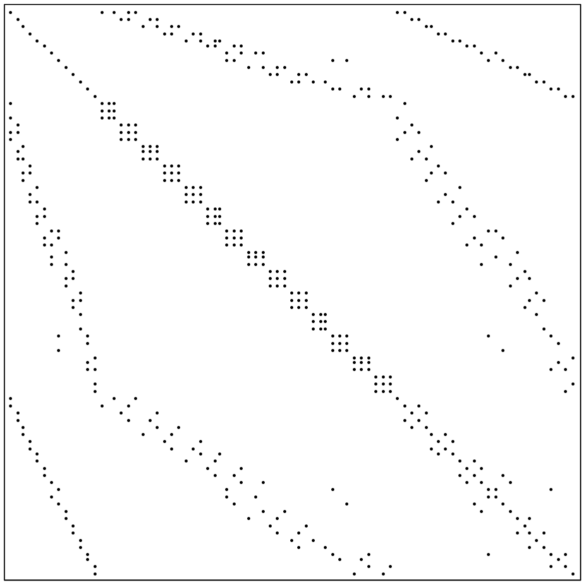
Manchester University Press Series in Algorithms and Architectures for Advanced Scientific Computing

NUMERICAL METHODS FOR LARGE EIGENVALUE PROBLEMS

Y. Saad



Contents

Preface **vii**

I Background in Matrix Theory and Linear Algebra **1**

1 . Matrices 2
2 . Square Matrices and Eigenvalues 3
3 . Types of Matrices 5
4 . Vector Inner Products and Norms 7
5 . Matrix Norms 9
6 . Subspaces 11
7 . Orthogonal Vectors and Subspaces 12
8 . Canonical Forms of Matrices 14
8.1. Reduction to the Diagonal Form. 16
8.2. The Jordan Canonical Form 17
8.3. The Schur Canonical Form 23
9 . Normal and Hermitian Matrices 26
9.1. Normal Matrices 26
9.2. Hermitian Matrices 29
10. Nonnegative Matrices 33

II Sparse Matrices **37**

1 . Introduction 38
2 . Storage Schemes 40

3 . Basic Sparse Matrix Operations 44
4 . Sparse Direct Solution Methods 46
5 . Test Problems 47
5.1. Random Walk Problem 48
5.2. Chemical Reactions 50
5.3. The Harwell-Boeing Collection 52
6 . SPARSKIT 53

III Perturbation Theory and Error Analysis 59
1 . Projectors and their Properties 60
1.1. Orthogonal Projectors 60
1.2. Oblique Projectors 63
1.3. Resolvent and Spectral Projector 65
1.4 Relations with the Jordan form 67
1.5. Linear Perturbations of A 71
2 . A-Posteriori Error Bounds 76
2.1. General Error Bounds 76
2.2. The Hermitian Case 79
2.3. The Kahan-Parlett-Jiang Theorem 86
3 . Conditioning of Eigen-problems 91
3.1. Conditioning of Eigenvalues 91
3.2. Conditioning of Eigenvectors 95
3.3. Conditioning of Invariant Subspaces 98
4 . Localization Theorems 101

IV The Tools of Spectral Approximation 109
1 . Single Vector Iterations 110
1.1. The Power Method 110
1.2. The Shifted Power Method 113
1.3. Inverse Iteration 114
2 . Deflation Techniques 117
2.1. Wielandt Deflation with One Vector 117
2.2. Optimality in Wieldant's Deflation 119
2.3. Deflation with Several Vectors. 122
2.4. Partial Schur Decomposition. 123
2.5. Practical Deflation Procedures 124

3 . General Projection Methods 126
3.1. Orthogonal Projection Methods 127
3.2. The Hermitian Case 131
3.3. Oblique Projection Methods 139
4 . Chebyshev Polynomials 142
4.1. Real Chebyshev Polynomials 142
4.2. Complex Chebyshev Polynomials 144

V Subspace Iteration 151
1 . Simple Subspace Iteration 152
2 . Subspace Iteration with Projection 156
3 . Practical Implementations 160
3.1. Locking 160
3.2. Linear Shifts 162
3.3. Preconditionings 163

VI Krylov Subspace Methods 167
1 . Krylov Subspaces 168
2 . Arnoldi's Method 172
2.1. The Basic Algorithm 172
2.2. Practical Implementations 176
2.3. Incorporation of Implicit Deflation 180
3 . The Hermitian Lanczos Algorithm 183
3.1. The Algorithm 184
3.2. Relation with Orthogonal Polynomials . . . 185
4 . Non-Hermitian Lanczos algorithm 186
4.1. The Algorithm 187
4.2. Practical Implementations 192
5 . Block Krylov Methods 196
6 . Convergence of the Lanczos Process 198
6.1. Distance between $\mathcal{K}_m$ and an Eigenvector . 199
6.2. Convergence of the Eigenvalues 201
6.3. Convergence of the Eigenvectors 203
7 . Convergence of the Arnoldi Process 204

VII Acceleration Techniques and Hybrid Methods 219

1 . The Basic Chebyshev Iteration 220

1.1. Convergence Properties. 224

2 . Arnoldi–Chebyshev Iteration 226

2.1. Purification by Arnoldi's Method 226

2.2. The Enhanced Initial Vector Approach . . 227

2.3. Computing an Optimal Ellipse 228

2.4. Starting the Chebyshev Iteration. 232

2.5. Choosing the Parameters m and k. 234

3 . Deflated Arnoldi-Chebyshev 235

4 . Chebyshev Subspace Iteration 237

4.1. Getting the Best Ellipse. 238

4.2. Parameters k and m. 238

4.3. Deflation 239

5 . Least Squares - Arnoldi 239

5.1. The Least Squares Polynomial 240

5.2. Use of Chebyshev Bases 243

5.3. The Gram Matrix 244

5.4. Computing the Best Polynomial 246

5.5. Least Squares Arnoldi Algorithms 251

VIII Preconditioning Techniques 257

1 . Shift-and-invert Preconditioning 258

1.1. General Concepts 258

1.2. Dealing with Complex Arithmetic 260

1.3. Shift-and-Invert Arnoldi 263

2 . Polynomial Preconditioning 267

3 . Davidson's Method 272

4 . Generalized Arnoldi Algorithms 276

IX Non-Standard Eigenvalue Problems 281

1 . Introduction 282

2 . Generalized Eigenvalue Problems 282

2.1. General Results 283

2.2. Reduction to Standard Form 290

2.3. Deflation 292

2.4. Shift-and-Invert 293
2.5. Projection Methods 294
2.6. The Hermitian Definite Case 295
3 . Quadratic Problems 299
3.1. From Quadratic to Generalized Problems . 300

X Origins of Matrix Eigenvalue Problems 303
1 . Introduction 304
2 . Mechanical Vibrations 305
3 . Electrical Networks. 311
4 . Quantum Chemistry 312
5 . Stability of Dynamical Systems 313
6 . Bifurcation Analysis 315
7 . Chemical Reactions 316
8 . Macro-economics 318
9 . Markov Chain Models 319

References 322

Index 342

Preface

Matrix eigenvalue problems arise in a large number of disciplines of sciences and engineering. They constitute the basic tool used in designing buildings, bridges, and turbines, that are resistent to vibrations. They allow to model queueing networks, and to analyze stability of electrical networks or fluid flow. They also allow the scientist to understand local physical phenonema or to study bifurcation patterns in dynamical systems. In fact the writing of this book was motivated mostly by the second class of problems.

Several books dealing with numerical methods for solving eigenvalue problems involving symmetric (or Hermitian) matrices have been written and there are a few software packages both public and commercial available. The book by Parlett [118] is an excellent treatise of the problem. Despite a rather strong demand by engineers and scientists there is little written on nonsymmetric problems and even less is available in terms of software. The 1965 book by Wilkinson [183] still constitutes an important reference. Certainly, science has evolved since the writing of Wilkinson's book and so has the computational environment and the demand for solving large matrix problems. Problems are becoming larger and more complicated while at the same time computers are able to deliver ever higher performances. This means in particular that methods that were deemed too demanding yesterday are now in the realm of the achievable. I hope that this book will be a small step in bridging the gap between the literature on what is available in the symmetric case and the nonsymmetric case. Both the Hermitian and the non-Hermitian case are covered, although non-Hermitian problems are given more emphasis.

This book attempts to achieve a good balance between theory and practice. I should comment that the theory is especially important in the nonsymmetric case. In essence what differentiates the Hermitian from the non-Hermitian eigenvalue problem is that in the first case we can always manage to compute an approximation whereas there are nonsymmetric problems that can be arbitrarily difficult to solve and can essentially make any algorithm fail. Stated more rigorously, the eigenvalue of a Hermitian matrix is always well-conditioned whereas this is not true for nonsymmetric matrices. On the practical side, I tried to give a general view of algorithms and tools that have proved efficient. Many of the algorithms described correspond to actual implementations of research software and have been tested on realistic problems. I have tried to convey our experience from the practice in using these techniques.

As a result of the partial emphasis on theory, there are a few chapters that may be found hard to digest for readers inexperienced with linear algebra. These are Chapter III and to some extent, a small part of Chapter IV. Fortunately, Chapter III is basically independent of the rest of the book. The minimal background needed to use the *algorithmic part* of the book, namely Chapters IV through VIII, is calculus and linear algebra at the undergraduate level. The book has been used twice to teach a special topics course; once in a Mathematics department and once in a Computer Science department. In a quarter period representing roughly 12 weeks of 2.5 hours lecture per week, Chapter I, III, and IV, to VI have been covered without much difficulty. In a semester period, 18 weeks of 2.5 hours lecture weekly, all chapters can be covered with various degrees of depth. Chapters II and X need not be treated in class and can be given as remedial reading.

Finally, I would like to extend my appreciation to a number of people to whom I am indebted. Françoise Chatelin, who was my thesis adviser, introduced me to numerical methods for eigenvalue problems. Her influence on my way of thinking is certainly reflected in this book. Beresford Parlett has been encouraging throughout my career and has always been a real inspiration. Part of the motivation in getting this book completed, rather than 'never finished', is owed to L. E. Scriven from the Chemical Engineering department and to many others in applied sciences who expressed interest in my work. I am indebted to Roland Freund who has read this manuscript with great care and has pointed out numerous mistakes.

Minneapolis, December 1991
Youcef Saad

Chapter I

Background in Matrix Theory and Linear Algebra

This chapter reviews basic matrix theory and introduces some of the elementary notation used throughout the book. Matrices are objects that represent linear mappings between vector spaces. The notions that will be predominantly used in this book are very intimately related to these linear mappings and it is possible to discuss eigenvalues of linear operators without ever mentioning their matrix representations. However, to the numerical analyst, or the engineer, any theory that would be developed in this manner would be insufficient in that it will not be of much help in developing or understanding computational algorithms. The abstraction of linear mappings on vector spaces does however provide very concise definitions and some important theorems.

