, a much better ratio than that of the B_- approach. Thus, the previous situation has been completely reversed. What is interesting is that this has occured in spite of keeping the distances of the two eigenvalues of A closest to the shift the same in both situations. In other words it is not only the distance of these eigenvalues that matters for the speed of inverse iteration, but also their relative location around the shift. This tells us that in practice it will be vain to try determining a-priori which of the two approaches is to be favored.

7. Summary and conclusion

We have considered five different ways of implementing shift and invert techniques for the generalized eigenvalue problem $Fz = \lambda Mz$, in the common situation where the shift is complex

while F and M are real and banded matrices. Method 1 is the most natural one that simply accepts complex arithmetic. Its defect is to force a fully complex implementation of the iterative method, be it Arnoldi, Lanczos or subspace iteration, with associated costs both obvious and subtle. There are better approaches.

If M is diagonal then Method 2 is strongly recommended. If M is banded then we have shown that Method 4 is majorized by Method 5. That leaves Methods 3 and 5. Both are preferable to Method 1 but we cannot make a definite choice between them. Method 3 employs an inner iteration to solve the linear system. It is therefore less well specified and demands a bigger programming effort. Nevertheless there will be problems where the resulting procedure is preferable to Method 5.

The attraction of Method 5 is its simple black box nature. It also carries the obligation to choose between B_+ and B_-. It is possible, but wasteful, to use both. For some problems the correct choice will be important, for others not at all. However, it should be emphasized that it is easy to write a program that will realize either of them according to the setting of a parameter. Consequently, experimentation is no extra burden. All our numerical experience confirms that the better of the B_+ and B_- approaches converges as well as Method 1. Yet Method 5 keeps the operator real.