1. Matrices

When dealing with eigenvalues it is more convenient, if not more relevant, to manipulate complex matrices rather than real matrices. A complex $n \times m$ matrix A is an $n \times m$ array of complex numbers

$$a_{ij}, \quad i = 1, \ldots, n, \ j = 1, \ldots, m.$$

The set of all $n \times m$ matrices is a complex vector space denoted by $\mathbb{C}^{n\times m}$. The main operations with matrices are the following:

- Addition: $C = A + B$, where A, B and C are matrices of size $n \times m$ and
$$c_{ij} = a_{ij} + b_{ij} \ ,$$
$i = 1, 2, \ldots n$, $j = 1, 2, \ldots m$.
- Multiplication by a scalar: $C = \alpha A$, where $c_{ij} = \alpha \ a_{ij}$.
- Multiplication by another matrix:
$$C = AB,$$
where $A \in \mathbb{C}^{n\times m}, B \in \mathbb{C}^{m\times p}, C \in \mathbb{C}^{n\times p}$, and
$$c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}.$$

A notation that is often used is that of column vectors and row vectors. The column vector $a_{.j}$ is the vector consisting of the j-th column of A, i.e., $a_{.j} = (a_{ij})_{i=1,\ldots,n}$. Similarly we will use the notation $a_{i.}$ to denote the i-th row of the matrix A. For example, we may write that

$$A = (a_{.1}, a_{.2}, \ldots, a_{.m}) \ .$$

or

$$A = \begin{pmatrix} a_{1.} \\ a_{2.} \\ . \\ . \\ a_{n.} \end{pmatrix}$$

The *transpose* of a matrix A in $\mathbb{C}^{n\times m}$ is a matrix C in $\mathbb{C}^{m\times n}$ whose elements are defined by $c_{ij} = a_{ji}, i = 1, \ldots, m,\ j = 1, \ldots, n$. The transpose of a matrix A is denoted by A^T. It is more relevant in eigenvalue problems to use the *transpose conjugate* matrix denoted by A^H and defined by

$$A^H = \bar{A}^T = \overline{A^T}$$

in which the bar denotes the (element-wise) complex conjugation.

Finally, we should recall that matrices are strongly related to linear mappings between vector spaces of finite dimension. They are in fact representations of these transformations with respect to two given bases; one for the initial vector space and the other for the image vector space.

2. Square Matrices and Eigenvalues

A matrix belonging to $\mathbb{C}^{n\times n}$ is said to be square. Some notions are only defined for square matrices. A square matrix which is very important is the identity matrix

$$I = \{\delta_{ij}\}_{i,j=1,\ldots,n}$$

where δ_{ij} is the Kronecker symbol. The identity matrix satisfies the equality $AI = IA = A$ for every matrix A of size n. The inverse of a matrix, when it exists, is a matrix C such that $CA = AC = I$. The inverse of A is denoted by A^{-1}.

The determinant of a matrix may be defined in several ways. For simplicity we adopt here the following recursive definition. The determinant of a 1×1 matrix (a) is defined as the scalar a. Then the determinant of an $n \times n$ matrix is given by

$$\det(A) = \sum_{j=1}^{n} (-1)^{j+1} a_{1j} \det(A_{1j})$$

where A_{1j} is an $(n-1) \times (n-1)$ matrix obtained by deleting the 1-st row and the $j-th$ column of A. The determinant of a matrix determines whether or not a matrix is singular since A is singular if and only if its determinant is zero. We have the following simple properties:

- $\det(AB) = \det(BA)$,
- $\det(A^T) = \det(A)$,
- $\det(\alpha A) = \alpha^n \det(A)$,
- $\det(\bar{A}) = \overline{\det(A)}$,
- $\det(I) = 1$.

From the above definition of the determinant it can be shown by induction that the function that maps a given complex value λ to the value $p_A(\lambda) = \det(A - \lambda I)$ is a polynomial of degree n (Problem P-1.6). This is referred to as the *characteristic polynomial* of the matrix A.

Definition 1.1 *A complex scalar λ is called an eigenvalue of the square matrix A if there exists a nonzero vector u of $\mathbb{C}^n$ such that $Au = \lambda u$. The vector u is called an eigenvector of A associated with λ. The set of all the eigenvalues of A is referred to as the spectrum of A and is denoted by $\sigma(A)$.*

An eigenvalue of A is a root of the characteristic polynomial. Indeed λ is an eigenvalue of A iff $\det(A - \lambda I) \equiv p_A(\lambda) = 0$. So there are at most n distinct eigenvalues. The maximum modulus of the eigenvalues is called *spectral radius* and is denoted by $\rho(A)$:

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|.$$

The *trace* of a matrix is equal to the sum of all its diagonal elements,

$$\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}.$$

It can be easily shown that this is also equal to the sum of its eigenvalues counted with their multiplicities as roots of the characteristic polynomial.

Proposition 1.1 *If λ is an eigenvalue of A then $\bar{\lambda}$ is an eigenvalue of A^H. An eigenvector v of A^H associated with the eigenvalue $\bar{\lambda}$ is called left eigenvector of A.*

When a distinction is necessary, an eigenvector of A is often called a right eigenvector. Thus the eigenvalue λ and the right and left eigenvectors, u and v, satisfy the relations

$$Au = \lambda u \ , \quad v^H A = \lambda v^H$$

or, equivalently,

$$u^H A^H = \bar{\lambda} u^H \ , \quad A^H v = \bar{\lambda} v \ .$$

3. Types of Matrices

The properties of eigenvalues and eigenvectors of square matrices will sometimes depend on special properties of the matrix A. For example, the eigenvalues or eigenvectors of the following types of matrices will all have some special properties.

- *Symmetric matrices:* $A^T = A$;
- *Hermitian matrices:* $A^H = A$;
- *Skew-symmetric matrices:* $A^T = -A$;
- *Skew-Hermitian matrices:* $A^H = -A$;
- *Normal matrices:* $A^H A = AA^H$;
- *Nonnegative matrices:* $a_{ij} \geq 0, \ i, j = 1, \ldots, n$ (similar definition for nonpositive, positive, and negative matrices);

- *Unitary matrices:* $Q^H Q = I.$

Often, a matrix Q such that $Q^H Q$ is diagonal is called orthogonal. It is worth noting that a unitary matrix Q is a matrix whose inverse is its transpose conjugate Q^H.

Some matrices have particular structures that are often convenient for computational purposes and play important roles in numerical analysis. The following list though incomplete, gives an idea of the most important special matrices arising in applications and algorithms.

- *Diagonal matrices:* $a_{ij} = 0$ for $j \neq i$. Notation:

$$A = \text{diag}\ (a_{11}, a_{22}, \ldots, a_{m}) .$$

- *Upper triangular matrices:* $a_{ij} = 0$ for $i > j$.
- *Lower triangular matrices:* $a_{ij} = 0$ for $i < j$.
- *Upper bidiagonal matrices:* $a_{ij} = 0$ for $j \neq i$ or $j \neq i + 1$.
- *Lower bidiagonal matrices:* $a_{ij} = 0$ for $j \neq i$ or $j \neq i - 1$.
- *Tridiagonal matrices:* $a_{ij} = 0$ for any pair i, j such that $|j - i|>1$. Notation:

$$A = \text{tridiag}\ (a_{i,i-1}, a_{ii}, a_{i,i+1}) .$$

- *Banded matrices:* there exist two integers m_l and m_u such that $a_{ij} \neq 0$ only if $i - m_l \leq j \leq i + m_u$. The number $m_l + m_u + 1$ is called the bandwidth of A.
- *Upper Hessenberg matrices:* $a_{ij} = 0$ for any pair i, j such that $i > j + 1$. One can define lower Hessenberg matrices similarly.
- *Outer product matrices:* $A = uv^H$, where both u and v are vectors.

- *Permutation matrices:* the columns of A are a permutation of the columns of the identity matrix.
- *Block diagonal matrices:* generalizes the diagonal matrix by replacing each diagonal entry by a matrix. Notation:

$$A = \text{diag } (A_{11}, A_{22}, \ldots, A_{mm}) .$$

- *Block tri-diagonal matrices:* generalizes the tri-diagonal matrix by replacing each nonzero entry by a square matrix. Notation:

$$A = \text{tridiag } (A_{i,i-1}, A_{ii}, A_{i,i+1}) .$$

The above properties emphasize structure, i.e., positions of the nonzero elements with respect to the zeros, and assume that there are many zero elements or that the matrix is of low rank. No such assumption is made for, say, orthogonal or symmetric matrices.

4. Vector Inner Products and Norms

We define the Hermitian inner product of the two vectors $x = (x_i)_{i=1,\ldots,n}$ and $y = (y_i)_{i=1,\ldots,n}$ of $\mathbb{C}^n$ as the complex number

$$(x, y) = \sum_{i=1}^{n} x_i \bar{y}_i, \tag{1.1}$$

which is often rewritten in matrix notation as

$$(x, y) = y^H x.$$

A vector norm on $\mathbb{C}^n$ is a real-valued function on $\mathbb{C}^n$, which satisfies the following three conditions,

$$\begin{aligned} &\|x\| \geq 0 \quad \forall\ x, \quad \text{and} \quad \|x\| = 0 \text{ iff } x = 0; \\ &\|\alpha x\| = |\alpha| \|x\|, \quad \forall\ x \in \mathbb{C}^n, \quad \forall \alpha \in \mathbb{C}\ ; \\ &\|x + y\| \leq \|x\| + \|y\|, \quad \forall x, y \in \mathbb{C}^n\ . \end{aligned}$$

Associated with the inner product (1.1) is the Euclidean norm of a complex vector defined by

$$\|x\|_2 = (x,x)^{1/2} \ .$$

A fundamental additional property in matrix computations is the simple relation

$$(Ax,y) = (x, A^H y) \quad \forall x,y \in \mathbb{C}^n \tag{1.2}$$

the proof of which is straightforward. The following proposition is a consequence of the above equality.

Proposition 1.2 *Unitary matrices preserve the Hermitian inner product, i.e., $(Qx,Qy) = (x,y)$ for any unitary matrix Q.*

Proof. Indeed $(Qx,Qy) = (x, Q^H Qy) = (x,y)$. ■

In particular a unitary matrix preserves the 2-norm metric, i.e., it is isometric with respect to the 2-norm.

The most commonly used vector norms in numerical linear algebra are special cases of the Hölder norms

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \ . \tag{1.3}$$

Note that the limit of $\|x\|_p$ when p tends to infinity exists and is equal to the maximum modulus of the x_i's. This defines a norm denoted by $\|.\|_\infty$. The cases $p = 1$, $p = 2$, and $p = \infty$ lead to the most important norms in practice,

$$\begin{aligned}
\|x\|_1 &= |x_1| + |x_2| + \cdots + |x_n| \\
\|x\|_2 &= \left[|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2 \right]^{1/2} \\
\|x\|_\infty &= \max_{i=1,..,n} |x_i| \ .
\end{aligned}$$

A useful relation concerning the 2-norm is the so-called Cauchy-Schwartz inequality:

$$|(x,y)| \le \|x\|_2 \|y\|_2.$$

5. Matrix Norms

For a general matrix A in $\mathbb{C}^{n\times m}$ we define a special set of norms of matrices as follows

$$\|A\|_{pq} = \max_{x\in\mathbb{C}^m,\ x\neq 0} \frac{\|Ax\|_p}{\|x\|_q} \ . \tag{1.4}$$

We say that the norms $\|.\|_{pq}$ are induced by the two norms $\|.\|_p$ and $\|.\|_q$. These norms satisfy the usual properties of norms, i.e.,

$$\begin{aligned}
&\|A\| \geq 0 \quad \forall A \ \in \mathbb{C}^{n\times m} \quad \text{and} \quad \|A\| = 0 \quad \text{iff} \quad A = 0 \ ; \\
&\|\alpha A\| = |\alpha|\|A\|, \forall A \ \in \mathbb{C}^{n\times m}, \quad \forall \alpha \in \mathbb{C} \ ; \\
&\|A + B\| \leq \|A\| + \|B\|, \quad \forall A, B \ \in \mathbb{C}^{n\times m} \ .
\end{aligned}$$

Again the most important cases are the ones associated with the cases $p, q = 1, 2, \infty$. The case $q = p$ is of particular interest and the associated norm $\|.\|_{pq}$ is simply denoted by $\|.\|_p$.

A fundamental property of these norms is that

$$\|AB\|_p \leq \|A\|_p\|B\|_p,$$

which is an immediate consequence of the definition (1.4). Matrix norms that satisfy the above property are sometimes called *consistent*. As a result of the above inequality, for example, we have that for any square matrix A,

$$\|A^n\|_p \leq \|A\|_p^n \ ,$$

which implies in particular that the matrix A^n converges to zero if *any* of its p-norms is less than 1.

The Frobenius norm of a matrix is defined by

$$\|A\|_F = \left(\sum_{j=1}^{m} \sum_{i=1}^{n} |a_{ij}|^2 \right)^{1/2} \ . \tag{1.5}$$

This can be viewed as the 2-norm of the column (or row) vector in $\mathbb{C}^{n^2}$ consisting of all the columns (resp. rows) of A listed from

1 to m (resp. 1 to n). It can easily be shown that this norm is also consistent, in spite of the fact that is not induced by a pair of vector norms, i.e., it is not derived from a formula of the form (1.4), see Problem P-1.3. However, it does not satisfy some of the other properties of the p-norms. For example, the Frobenius norm of the identity matrix is not unity. To avoid these difficulties, *we will only use the term matrix norm for a norm that is induced by two norms as in the definition (1.4).* Thus, we will not consider the Frobenius norm to be a proper matrix norm, according to our conventions, even though it is consistent.

It can be shown that the norms of matrices defined above satisfy the following equalities that lead to alternative definitions that are often easier to work with.

$$\|A\|_1 = \max_{j=1,..,m} \sum_{i=1}^{n} |a_{ij}| \; ; \tag{1.6}$$

$$\|A\|_\infty = \max_{i=1,..,n} \sum_{j=1}^{m} |a_{ij}| \; ; \tag{1.7}$$

$$\|A\|_2 = \left[\rho(A^H A)\right]^{1/2} = \left[\rho(AA^H)\right]^{1/2} \; ; \tag{1.8}$$

$$\|A\|_F = \left[\mathrm{tr}(A^H A)\right]^{1/2} = \left[\mathrm{tr}(AA^H)\right]^{1/2} \; . \tag{1.9}$$

As will be shown in Section 5, the eigenvalues of $A^H A$ are nonnegative. Their square roots are called singular values of A and are denoted by $\sigma_i, i = 1, \ldots, m$. Thus, the relation (1.8) shows that $\|A\|_2$ is equal to σ_1, the largest singular value of A.

Example 1.1 From the above properties, it is clear that the spectral radius $\rho(A)$ is equal to the 2-norm of a matrix when the matrix is Hermitian. However, it is not a matrix norm in general. For example, the first property of norms is not satisfied, since for

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

we have $\rho(A) = 0$ while $A \neq 0$. The triangle inequality is also not satisfied for the pair A, and $B = A^T$ where A is defined above. Indeed,

$$\rho(A + B) = 1 \quad \text{while} \quad \rho(A) + \rho(B) = 0.$$

6. Subspaces

A subspace of $\mathbb{C}^n$ is a subset of $\mathbb{C}^n$ that is also a complex vector space. The set of all linear combinations of a set of vectors $G = \{a_1, a_2, ..., a_q\}$ of $\mathbb{C}^n$ is a vector subspace called the linear span of G,

$$\begin{aligned} \mathrm{span}\{G\} &= \mathrm{span}\,\{a_1, a_2, \ldots, a_q\} \\ &= \left\{ z \ \in \mathbb{C}^n \mid z = \sum_{i=1}^{q} \alpha_i a_i \ ; \ \{\alpha\}_{i=1,\ldots,q} \in \mathbb{C}^q \right\} . \end{aligned}$$

If the a_i's are linearly independent, then each vector of span$\{G\}$ admits a unique expression as a linear combination of the a_i's. The set G is then called a basis of the subspace *span*$\{G\}$.

Given two vector subspaces S_1 and S_2, their sum S is a subspace defined as the set of all vectors that are equal to the sum of a vector of S_1 and a vector of S_2. The intersection of two subspaces is also a subspace. If the intersection of S_1 and S_2 is reduced to $\{0\}$ then the sum of S_1 and S_2 is called their direct sum and is denoted by $S = S_1 \bigoplus S_2$. When S is equal to $\mathbb{C}^n$ then every vector x of $\mathbb{C}^n$ can be decomposed in a unique way as the sum of an element x_1 of S_1 and an element x_2 of S_2. The transformation P that maps x into x_1 is a linear transformation that is idempotent ($P^2 = P$). It is called a *projector*, onto S_1 along S_2.

Two subspaces of importance that are associated with a matrix A of $\mathbb{C}^{n\times m}$ are its *range* defined by

$$\mathrm{Ran}(A) = \{Ax \mid x \in \mathbb{C}^m\} \tag{1.10}$$

and its *kernel* or null space

$$\mathrm{Ker}(A) = \{x \in \mathbb{C}^m \mid Ax = 0 \} .$$

The range of A is clearly equal to the linear *span* of its columns. The *rank* of a matrix is equal to the dimension of the range of A.

A subspace S is said to be *invariant* under a (square) matrix A whenever $AS \subset S$. In particular for any eigenvalue λ of A

the subspace $\mathrm{Ker}(A - \lambda I)$ is invariant under A. The subspace $\mathrm{Ker}(A-\lambda I)$ is called the eigenspace associated with λ and consists of all the eigenvectors of A associated with λ and the vector 0.

7. Orthogonal Vectors and Subspaces

A set of vectors $G = \{a_1, a_2, \ldots, a_r\}$ is said to be *orthogonal* if

$$(a_i, a_j) = 0 \quad \text{when} \quad i \neq j$$

It is *orthonormal* if in addition every vector of G has a 2-norm equal to unity. Every subspace admits an orthonormal basis which is obtained by taking any basis and "orthonormalizing" it. The orthonormalization can be achieved by an algorithm referred to as the Gram-Schmidt process which we now describe. Given a set of linearly independent vectors $\{x_1, x_2, \ldots, x_r\}$, we first normalize the vector x_1, i.e., we divide it by its 2-norm, to obtain the scaled vector q_1. Then x_2 is orthogonalized against the vector q_1 by subtracting from x_2 a multiple of q_1 to make the resulting vector orthogonal to q_1, i.e.,

$$x_2 \leftarrow x_2 - (x_2, q_1) q_1 .$$

The resulting vector is again normalized to yield the second vector q_2. The i-th step of the Gram-Schmidt process consists of orthogonalizing the vector x_i against all previous vectors q_j.

ALGORITHM **1.1 Gram-Schmidt**

```
1. Start: Compute $r_{11} := \|x_1\|_2$. If $r_{11} = 0$ stop, else $q_1 := x_1/r_{11}$.
2. Loop: For $j = 2, \ldots, r$ do:
    (a) Compute $r_{ij} := (x_j, q_i)$   for $i = 1, 2, \ldots, j-1$,
    (b) $\hat{q} := x_j - \sum_{i=1}^{j-1} r_{ij} q_i$ ,
    (c) $r_{jj} := \|\hat{q}\|_2$ ,
    (d) If $r_{jj} = 0$ then stop, else $q_j := \hat{q}/r_{jj}$.
```

It is easy to prove that the above algorithm will not break down, i.e., all r steps will be completed, if and only if the family of vectors $x_1, x_2, \ldots, x_r$ is linearly independent. From 2-(b) and 2-(c) it is clear that at every step of the algorithm the following relation holds:

$$x_j = \sum_{i=1}^{j} r_{ij} q_i \ .$$

If we let $X = [x_1, x_2, \ldots, x_r]$, $Q = [q_1, q_2, \ldots, q_r]$, and if R denotes the $r \times r$ upper triangular matrix whose nonzero elements are the r_{ij} defined in the algorithm, then the above relation can be written as

$$X = QR \ . \tag{1.11}$$

This is called the QR decomposition of the $n \times r$ matrix X. Thus, from what was said above the QR decomposition of a matrix exists whenever the column vectors of X form a linearly independent set of vectors.

The above algorithm is the standard Gram-Schmidt process. There are other formulations of the same algorithm which are mathematically equivalent but have better numerical properties. The Modified Gram-Schmidt algorithm (MGSA) is one such alternative.

Algorithm **1.2 Modified Gram-Schmidt**

```
1. Start: define $r_{11} := \|x_1\|_2$. If $r_{11} = 0$ stop, else $q_1 := x_1/r_{11}$.
2. Loop: For $j = 2, \ldots, r$ do:
    (a) Define $\hat{q} := x_j$,
    (b) For $i = 1, \ldots, j-1$, do $\left\{ \begin{array}{l} r_{ij} := (\hat{q}, q_i) \\ \hat{q} := \hat{q} - r_{ij} q_i \end{array} \right.$
    (c) Compute $r_{jj} := \|\hat{q}\|_2$,
```

(d) *If* $r_{jj} = 0$ *then stop, else* $q_j := \hat{q}/r_{jj}$.

A vector that is orthogonal to all the vectors of a subspace S is said to be orthogonal to that subspace. The set of all the vectors that are orthogonal to S is a vector subspace called the *orthogonal complement* of S and denoted by $S^{\perp}$. The space $\mathbb{C}^n$ is the direct sum of S and its orthogonal complement. The projector onto S along its orthogonal complement is called an *orthogonal projector* onto S. If $V = [v_1, v_2, \ldots, v_r]$ is an orthonormal matrix then $V^H V = I$, i.e., V is orthogonal. However, VV^H is not the identity matrix but represents the orthogonal projector onto span$\{V\}$, see Section 1 of Chapter III for details.

8. Canonical Forms of Matrices

In this section we will be concerned with the reduction of square matrices into matrices that have simpler forms, such as diagonal or bidiagonal, or triangular. By reduction we mean a transformation that preserves the eigenvalues of a matrix.

Definition 1.2 *Two matrices A and B are said to be similar if there is a nonsingular matrix X such that*

$$A = XBX^{-1}$$

The mapping $B \rightarrow A$ is called a similarity transformation.

It is clear that *similarity* is an equivalence relation. Similarity transformations preserve the eigenvalues of matrix. An eigenvector u_B of B is transformed into the eigenvector $u_A = Xu_B$ of A. In effect, a similarity transformation amounts to representing the matrix B in a different basis.

We now need to define some terminology.

1. An eigenvalue λ of A is said to have *algebraic multiplicity* μ if it is a root of multiplicity μ of the characteristic polynomial.

2. If an eigenvalue is of algebraic multiplicity one it is said to be *simple*. A nonsimple eigenvalue is said to be *multiple*.

3. An eigenvalue λ of A is said to have *geometric multiplicity* γ if the maximum number of independent eigenvectors associated with it is γ. In other words the geometric multiplicity γ is the dimension of the eigenspace Ker $(A - \lambda I)$.

4. A matrix is said to be *derogatory* if the geometric multiplicity of at least one of its eigenvalues is larger than one.