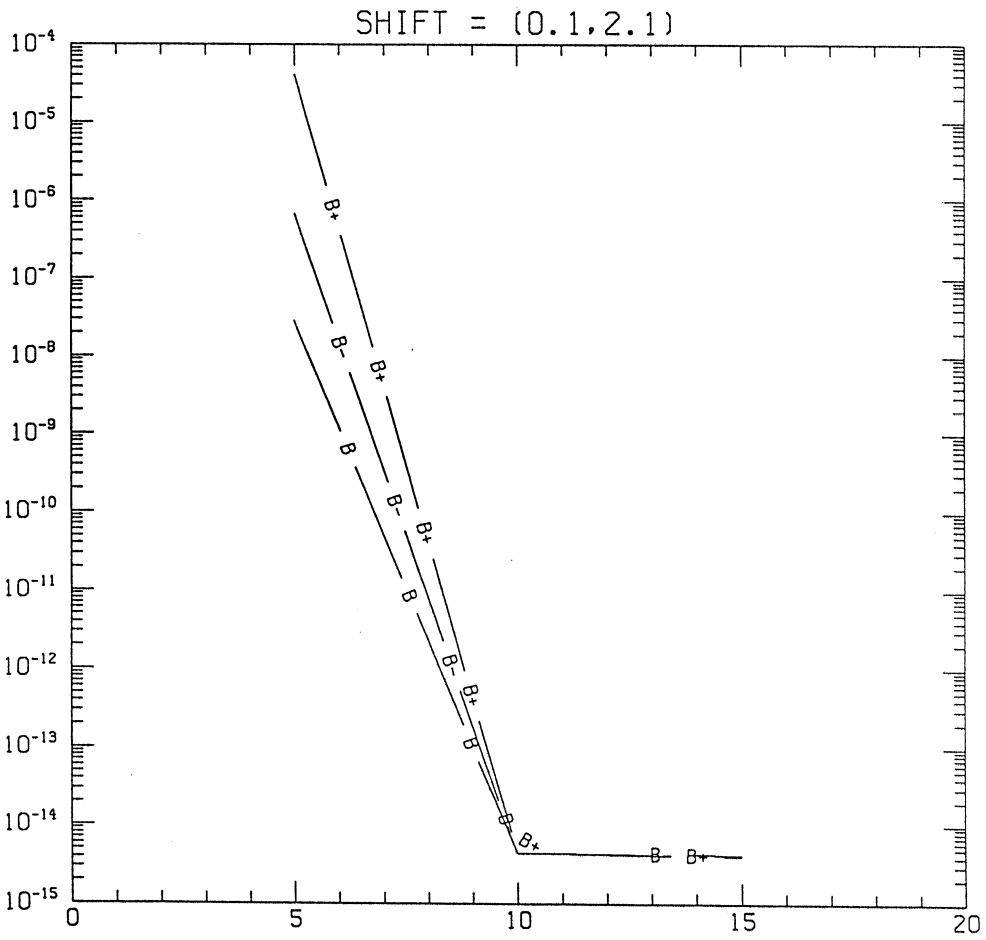


Figure 1: Convergence history for $\sigma = 0.1 + 2.1\ i$

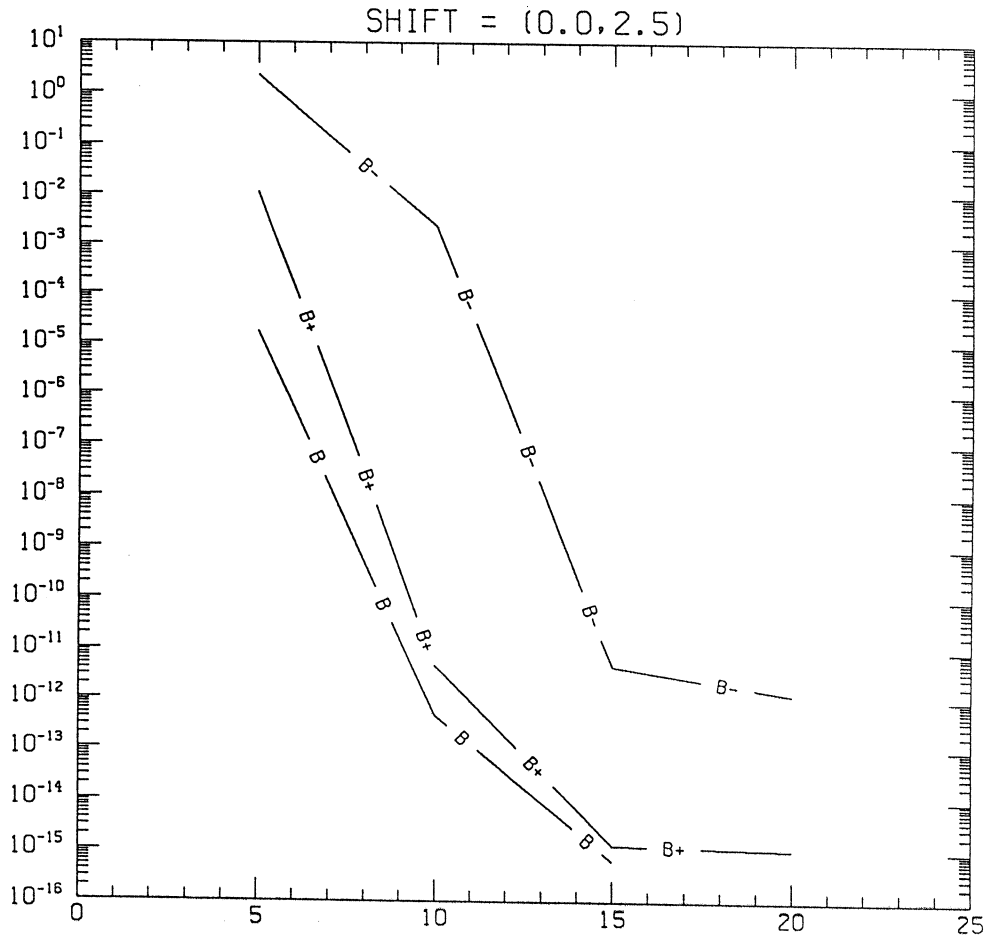


Figure 2: Convergence history for $\sigma = 0.0 + 2.5\ i$

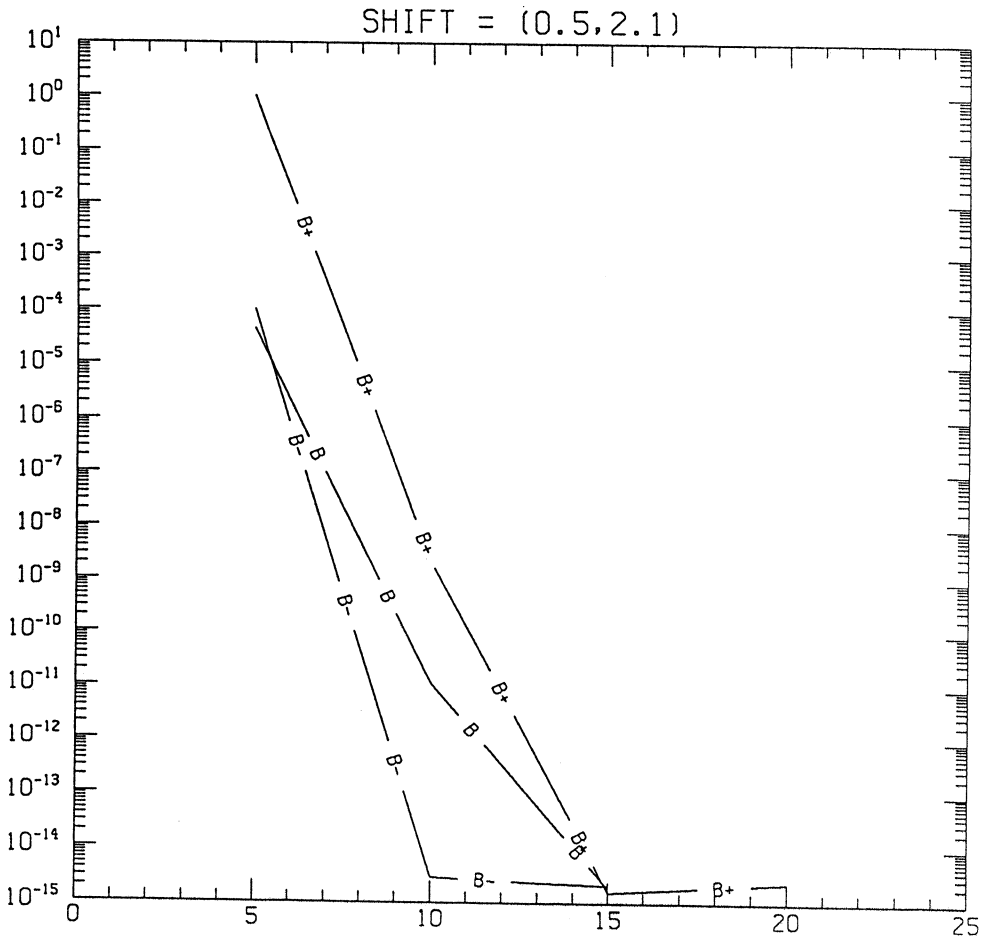


Figure 3: Convergence history for $\sigma = 0.5 + 2.1\ i$