5. An eigenvalue is said to be *semi-simple* if its algebraic multiplicity is equal to its geometric multiplicity. An eigenvalue that is not semi-simple is called *defective* .

We will often denote by $\lambda_1, \lambda_2, \ldots, \lambda_p$, $(p \leq n)$, all the *distinct* eigenvalues of A. It is a simple exercise to show that the characteristic polynomials of two similar matrices are identical, see Exercise P-1.7. Therefore, the eigenvalues of two similar matrices are equal and so are their algebraic multiplicities. Moreover if v is an eigenvector of B then Xv is an eigenvector of A and, conversely, if y is an eigenvector of A then $X^{-1}y$ is an eigenvector of B. As a result the number of independent eigenvectors associated with a given eigenvalue is the same for two similar matrices, i.e., their geometric multiplicity is also the same.

The possible desired forms are numerous but they all have the common goal of attempting to simplify the original eigenvalue problem. Here are some possibilities with comments as to their usefulness.

- *Diagonal:* the simplest and certainly most desirable choice but it is not always achievable.

- *Jordan:* this is an upper bidiagonal matrix with ones or zeroes on the super diagonal. Always possible but not numerically trustworthy.

- *Upper triangular:* in practice this is the most reasonable compromise as the similarity from the original matrix to a triangular form can be chosen to be isometric and therefore the transformation can be achieved via a sequence of elementary unitary transformations which are numerically stable.

8.1. Reduction to the Diagonal Form.

The simplest form in which a matrix can be reduced is undoubtedly the diagonal form but this reduction is, unfortunately, not always possible. A matrix that can be reduced to the diagonal form is called diagonalizable. The following theorem characterizes such matrices.

Theorem 1.1 *A matrix of dimension n is diagonalizable if and only if it has n linearly independent eigenvectors.*

Proof. A matrix A is diagonalizable if and only if there exists a nonsingular matrix X and a diagonal matrix D such that $A = XDX^{-1}$ or equivalently $AX = XD$, where D is a diagonal matrix. This is equivalent to saying that there exist n linearly independent vectors – the n column-vectors of X – such that $Ax_i = d_i x_i$, i.e., each of these column-vectors is an eigenvector of A. ■

A matrix that is diagonalizable has only semi-simple eigenvalues. Conversely, if all the eigenvalues of a matrix are semi-simple then there exist n eigenvectors of the matrix A. It can be easily shown that these eigenvectors are linearly independent, see Exercise P-1.1. As a result we have the following proposition.

Proposition 1.3 *A matrix is diagonalizable if and only if all its eigenvalues are semi-simple.*

Since every simple eigenvalue is semi-simple, an immediate corollary of the above result is that when A has n distinct eigenvalues then it is diagonalizable.

8.2. The Jordan Canonical Form

From the theoretical viewpoint, one of the most important canonical forms of matrices is the well-known Jordan form. In what follows, the main constructive steps that lead to the Jordan canonical decomposition are outlined. For details, the reader is referred to a standard book on matrix theory or linear algebra.

• For every integer l and each eigenvalue λ_i it is true that

$$\mathrm{Ker}(A - \lambda_i I)^{l+1} \supset \mathrm{Ker}(A - \lambda_i I)^l \ .$$

• Because we are in a finite dimensional space the above property implies that there is a first integer l_i such that

$$\mathrm{Ker}(A - \lambda_i I)^{l_i+1} = \mathrm{Ker}(A - \lambda_i I)^{l_i},$$

and in fact $\mathrm{Ker}(A - \lambda_i I)^l = \mathrm{Ker}(A - \lambda_i I)^{l_i}$ for all $l \geq l_i$. The integer l_i is called the index of λ_i.

• The subspace $M_i = \mathrm{Ker}(A - \lambda_i I)^{l_i}$ is invariant under A. Moreover, the space $\mathbb{C}^n$ is the direct sum of the subspaces M_i's, for $i = 1, 2, \ldots, p$. Let $m_i = \dim(M_i)$.

• In each invariant subspace M_i there are γ_i independent eigenvectors, i.e., elements of $\mathrm{Ker}(A - \lambda_i I)$, with $\gamma_i \leq m_i$. It turns out that this set of vectors can be completed to form a basis of M_i by adding to it elements of $\mathrm{Ker}(A - \lambda_i I)^2$, then elements of $\mathrm{Ker}(A - \lambda_i I)^3$, and so on. These elements are generated by starting separately from each eigenvector u, i.e., an element of $\mathrm{Ker}(A - \lambda_i I)$, and then seeking an element that satisfies $(A - \lambda_i I)z_1 = u$. Then, more generally we construct z_{i+1} by solving the equation $(A - \lambda_i I)z_{i+1} = z_i$ when possible. The vector z_i belongs to $\mathrm{Ker}(A - \lambda_i I)^{i+1}$ and is called a principal vector (sometimes generalized eigenvector). The process is continued until no more principal vectors are found. There are at most l_i principal vectors for each of the γ_i eigenvectors.

- The final step is to represent the original matrix A with respect to the basis made up of the p bases of the invariant subspaces M_i defined in the previous step.

The matrix representation J of A in the new basis described above has the block diagonal structure,

$$X^{-1}AX = J = \begin{pmatrix} J_1 & & & & & \\ & J_2 & & & & \\ & & \ddots & & & \\ & & & J_i & & \\ & & & & \ddots & \\ & & & & & J_p \end{pmatrix}$$

where each J_i corresponds to the subspace M_i associated with the eigenvalue λ_i. It is of size m_i and it has itself the following structure,

$$J_i = \begin{pmatrix} J_{i1} & & & \\ & J_{i2} & & \\ & & \ddots & \\ & & & J_{i\gamma_i} \end{pmatrix} \text{ with } J_{ik} = \begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{pmatrix}.$$

Each of the blocks J_{ik} corresponds to a different eigenvector associated with the eigenvalue λ_i. Its size is equal to the number of principal vectors found for the eigenvector to which the block is associated and does not exceed l_i.

Theorem 1.2 *Any matrix A can be reduced to a block diagonal matrix consisting of p diagonal blocks, each associated with a distinct eigenvalue. Each diagonal block number i has itself a block diagonal structure consisting of γ_i subblocks, where γ_i is the geometric multiplicity of the eigenvalue λ_i. Each of the subblocks, referred to as a Jordan block, is an upper bidiagonal matrix of size not exceeding l_i, with the constant λ_i on the diagonal and the constant one on the super diagonal.*

We refer to the i-th diagonal block, $i = 1, \ldots, p$ as the i-th Jordan submatrix (sometimes "Jordan Box"). The Jordan submatrix number i starts in column $j_i \equiv m_1 + m_2 + \cdots + m_{i-1} + 1$. From the above form it is not difficult to see that $M_i = \text{Ker}(A - \lambda_i I)^{l_i}$ is merely the span of the columns $j_i, j_i + 1, \ldots, j_{i+1} - 1$ of the matrix X. These vectors are all the eigenvectors and the principal vectors associated with the eigenvalue λ_i.

Since A and J are similar matrices their characteristic polynomials are identical. Hence, it is clear that the algebraic multiplicity of an eigenvalue λ_i is equal to the dimension of M_i:

$$\mu_i = m_i \equiv \dim\,(M_i)\ .$$

As a result,

$$\mu_i \geq \gamma_i.$$

Because $\mathbb{C}^n$ is the direct sum of the subspaces $M_i, i = 1, \ldots, p$ each vector x can be written in a unique way as

$$x = x_1 + x_2 + \cdots + x_i + \cdots + x_p,$$

where x_i is a member of the subspace M_i. The linear transformation defined by

$$P_i : x \rightarrow x_i$$

is a projector onto M_i along the direct sum of the subspaces $M_j, j \neq i$. The family of projectors $P_i, i = 1, \ldots, p$ satisfies the following properties,

$$\text{Ran}(P_i) = M_i \tag{1.12}$$

$$P_i P_j = P_j P_i = 0, \text{ if } i \neq j \tag{1.13}$$

$$\sum_{i=1}^{p} P_i = I \tag{1.14}$$

In fact it is easy to see that the above three properties define a decomposition of $\mathbb{C}^n$ into a direct sum of the images of the projectors P_i in a unique way. More precisely, any family of projectors that satisfies the above three properties is uniquely determined and is associated with the decomposition of $\mathbb{C}^n$ into the direct sum of the images of the P_i 's.

It is helpful for the understanding of the Jordan canonical form to determine the matrix representation of the projectors P_i. Consider the matrix $\hat{J}_i$ which is obtained from the Jordan matrix by replacing all the diagonal submatrices by zero blocks except the i^{th} submatrix which is replaced by the identity matrix.

$$\hat{J}_i = \begin{pmatrix} 0 & & & & \\ & 0 & & & \\ & & I & & \\ & & & 0 & \\ & & & & 0 \end{pmatrix}$$

In other words if each i-th Jordan submatrix starts at the column number j_i, then the columns of $\hat{J}_i$ will be zero columns except columns $j_i, \ldots, j_{i+1} - 1$ which are the corresponding columns of the identity matrix. Let $\hat{P}_i = X \hat{J}_i X^{-1}$. Then it is not difficult to verify that $\hat{P}_i$ is a projector and that,

1. The range of $\hat{P}_i$ is the span of columns $j_i, \ldots, j_{i+1} - 1$ of the matrix X. This is the same subspace as M_i.
2. $\hat{P}_i \hat{P}_j = \hat{P}_j \hat{P}_i = 0$ whenever $i \neq j$
3. $\hat{P}_1 + \hat{P}_2 + \cdots + \hat{P}_p = I$

According to our observation concerning the uniqueness of a family of projectors that satisfy (1.12) - (1.14) this implies that

$$\hat{P}_i = P_i \quad , \quad i = 1, \ldots, p$$

Example 1.2 Let us assume that the eigenvalue λ_i is simple. Then,

$$P_i = X e_i e_i^H X^{-1} \equiv u_i w_i^H,$$

in which we have defined $u_i = X e_i$ and $w_i = X^{-H} e_i$. It is easy to show that u_i and w_i are right and left eigenvectors, respectively, associated with λ_i and normalized so that $w_i^H u_i = 1$.

Consider now the matrix $\hat{D}_i$ obtained from the Jordan form of A by replacing each Jordan submatrix by a zero matrix except the i-th submatrix which is obtained by zeroing its diagonal elements, i.e.,

$$\hat{D}_i = \begin{pmatrix} 0 & & & & & \\ & 0 & & & & \\ & & \ddots & & & \\ & & & J_i - \lambda_i I & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

Define $D_i = X\hat{D}_i X^{-1}$. Then it is a simple exercise to show by means of the explicit expression for $\hat{P}_i$, that

$$D_i = (A - \lambda_i I)P_i. \tag{1.15}$$

Moreover, $D_i^{l_i} = 0$, i.e., D_i is a *nilpotent matrix* of index l_i. We are now ready to state the following important theorem which can be viewed as an alternative mathematical formulation of Theorem 1.2 on Jordan forms.

Theorem 1.3 *Every square matrix A admits the decomposition*

$$A = \sum_{i=1}^{p}(\lambda_i P_i + D_i) \tag{1.16}$$

where the family of projectors $\{P_i\}_{i=1,\ldots,p}$ satisfies the conditions (1.12), (1.13), and (1.14), and where $D_i = (A - \lambda_i I)P_i$ is a nilpotent operator of index l_i.