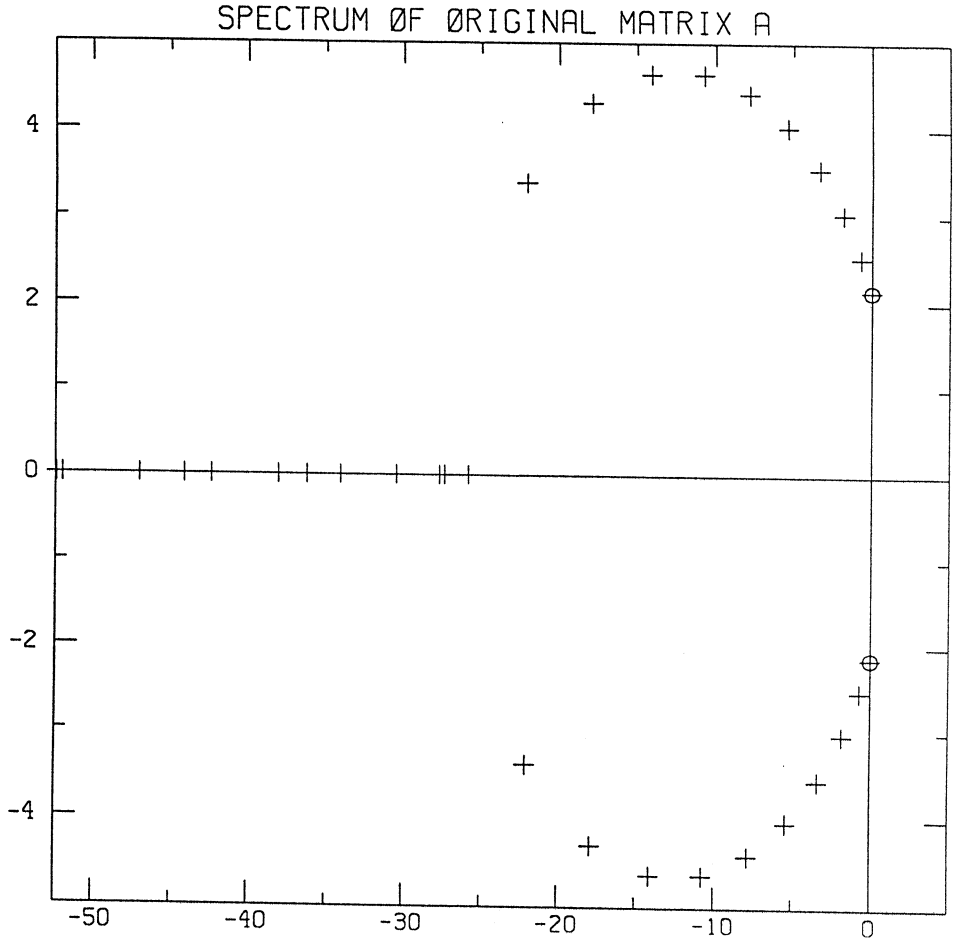
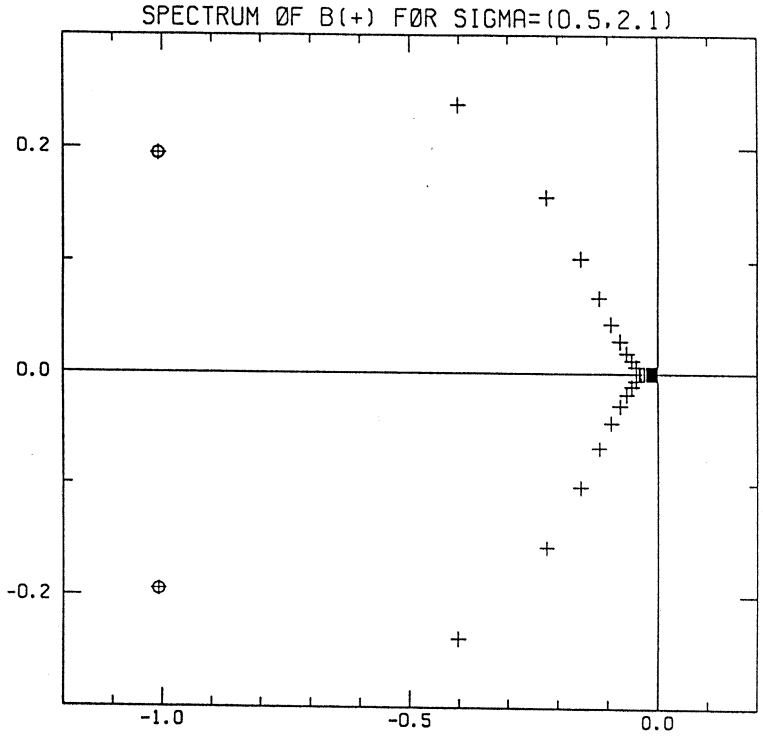


Figure 4: Spectrum of the original matrix A for $n = 100$.