Proof. From (1.15), we have

$$AP_i = \lambda_i P_i + D_i \quad i = 1, 2, \ldots, p$$

Summing up the above equalities for $i = 1, 2, \ldots, p$ we get

$$A\sum_{i=1}^{p} P_i = \sum_{i=1}^{p}(\lambda_i P_i + D_i)$$

The proof follows by substituting (1.14) into the left-hand-side. ■

The projector P_i is called the *spectral projector* associated with the eigenvalue λ_i. The linear operator D_i is called the *nilpotent* associated with λ_i. The decomposition (1.16) is referred to as the spectral decomposition of A. Additional properties that are easy to prove from the various expressions of P_i and D_i are the following

$$
\begin{aligned}
& P_i D_j = D_j P_i = \delta_{ij} P_i && (1.17)\\
& AP_i = P_i A = P_i A P_i = \lambda_i P_i + D_i && (1.18)\\
& A^k P_i = P_i A^k = P_i A^k P_i = \\
& \qquad P_i(\lambda_i I + D_i)^k = (\lambda_i I + D_i)^k P_i && (1.19)\\
& AP_i = [x_{j_i}, \ldots, x_{j_{i+1}-1}] B_i [y_{j_i}, \ldots, y_{j_{i+1}-1}]^H && (1.20)
\end{aligned}
$$

where B_i is the i-th Jordan submatrix and where the columns y_j are the columns of the matrix X^{-H}.

Corollary 1.1 *For any matrix norm $\|.\|$, the following relation holds*

$$\lim_{k \to \infty} \|A^k\|^{1/k} \;=\; \rho(A) \; . \tag{1.21}$$

Proof. The proof of this corollary is the subject of exercise P-1.8. ■

Another way of stating the above corollary is that there is a sequence ϵ_k such that

$$\|A^k\| = (\rho(A) + \epsilon_k)^k$$

where $\lim_{k \to \infty} \epsilon_k = 0$.

8.3. The Schur Canonical Form

We will now show that any matrix is unitarily similar to an upper triangular matrix. The only result needed to prove the following theorem is that any vector of 2-norm one can be completed by $n-1$ additional vectors to form an orthonormal basis of $\mathbb{C}^n$.

Theorem 1.4 *For any given matrix A there exists a unitary matrix Q such that $Q^H AQ = R$ is upper triangular.*

Proof. The proof is by induction over the dimension n. The result is trivial for $n = 1$. Let us assume that it is true for $n-1$ and consider any matrix A of size n. The matrix admits at least one eigenvector u that is associated with an eigenvalue λ. We assume without loss of generality that $\|u\|_2 = 1$. We can complete the vector u into an orthonormal set, i.e., we can find an $n \times (n-1)$ matrix V such that the $n \times n$ matrix $U = [u, V]$ is unitary. Then we have $AU = [\lambda u, AV]$ and hence,

$$U^H AU = \begin{bmatrix} u^H \\ V^H \end{bmatrix} \quad [\lambda u, AV] = \begin{pmatrix} \lambda & u^H AV \\ 0 & V^H AV \end{pmatrix} \tag{1.22}$$

We now use our induction hypothesis for the $(n-1) \times (n-1)$ matrix $B = V^H AV$: there exists an $(n-1) \times (n-1)$ unitary matrix Q_1 such that $Q_1^H BQ_1 = R_1$ is upper triangular. Let us define the $n \times n$ matrix

$$\hat{Q}_1 = \begin{pmatrix} 1 & 0 \\ 0 & Q_1 \end{pmatrix}$$

and multiply both members of (1.22) by $\hat{Q}_1^H$ from the left and $\hat{Q}_1$ from the right. The resulting matrix is clearly upper triangular and this shows that the result is true for A, with $Q = \hat{Q}_1 U$ which is a unitary $n \times n$ matrix. ■

A simpler proof that uses the Jordan canonical form and the QR decomposition is the subject of Exercise P-1.5. Since the matrix

R is triangular and similar to A, its diagonal elements are equal to the eigenvalues of A ordered in a certain manner. In fact it is easy to extend the proof of the theorem to show that we can obtain this factorization with *any order* we want for the eigenvalues. One might ask the question as to which order might be best numerically but the answer to the question goes beyond the scope of this book. Despite its simplicity, the above theorem has far reaching consequences some of which will be examined in the next section.

It is important to note that for any $k \leq n$ the subspace spanned by the first k columns of Q is invariant under A. This is because from the Schur decomposition we have, for $1 \leq j \leq k$,

$$Aq_j = \sum_{i=1}^{i=j} r_{ij} q_i \ .$$

In fact, letting $Q_k = [q_1, q_2, \ldots, q_k]$ and R_k be the principal leading submatrix of dimension k of R, the above relation can be rewritten as

$$AQ_k = Q_k R_k$$

which we refer to as the partial Schur decomposition of A. The simplest case of this decomposition is when $k = 1$, in which case q_1 is an eigenvector. The vectors q_i are usually referred to as Schur vectors. Note that the Schur vectors are not unique and in fact they depend on the order chosen for the eigenvalues.

A slight variation on the Schur canonical form is the quasi Schur form, also referred to as the real Schur form. Here, diagonal blocks of size 2 x 2 are allowed in the upper triangular matrix R. The reason for this is to avoid complex arithmetic when the original matrix is real. A 2×2 block is associated with each complex conjugate pair of eigenvalues of the matrix.

Example 1.3 Consider the 3×3 matrix

$$A = \begin{pmatrix} 1 & 10 & 0 \\ -1 & 3 & 1 \\ -1 & 0 & 1 \end{pmatrix}$$

The matrix A has the pair of complex conjugate eigenvalues

$$2.4069.. \pm i \times 3.2110..$$

and the real eigenvalue 0.1863... The standard (complex) Schur form is given by the pair of matrices

$$V = \begin{pmatrix} 0.3381 - 0.8462i & 0.3572 - 0.1071i & 0.1749 \\ 0.3193 - 0.0105i & -0.2263 - 0.6786i & -0.6214 \\ 0.1824 + 0.1852i & -0.2659 - 0.5277i & 0.7637 \end{pmatrix}$$

and

$$S = \begin{pmatrix} 2.4069 + 3.2110i & 4.6073 - 4.7030i & -2.3418 - 5.2330i \\ 0 & 2.4069 - 3.2110i & -2.0251 - 1.2016i \\ 0 & 0 & 0.1863 \end{pmatrix} .$$

It is possible to avoid complex arithmetic by using the quasi-Schur form which consists of the pair of matrices

$$U = \begin{pmatrix} -0.9768 & 0.1236 & 0.1749 \\ -0.0121 & 0.7834 & -0.6214 \\ 0.2138 & 0.6091 & 0.7637 \end{pmatrix}$$

and

$$R = \begin{pmatrix} 1.3129 & -7.7033 & 6.0407 \\ 1.4938 & 3.5008 & -1.3870 \\ 0 & 0 & 0.1863 \end{pmatrix} .$$

We would like to conclude this section by pointing out that the Schur and the quasi Schur forms of a given matrix are in no way unique. In addition to the dependence on the ordering of the eigenvalues, any column of Q can be multiplied by a complex sign $e^{i\theta}$ and a new corresponding R can be found. For the quasi Schur form there are infinitely many ways of selecting the 2×2 blocks, corresponding to applying arbitrary rotations to the columns of Q associated with these blocks.

9. Normal and Hermitian Matrices

In this section we look at the specific properties of normal matrices and Hermitian matrices regarding among other things their spectra and some important optimality properties of their eigenvalues. The most common normal matrices that arise in practice are Hermitian or skew-Hermitian. In fact, symmetric real matrices form a large part of the matrices that arise in practical eigenvalue problems.

9.1. Normal Matrices

By definition a matrix is said to be normal if it satisfies the relation

$$A^H A = AA^H. \tag{1.23}$$

An immediate property of normal matrices is stated in the following proposition.

Proposition 1.4 *If a normal matrix is triangular then it is necessarily a diagonal matrix.*

Proof. Assume for example that A is upper triangular and normal and let us compare the first diagonal element of the left hand side matrix of (1.23) with the corresponding element of the matrix on the right hand side. We obtain that

$$|a_{11}|^2 = \sum_{j=1}^{n} |a_{1j}|^2,$$

which shows that the elements of the first row are zeros except for the diagonal one. The same argument can now be used for the second row, the third row, and so on to the last row, to show that $a_{ij} = 0$ for $i \neq j$. ■

As a consequence of this we have the following important result.

Theorem 1.5 *A matrix is normal if and only if it is unitarily similar to a diagonal matrix.*

Proof. It is straightforward to verify that a matrix which is unitarily similar to a diagonal matrix is normal. Let us now show that any normal matrix A is unitarily similar to a diagonal matrix. Let $A = QRQ^H$ be the Schur canonical form of A where we recall that Q is unitary and R is upper triangular. By the normality of A we have

$$QR^HQ^HQRQ^H = QRQ^HQR^HQ^H$$

or,

$$QR^HRQ^H = QRR^HQ^H$$

Upon multiplication by Q^H on the left and Q on the right this leads to the equality $R^HR = RR^H$ which means that R is normal, and according to the previous proposition this is only possible if R is diagonal. ■

Thus, any normal matrix is diagonalizable and admits an orthonormal basis of eigenvectors, namely the column vectors of Q.

Clearly, Hermitian matrices are just a particular case of normal matrices. Since a normal matrix satisfies the relation $A = QDQ^H$, with D diagonal and Q unitary, the eigenvalues of A are the diagonal entries of D. Therefore, if these entries are real it is clear that we will have $A^H = A$. This is restated in the following corollary.

Corollary 1.2 *A normal matrix whose eigenvalues are real is Hermitian.*

As will be seen shortly the converse is also true, in that a Hermitian matrix has real eigenvalues.

An eigenvalue λ of any matrix satisfies the relation

$$\lambda = \frac{(Au, u)}{(u, u)}$$

where u is an associated eigenvector. More generally one might consider the complex scalars,

$$\mu(x) = \frac{(Ax, x)}{(x, x)} \tag{1.24}$$

defined for any nonzero vector in $\mathbb{C}^n$. These ratios are referred to as *Rayleigh quotients* and are important both from theoretical and practical purposes. The set of all possible Rayleigh quotients as x runs over $\mathbb{C}^n$ is called the *field of values* of A. This set is clearly bounded since each $|\mu(x)|$ is bounded by the the 2-norm of A, i.e., $|\mu(x)| \leq \|A\|_2$ for all x.

If a matrix is normal then any vector x in $\mathbb{C}^n$ can be expressed as

$$\sum_{i=1}^{n} \xi_i q_i$$

where the vectors q_i form an orthogonal basis of eigenvectors, and the expression for $\mu(x)$ becomes,

$$\mu(x) = \frac{(Ax, x)}{(x, x)} = \frac{\sum_{k=1}^{n} \lambda_k |\xi_k|^2}{\sum_{k=1}^{n} |\xi_k|^2} \equiv \sum_{k=1}^{n} \beta_k \lambda_k \tag{1.25}$$

where

$$0 \leq \beta_i = \frac{|\xi_i|^2}{\sum_{k=1}^{n} |\xi_k|^2} \leq 1 \ , \quad \text{and} \quad \sum_{i=1}^{n} \beta_i = 1$$

From a well-known characterization of convex hulls due to Hausdorff, (Hausdorff's convex hull theorem) this means that the set of all possible Rayleigh quotients as x runs over all of $\mathbb{C}^n$ is equal to the convex hull of the λ_i's. This leads to the following theorem.

Theorem 1.6 *The field of values of a normal matrix is equal to the convex hull of its spectrum.*

The question that arises next is whether or not this is also true for non-normal matrices and the answer is no, i.e., the convex hull of the eigenvalues and the field of values of a non-normal matrix are different in general, see Exercise P-1.10 for an example. As a generic example, one can take any nonsymmetric real matrix that has real eigenvalues only; its field of values will contain imaginary values. It has been shown (Hausdorff) that the field of values of a matrix is a convex set. Since the eigenvalues are members of the field of values, their convex hull is contained in the field of values. This is summarized in the following proposition.

Proposition 1.5 *The field of values of an arbitrary matrix is a convex set which contains the convex hull of its spectrum. It is equal to the convex hull of the spectrum when the matrix in normal.*

9.2. Hermitian Matrices

A first and important result on Hermitian matrices is the following.

Theorem 1.7 *The eigenvalues of a Hermitian matrix are real, i.e.,* $\sigma(A) \subset \mathbb{R}$.

Proof. Let λ be an eigenvalue of A and u an associated eigenvector or 2-norm unity. Then

$$\lambda = (Au, u) = (u, Au) = \overline{(Au, u)} = \overline{\lambda}$$

■

Moreover, it is not difficult to see that if, in addition, the matrix is real then the eigenvectors can be chosen to be real, see Exercise P-1.16. Since a Hermitian matrix is normal an immediate consequence of Theorem 1.5 is the following result.

Theorem 1.8 *Any Hermitian matrix is unitarily similar to a real diagonal matrix.*

In particular a Hermitian matrix admits a set of orthonormal eigenvectors that form a basis of $\mathbb{C}^n$.

In the proof of Theorem 1.6 we used the fact that the inner products (Au, u) are real. More generally it is clear that any Hermitian matrix is such that (Ax, x) is real for any vector $x \in \mathbb{C}^n$. It turns out that the converse is also true, i.e., it can be shown that if (Az, z) is real for all vectors z in $\mathbb{C}^n$ then the matrix A is Hermitian, see Problem P-1.14.

Eigenvalues of Hermitian matrices can be characterized by optimality properties of the Rayleigh quotients (1.24). The best known of these is the Min-Max principle. Let us order all the eigenvalues of A in descending order:

$$\lambda_1 \ \geq \ \lambda_2 \ \ldots \ \geq \ \lambda_n .$$

Here the eigenvalues are not necessarily distinct and they are repeated, each according to its multiplicity. In what follows, we denote by S a generic subspace of $\mathbb{C}^n$. Then we have the following theorem.

Theorem 1.9 (Min-Max theorem) *The eigenvalues of a Hermitian matrix A are characterized by the relation*

$$\lambda_k = \min_{S,\ \dim\ (S)=n-k+1} \ \max_{x \in S, x \neq 0} \frac{(Ax, x)}{(x, x)} \tag{1.26}$$

Proof. Let $\{q_i\}_{i=1,\ldots,n}$ be an orthonormal basis of $\mathbb{C}^n$ consisting of eigenvectors of A associated with $\lambda_1, \ldots, \lambda_n$ respectively. Let S_k be the subspace spanned by the first k of these vectors and denote by $\mu(S)$ the maximum of $(Ax, x)/(x, x)$ over all nonzero vectors of a subspace S. Since the dimension of S_k is k, a well-known theorem of linear algebra shows that its intersection with any subspace S of dimension $n - k + 1$ is not reduced to $\{0\}$, i.e., there is vector x in $S \bigcap S_k$. For this $x = \sum_{i=1}^{k} \xi_i q_i$ we have

$$\frac{(Ax, x)}{(x, x)} \ = \ \frac{\sum_{i=1}^{k} \lambda_i |\xi_i|^2}{\sum_{i=1}^{k} |\xi_i|^2} \quad \geq \quad \lambda_k$$

so that $\mu(S) \ \geq \ \lambda_k$.

Consider on the other hand the particular subspace S_0 of dimension $n-k+1$ which is spanned by $q_k, \ldots, q_n$. For each vector x in this subspace we have

$$\frac{(Ax,x)}{(x,x)} = \frac{\sum_{i=k}^n \lambda_i |\xi_i|^2}{\sum_{i=k}^n |\xi_i|^2} \quad \leq \quad \lambda_k$$

so that $\mu(S_0) \leq \lambda_k$. In other words, as S runs over all $n-k+1$-dimensional subspaces $\mu(S)$ is always $\ \geq \ \lambda_k$ and there is at least one subspace S_0 for which $\mu(S_0) \leq \lambda_k$ which shows the result. ■

This result is attributed to Courant and Fisher, and to Poincaré and Weyl. It is often referred to as Courant-Fisher min-max principle or theorem. As a particular case, the largest eigenvalue of A satisfies

$$\lambda_1 = \max_{x \neq 0} \frac{(Ax,x)}{(x,x)} \ . \tag{1.27}$$

Actually, there are four different ways of rewriting the above characterization. The second formulation is

$$\lambda_k = \max_{S, \ \dim\ (S)=k} \ \min_{x \in S, x \neq 0} , \ \frac{(Ax,x)}{(x,x)} \tag{1.28}$$

and the two other ones can be obtained from the above two formulations by simply relabeling the eigenvalues increasingly instead of decreasingly. Thus, with our labeling of the eigenvalues in descending order, (1.28) tells us that the smallest eigenvalue satisfies,

$$\lambda_n = \min_{x \neq 0} \frac{(Ax,x)}{(x,x)} \ .$$

with λ_n replaced by λ_1 if the eigenvalues are relabeled increasingly.

In order for all the eigenvalues of a Hermitian matrix to be positive it is necessary and sufficient that

$$(Ax,x) > 0, \quad \forall \ x \in \mathbb{C}^n, \quad x \neq 0.$$

Such a matrix is called *positive definite.* A matrix that satisfies $(Ax, x) \geq 0$ for any x is said to be *positive semi-definite.* In particular the matrix $A^H A$ is semi-positive definite for any rectangular matrix, since

$$(A^H Ax, x) = (Ax, Ax) \ \geq \ 0 \quad \forall \ x.$$

Similarly, AA^H is also a Hermitian semi-positive definite matrix. The square roots of the eigenvalues of $A^H A$ for a general rectangular matrix A are called the *singular values* of A and are denoted by σ_i. In Section 1.5 we have stated without proof that the 2-norm of any matrix A is equal to the largest singular value σ_1 of A. This is now an obvious fact, because

$$\|A\|_2^2 = \max_{x\neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} \ = \ \max_{x\neq 0} \frac{(Ax, Ax)}{(x, x)} \ = \ \max_{x\neq 0} \frac{(A^H Ax, x)}{(x, x)} \ = \ \sigma_1^2$$

which results from (1.27).

Another characterization of eigenvalues, known as the Courant characterization, is stated in the next theorem. In contrast with the min-max theorem this property is recursive in nature.

Theorem 1.10 *The eigenvalue λ_i and the corresponding eigenvector q_i of a Hermitian matrix are such that*

$$\lambda_1 = \frac{(Aq_1, q_1)}{(q_1, q_1)} \ = \ \max_{x\in \mathbb{C}^n, x\neq 0} \frac{(Ax, x)}{(x, x)}$$

and for $k > 1$:

$$\lambda_k = \frac{(Aq_k, q_k)}{(q_k, q_k)} \ = \ \max_{x\neq 0, q_1^H x = \ldots = q_{k-1}^H x = 0} \frac{(Ax, x)}{(x, x)} \ . \qquad (1.29)$$

In other words, the maximum of the Rayleigh quotient over a subspace that is orthogonal to the first $k-1$ eigenvectors is equal to λ_k and is achieved for the eigenvector q_k associated with λ_k. The proof follows easily from the expansion (1.25) of the Rayleigh quotient.

10. Nonnegative Matrices

A nonnegative matrix is a matrix whose entries are nonnegative,

$$a_{ij} \geq 0 \ .$$

Nonnegative matrices arise in many applications and play a crucial role in the theory of matrices. They play for example a key role in the analysis of convergence of iterative methods for partial differential equations. They also arise in economics, queuing theory, chemical engineering, etc..

A matrix is said to be reducible if, there is a permutation matrix P such that PAP^T is block upper triangular. An important result concerning nonnegative matrices is the following theorem known as the Perron-Frobenius theorem.

Theorem 1.11 *Let A be a real $n \times n$ nonnegative irreducible matrix. Then $\lambda \equiv \rho(A)$, the spectral radius of A, is a simple eigenvalue of A. Moreover, there exists an eigenvector u with positive elements associated with this eigenvalue.*

Problems

P-1.1 Show that two eigenvectors associated with two distinct eigenvalues are linearly independent. More generally show that a family of eigenvectors associated with distinct eigenvalues forms a linearly independent family.

P-1.2 Show that if λ is any eigenvalue of the matrix AB then it is also an eigenvalue of the matrix BA. Start with the particular case where A and B are square and B is nonsingular then consider the more general case where A, B may be singular or even rectangular (but such that AB and BA are square).

P-1.3 Show that the Frobenius norm is consistent. Can this norm be associated to two vector norms via (1.4)? What is the Frobenius norm of a diagonal matrix? What is the p-norm of a diagonal matrix (for any p)?

P-1.4 Find the Jordan canonical form of the matrix:

$$A = \begin{pmatrix} 1 & 2 & -4 \\ 0 & 1 & 2 \\ 0 & 0 & 2 \end{pmatrix}.$$

Same question for the matrix obtained by replacing the element a_{33} by 1.

P-1.5 Give an alternative proof of Theorem 1.4 on the Schur form by starting from the Jordan canonical form. [Hint: write $A = XJX^{-1}$ and use the QR decomposition of X.]

P-1.6 Show from the definition of determinants used in Section (1.2) that the characteristic polynomial is a polynomial of degree n for an $n \times n$ matrix.

P-1.7 Show that the characteristic polynomials of two similar matrices are equal.

P-1.8 Show that

$$\lim_{k \to \infty} \|A^k\|^{1/k} = \rho(A),$$

for any matrix norm. [Hint: use the Jordan canonical form or Theorem 1.3]

P-1.9 Let X be a nonsingular matrix and, for any matrix norm $\|.\|$, define $\|A\|_X = \|AX\|$. Show that this is indeed a matrix norm. Is this matrix norm consistent? Similar questions for $\|XA\|$ and $\|YAX\|$ where Y is also a nonsingular matrix. These norms are not, in general, associated with any vector norms, i.e., they can't be defined by a formula of the form (1.4). Why? What about the particular case $\|A\|' = \|XAX^{-1}\|$?

P-1.10 Find the field of values of the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

and verify that it is not equal to the convex hull of its eigenvalues.

P-1.11 Show that any matrix can be written as the sum of a Hermitian and a skew-Hermitian matrix (or the sum of a symmetric and a skew-symmetric matrix).

P-1.12 Show that for a skew-Hermitian matrix S, we have

$$\Re e(Sx, x) = 0 \quad \text{for any } x \in \mathbb{C}^n.$$

P-1.13 Given an arbitrary matrix S, show that if $(Sx, x) = 0$ for all x in $\mathbb{C}^n$ then we must have

$$(Sy, z) + (Sz, y) = 0 \quad \forall\ y\ ,\ z\ \in\ \mathbb{C}^n.$$

[Hint: expand $(S(y+z), y+z)$].