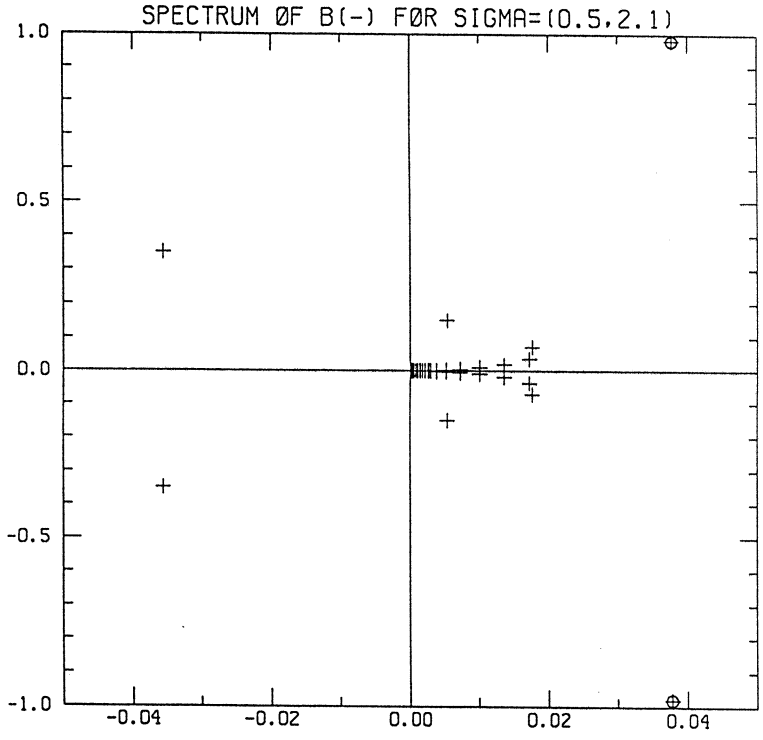
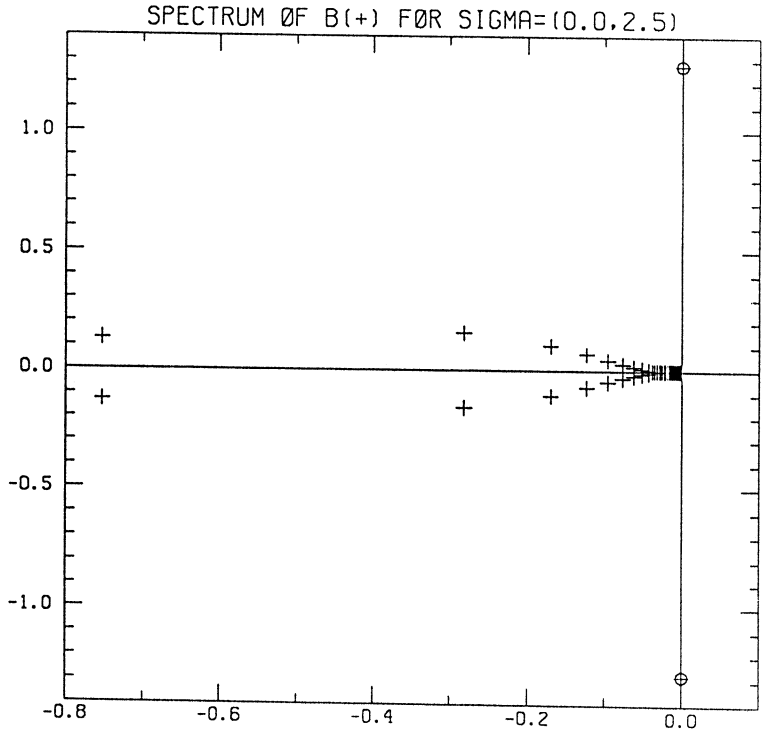