P-1.14 Using the result of the previous two problems, show that if (Ax, x) is real for all x in $\mathbb{C}^n$, then A must be Hermitian. Would this result be true if we were to replace the assumption by: (Ax, x) *is real for all real* x? Explain.

P-1.15 The definition of a positive definite matrix is that (Ax, x) be real and positive for all real vectors x. Show that this is equivalent to requiring that the Hermitian part of A, namely $\frac{1}{2}(A + A^H)$, be (Hermitian) positive definite.

P-1.16 Let A be a real symmetric matrix and λ an eigenvalue of A. Show that if u is an eigenvector associated with λ then so is $\bar{u}$. As a result, prove that for any eigenvalue of a real symmetric matrix, there is an associated eigenvector which is real.

P-1.17 Show that a Hessenberg matrix H such that $h_{j+1,j} \neq 0, j = 1, 2, \ldots, n-1$ cannot be derogatory.

NOTES AND REFERENCES. For additional reading on the material presented in this Chapter, see Golub and Van Loan [63] and Stewart [167]. More details on matrix eigenvalue problems can be found in Gantmacher's book [54] and Wilkinson [183]. Stewart and Sun's recent book [172] devotes a separate chapter to matrix norms and contains a wealth of information. Some of the terminology we used is borrowed from Chatelin [14, 15] and Kato [85]. For a good overview of the linear algebra aspects of matrix theory and a complete proof of Jordan's canonical form we recommend Halmos' book [69]. ♠

Chapter II

Sparse Matrices

The eigenvalue problems that arise in practice often involve very large matrices. The meaning of ‘large’ is relative and it is changing rapidly with the progress of computer technology. A matrix of size a few hundreds can be considered large if one is working on a workstation, while, similarly, a matrix whose size is in the millions can be considered large if one is using a supercomputer. Fortunately, many of these matrices are also sparse, i.e., they have very few nonzeros. Again, it is not clear how ‘few’ nonzeros a matrix must have before it can be called sparse. A commonly used definition due to Wilkinson is to say that a matrix is sparse *whenever it is possible to take advantage of the number and location of its nonzero entries.* By this definition a tridiagonal matrix is sparse, but so would also be a triangular matrix, which may not be as convincing. It is probably best to leave this notion somewhat vague, since the decision as to whether or not a matrix should be considered sparse is a practical one that is ultimately made by the user.

1. Introduction

The natural idea of taking advantage of the zeros of a matrix and their location has been exploited for a long time. In the simplest situation, such as for banded or tridiagonal matrices, special techniques are straightforward to develop. However, the notion of exploiting sparsity for general sparse matrices, i.e., sparse matrices with irregular structure, has become popular only after the 1960's. The main issue, and the first one to be addressed by sparse matrix technology, is to devise direct solution methods for linear systems, that are economical both in terms of storage and computational effort. These sparse direct solvers allow to handle very large problems that could not be tackled by the usual 'dense' solvers. We will briefly discuss the solution of large sparse linear systems in Section 4 of this Chapter.

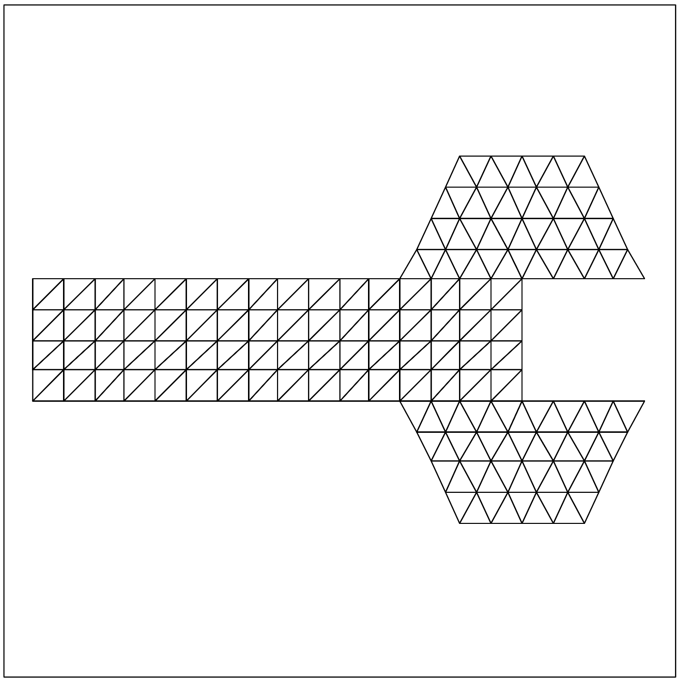


Figure 2.1 A finite element grid model

There are basically two broad types of sparse matrices: *structured* and *unstructured*. A structured sparse matrix is one whose nonzero entries, or square blocks of nonzero entries, form a regular pattern, often along a small number of diagonals. A matrix with irregularly located entries is said to be irregularly structured. The best example of a regularly structured matrix is that of a matrix that consists only of a few diagonals. Figure 2.2 shows a small irregularly structured sparse matrix associated with the finite element grid problem shown in Figure 2.1.

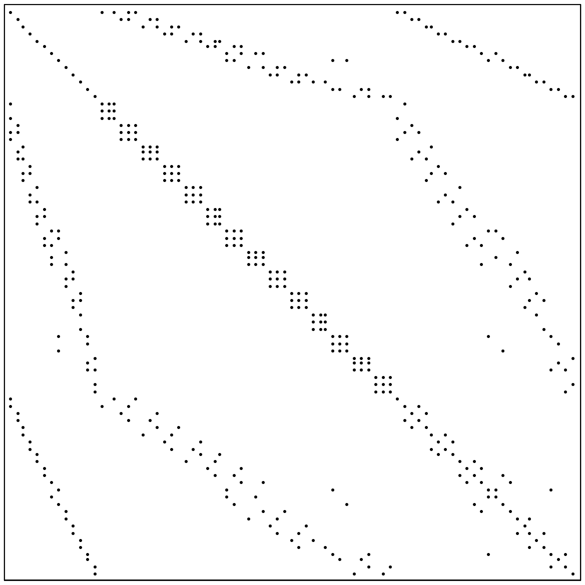


Figure 2.2 Sparse matrix associated with the finite element grid of Figure 2.1

Although the difference between the two types of matrices may not matter that much for direct solvers, it may be important for eigenvalue methods or iterative methods for solving linear systems. In these methods, one of the essential operations are matrix by vector products. The performance of these operations on supercomputers can differ significantly from one data structure to another. For example, diagonal storage schemes are ideal for vector machines, whereas more general schemes, may suffer on such machines because of the need to use indirect addressing.

In the next section we will discuss some of the storage schemes used for sparse matrices. Then we will see how some of the simplest matrix operations with sparse matrices can be performed. We will then give an overview of sparse linear system solution methods. The last two sections discuss test matrices and a set of tools for working with sparse matrices called SPARSKIT.

2. Storage Schemes

In order to take advantage of the large number of zero elements special schemes are required to store sparse matrices. Clearly, the main goal is to represent only the nonzero elements, and be able at the same time to perform the commonly needed matrix operations. In the following we will denote by Nz the total number of nonzero elements. We describe only the most popular schemes but additional details can be found in the book by Duff, Erisman, and Reid [38].

The simplest storage scheme for sparse matrices is the so-called coordinate format. The data structure consists of three arrays: a real array containing all the real (or complex) values of the nonzero elements of A in any order, an integer array containing their row indices and a second integer array containing their column indices. All three arrays are of length Nz. Thus the matrix

$$A = \begin{pmatrix} 1. & 0. & 0. & 2. & 0. \\ 3. & 4. & 0. & 5. & 0. \\ 6. & 0. & 7. & 8. & 9. \\ 0. & 0. & 10. & 11. & 0. \\ 0. & 0. & 0. & 0. & 12. \end{pmatrix} \tag{2.1}$$

will be represented (for example) by

AA	=	12.	9.	7.	5.	1.	2.	11.	3.	6.	4.	8.	10.
JR	=	5	3	3	2	1	1	4	2	3	2	3	4
JC	=	5	5	3	4	1	4	4	1	1	2	4	3

In the above example we have, on purpose, listed the elements in an arbitrary order. In fact it would have been more natural to list the elements by row or columns. If we listed the elements row-wise, we would notice that the array JC contains redundant information, and may be replaced by an array that points to the beginning of each row instead. This would entail nonnegligible savings in storage. The new data structure consists of three arrays with the following functions.

- A real array AA contains the real values a_{ij} stored row by row, from row 1 to n. The length of AA is Nz.
- An integer array JA contains the column indices of the elements a_{ij} as stored in the array AA. The length of JA is Nz.
- An integer array IA contains the pointers to the beginning of each row in the arrays AA and JA. Thus the content of $IA(i)$ is the position in arrays AA and JA where the i-th row starts. The length of IA is $n+1$ with $IA(n+1)$ containing the number $IA(1)+Nz$, i.e., the address in A and JA of the beginning of a fictitious row $n+1$.

Thus, the above matrix could be stored as follows.

AA	=	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
JA	=	1	4	1	2	4	1	3	4	5	3	4	5
IA	=	1	3	6	10	12	13						

This format is probably the most commonly used to store general sparse matrices. We will refer to it as the *Compressed*

Sparse Row (CSR) format. An advantage of this scheme over the coordinate scheme is that it is often more amenable to perform typical computations. On the other hand the coordinate scheme is attractive because of its simplicity and its flexibility. For this reason it is used as the 'entry' format in software packages such as the Harwell library.

There are a number of variations to the Compressed Sparse Row format. The most obvious variation is to store the columns instead of the rows. The corresponding scheme will be called the *Compressed Sparse Column* (CSC) scheme Another common variation exploits the fact that the diagonal elements of many matrices are usually all nonzero and/or that they are accessed more often than the rest of the elements. As a result they can be stored separately. In fact, what we refer to as the *Modified Sparse Row* (MSR) format, consists of only two arrays: a real array AA and an integer array JA. The first n positions in AA contain the diagonal elements of the matrix, in order. The position $n{+}1$ of the array AA is not used, or may sometimes be used to carry some other information concerning the matrix. Starting at position $n{+}2$, the nonzero elements of AA, excluding its diagonal elements, are stored row-wise. Corresponding to each element $AA(k)$ the integer $JA(k)$ is the column index of the element $A(k)$ in the matrix AA. The $n+1$ first positions of JA contain the pointer to the beginning of each row in AA and JA. Thus, for the above example the two arrays will be as follows.

AA =	1.	4.	7.	11.	12.	*	2.	3.	5.	6.	8.	9.	10.
JA =	7	8	10	13	14	14	4	1	4	1	4	5	3

The star denotes an unused location. Notice that $JA(n) = JA(n+1) = 14$, indicating that the last row, is a zero row, once the diagonal element has been removed.

There are a number of applications that lead to regularly structured matrices. Among these matrices one can distinguish two different types: block matrices, and diagonally structured matrices. Here we discuss only diagonally structured matrices which are matrices whose nonzero elements are located along a small number of diagonals. To store such matrices we may store the diagonals in a rectangular array $DIAG(1:n,1:Nd)$ where Nd is the number of diagonals. We also need to know the offsets of each of the diagonals with respect to the main diagonal. These will be stored in an array $IOFF(1:Nd)$. Thus, in position (i,j) of the array $DIAG$ is located the element $a_{i,i+\mathrm{IOFF}(j)}$ of the original matrix, i.e.,

$$DIAG(i,j) \leftarrow a_{i,i+\mathrm{ioff}(j)}.$$

The order in which the diagonals are stored in the columns of $DIAG$ is unimportant in general. If many more operations are performed with the main diagonal there may be a slight advantage in storing it in the first column. Note also that all the diagonals except the main diagonal have fewer than n elements, so there are positions in $DIAG$ that will not be used.