Figure 5: Spectra of B_+ and B_- for $\sigma = 0.5 + 2.1\,i$



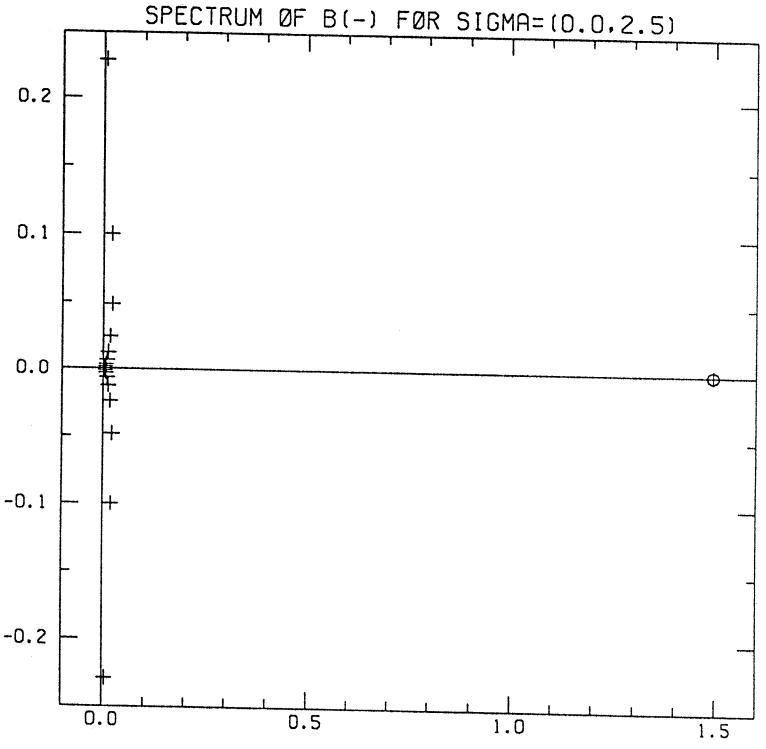


Figure 6: Spectra of B_+ and B_- for $\sigma = 0.0 + 2.5\ i$

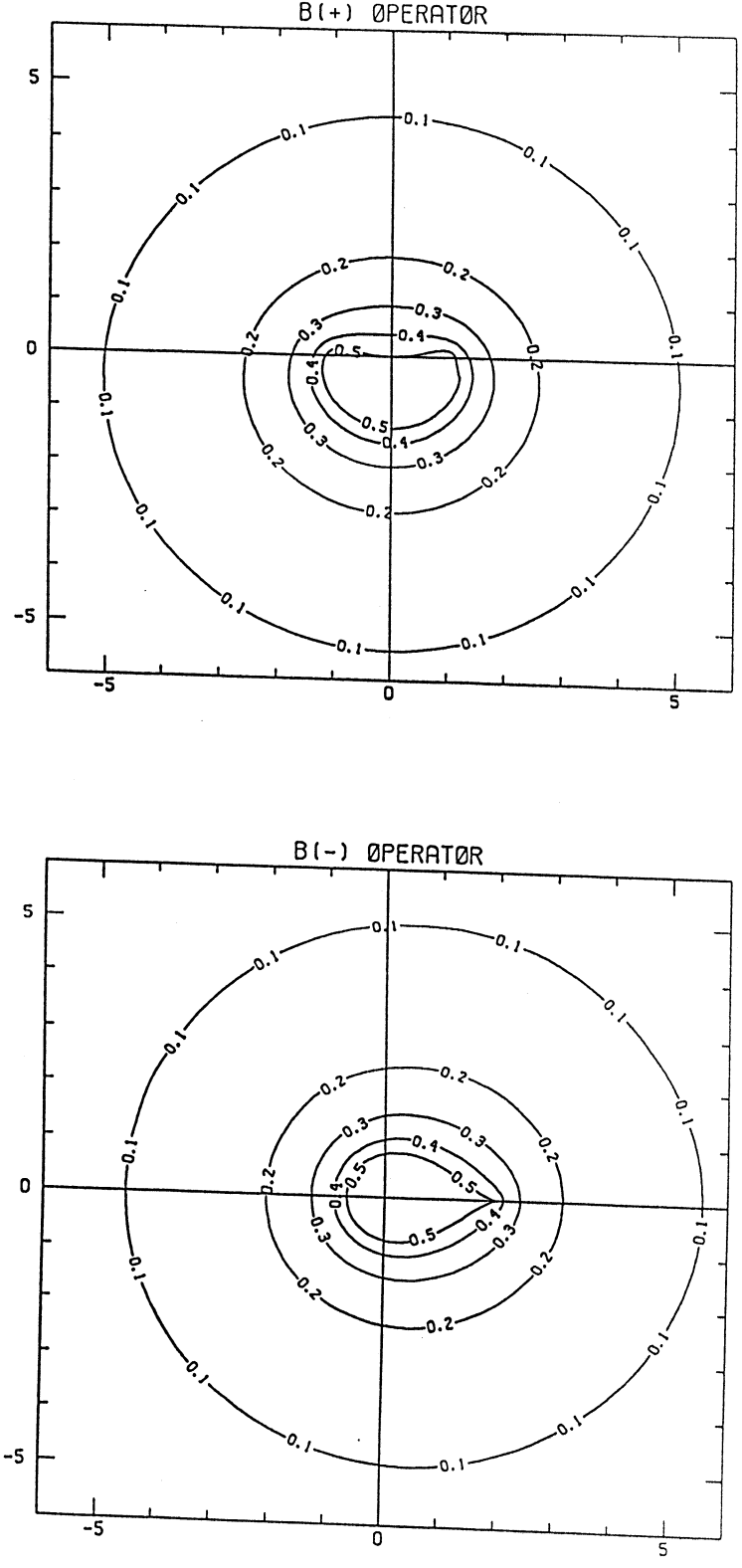


Figure 7: Images of circles by the μ_+ and μ_- transforms.

References

[1] W.E. Arnoldi, *The principle of minimized iteration in the solution of the matrix eigenvalue problem,* Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] J. Cullum and R. Willoughby, A Lanczos procedure for the modal analysis of very large nonsymmetric matrices, *Proceedings of the 23rd Conference on Decision and Control, Las Vegas,* 1984.

[3] H.C. Elman, *Iterative Methods for Large Sparse Nonsymmetric Systems of Linear Equations,* Ph.D. Thesis, Yale University, Computer Science Dpt., 1982.

[4] J. G. F. Francis, *The QR transformations, Parts I and II,* Computer J., 4 (1961-1962), pp. 362–363,and332–345.

[5] A. Jennings and W.J. Stewart, *A simultaneous iteration algorithm for real matrices,* ACM, Trans. of Math. Software, 7 (1981), pp. 184–198.

[6] B.N. Parlett, D. R. Taylor, and Z.S. Liu, *A look-ahead Lanczos algorithm for nonsymmetric matrices,* Mathematics of Computation, 44 (1985), pp. 105–124.

[7] Raschman P., M. Kubicek and M. Maros, Waves in distributed chemical systems: experiments and computations, P.J. Holmes ed., *New Approaches to Nonlinear Problems in Dynamics - Proceedings of the Asilomar Conference Ground, Pacific Grove, California 1979,* The Engineering Foundation, SIAM, 1980, pp. 271–288.

[8] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices,* Lin. Alg. Appl., 34 (1980), pp. 269–295.

[9] J. H. Wilkinson, *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford, 1965.