For example the following matrix which has three diagonals

$$A = \begin{pmatrix} 1. & 0. & 2. & 0. & 0. \\ 3. & 4. & 0. & 5. & 0. \\ 0. & 6. & 7. & 0. & 8. \\ 0. & 0. & 9. & 10. & 0. \\ 0. & 0. & 0. & 11. & 12. \end{pmatrix} \qquad (2.2)$$

will be represented the two arrays

$$\mathrm{DIAG} = \begin{array}{|ccc|} \hline * & 1. & 2. \\ 3. & 4. & 5. \\ 6. & 7. & 8. \\ 9. & 10. & * \\ 11 & 12. & * \\ \hline \end{array} \qquad \mathrm{IOFF} = \boxed{\text{-1} \;\; 0 \;\; 2}$$

A more general scheme that has been popular on vector machines is the so-called Ellpack-Itpack format. The assumption in this scheme is that we have at most Nd nonzero elements per row, where Nd is small. Then two rectangular arrays of dimension $n \times Nd$ each are required, one real and one integer. The first, $COEF$, is similar to $DIAG$ and contains the nonzero elements of A. We can store the nonzero elements of each row of the matrix in a row of the array $COEF(1:n, 1:Nd)$ completing the row by zeros if necessary. Together with $COEF$ we need to store an integer array $JCOEF(1:n, 1:Nd)$ which contains the column positions of each entry in $COEF$. Thus, for the above matrix, we would have,

$$\text{COEF} = \begin{array}{|ccc|} \hline 1. & 2. & 0. \\ 3. & 4. & 5. \\ 6. & 7. & 8. \\ 9. & 10. & 0. \\ 11 & 12. & 0. \\ \hline \end{array} \qquad \text{JCOEF} = \begin{array}{|ccc|} \hline 1 & 3 & 1 \\ 1 & 2 & 4 \\ 2 & 3 & 5 \\ 3 & 4 & 4 \\ 4 & 5 & 5 \\ \hline \end{array} \ .$$

Note that in the above $JCOEF$ array we have put a column number equal to the row number, for the zero elements that have been added to pad the rows of $DIAG$ that correspond to shorter rows in the matrix A. This is somewhat arbitrary, and in fact any integer between 1 and n would be acceptable, except that there may be good reasons for not putting the same integers too often, for performance considerations.

3. Basic Sparse Matrix Operations

One of the most important operations required in many of the algorithms for computing eigenvalues of sparse matrices is the matrix-by-vector product. We do not intend to show how these are performed for each of the storage schemes considered earlier, but only for a few important ones.

The following Fortran 8-X segment shows the main loop of the matrix by vector operation for matrices stored in the Compressed Sparse Row stored format.

```
DO I=1, N
   K1 = IA(I)
   K2 = IA(I+1)-1
   Y(I) = DOTPRODUCT(A(K1:K2),X(JA(K1:K2)))
ENDDO
```

Notice that each iteration of the loop computes a different component of the resulting vector. This has the obvious advantage that each of these iterations can be performed independently. If the matrix is stored column-wise, then we would use the following code instead.

```
DO J=1, N
   K1 = IA(J)
   K2 = IA(J+1)-1
   Y(JA(K1:K2)) = Y(JA(K1:K2))+X(J)*A(K1:K2)
ENDDO
```

In each iteration of the loop a multiple of the j-th column is added to the result, which is assumed to have been set initially to zero. Notice now that the outer loop is no longer parallelizable. Barring the use of a different data structure, the only alternative left to improve parallelization is to attempt to split the vector operation in each inner loop, which has few operations, in general. The point of this comparison is that we may have to change data structures to improve performance when dealing with supercomputers.

We now consider the matrix-vector product in diagonal storage.

```
DO J=1, NDIAG
   JOFF = IOFF(J)
   DO I=1, N
      Y(I) = Y(I) + DIAG(I,J)*X(JOFF+I)
   ENDDO
 ENDDO
```

Here, each of the diagonals is multiplied by the vector x and the result added to the vector y. It is again assumed that the vector y has been filled with zero elements before the start of the loop. From the point of view of parallelization and/or vectorization the above code is probably the one that has the most to offer. On the other hand, its drawback is that it is not general enough.

Another important 'kernel' in sparse matrix computations is that of solving a lower or upper triangular system. The following segment shows a simple routine for solving a unit lower triangular system.

```
X(1) = Y(1)
DO K = 2, N
   K1 = IAL(K)
   k2 = IAL(K+1)-1
   X(K)=Y(K)-DOTPRODUCT(AL(K1:K2),X(JAL(K1:K2)))
ENDDO
```

4. Sparse Direct Solution Methods

Solution methods for large sparse linear systems of equations are important in eigenvalue calculations mainly because they are needed in the context of the shift-and-invert techniques, described in Chapter IV. In these techniques the matrix that is used in the iteration process is $(A - \sigma I)^{-1}$ or $(A - \sigma B)^{-1}B$ for the generalized eigenvalue problem. In this section we give a brief overview of sparse matrix techniques for solving linear systems. The difficulty here is that we must deal with problems that are not only complex, since complex shifts are likely to occur, but also indefinite. There are two broad classes of methods that are commonly used: direct and iterative. Direct methods are more commonly used in the context of shift-and-invert techniques because of their robustness when dealing with indefinite problems.

Most direct methods for sparse linear systems perform an LU factorization of the original matrix and try to reduce cost by minimizing fill-ins, i.e., non-zero elements introduced during the elimination process in positions which were initially zeros. Typical codes in this category include MA28, see reference [36], from the Harwell library and the Yale Sparse Matrix Package (YSMP), see reference [163]. For a detailed view of sparse matrix techniques we refer to the book by Duff, Erisman, and Reid [38].

Currently, the most popular iterative methods are the preconditioned conjugate gradient type techniques. In these techniques an approximate factorization $A = LU + E$ of the original matrix is obtained and then the conjugate gradient method is applied to a preconditioned system, a form of which is $U^{-1}L^{-1}Ax = U^{-1}L^{-1}b$. The conjugate gradient method is a projection method related to the Lanczos algorithm, which will be described in Chapter VI. One difficulty with conjugate gradient-type methods is that they are designed for matrices that are positive real, i.e., matrices whose symmetric parts are positive definite, and as a result they will perform well for the types of problems that will arise in the context of shift-and-invert.

5. Test Problems

When developing algorithms for sparse matrix computations it is desirable to be able to use test matrices that are well documented and often used by other researchers. There are many different ways in which these test matrices can be useful but their most common use is for comparison purposes.

Two different ways of providing data sets consisting of large sparse matrices for test purposes have been used in the past. The first one is to collect sparse matrices in a well-specificed format, from various applications. This approach has is used in the well-known Harwell-Boeing collection of test matrices. The second approach is to collect subroutines or programs that generate such matrices. This approach is taken in the SPARSKIT package which we briefly describe in the next section.

In the course of the book we will often use two test problems in the examples. These are described in detail next. While these two examples are far from being representative of all the problems that occur they have the advantage of being easy to reproduce. They have also been extensively used in the literature.

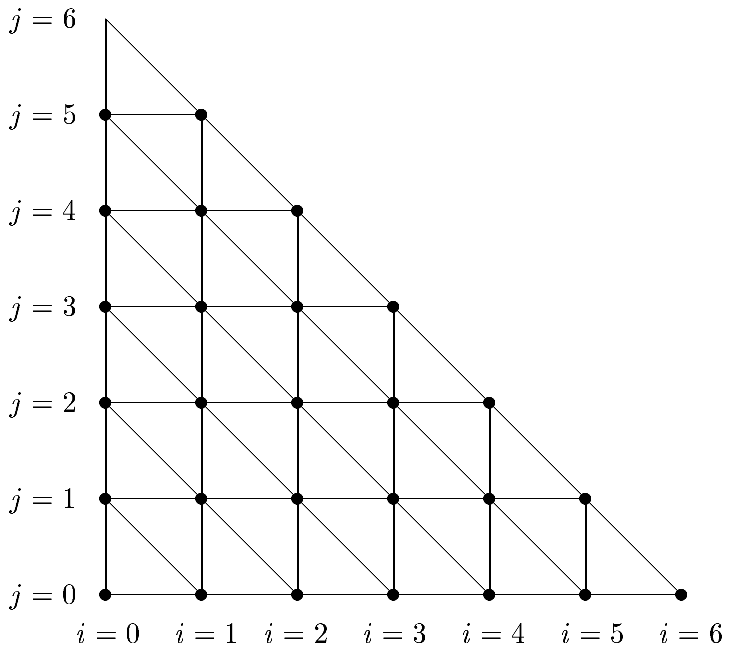


Figure 2.3 Random walk on a triangular grid

5.1. Random Walk Problem

The first test problem is issued from a Markov model of a random walk on a triangular grid. It was proposed by G. W. Stewart [170] and has been used in several papers for testing eigenvalue algorithms. The problem models a random walk on a $(k+1) \times (k+1)$ triangular grid as is shown in Figure 2.3.

We label by (i, j) the node of the grid with coordinates (ih, jh) where h is the grid spacing, for $i, j = 0, 1, ...k$. A particle moves randomly on the grid by jumping from a node (i, j) into either of its (at most 4) neighbors. The probability of jumping from node (i, j) to either node $(i - 1, j)$ or node $(i, j - 1)$ (down transition) is given by

$$\mathrm{pd}(i, j) = \frac{i + j}{2k}$$

this probability being doubled when either i or j is equal to zero. The probability of jumping from node (i, j) to either node $(i+1, j)$ or node $(i, j + 1)$ (up transition) is given by

$$\mathrm{pu}(i, j) = \frac{1}{2} - \mathrm{pd}(i, j).$$

Note that there cannot be an up transition when $i + j = k$, i.e., for nodes on the oblique boundary of the grid. This is reflected by the fact that in this situation $\mathrm{pu}(i, j) = 0$.

The problem is to compute the steady state probability distribution of the chain, i.e., the probabilities that the particle be located in each grid cell after a very long period of time. We number the nodes from the bottom up and from left to right, i.e., in the order,

$$(0, 0), (0, 1), \ldots, (0, k); (1, 0), (1, 1), \ldots (1, k - 1); .....; (k, 0)$$

The matrix P of transition probabilities is the matrix whose generic element $p_{k,q}$ is the probability that the particle jumps from node k to node q. This is a stochastic matrix, i.e., its elements are nonnegative and the sum of elements in the same row is equal to one. The vector $(1, 1, ..., 1)^T$ is an eigenvector of P associated with the eigenvalue unity. As is known the steady state probability distribution vector is the appropriately scaled eigenvector of the transpose of P associated with the eigenvalue one. Note that the number of different states is $\frac{1}{2}(k + 1)(k + 2)$, which is the dimension of the matrix. We will denote by Mark(k+1) the corresponding matrix. Figure 2.4 shows the sparsity pattern of Mark(15) which is a matrix of dimension $n = 120$ with $nz = 420$ nonzero elements.

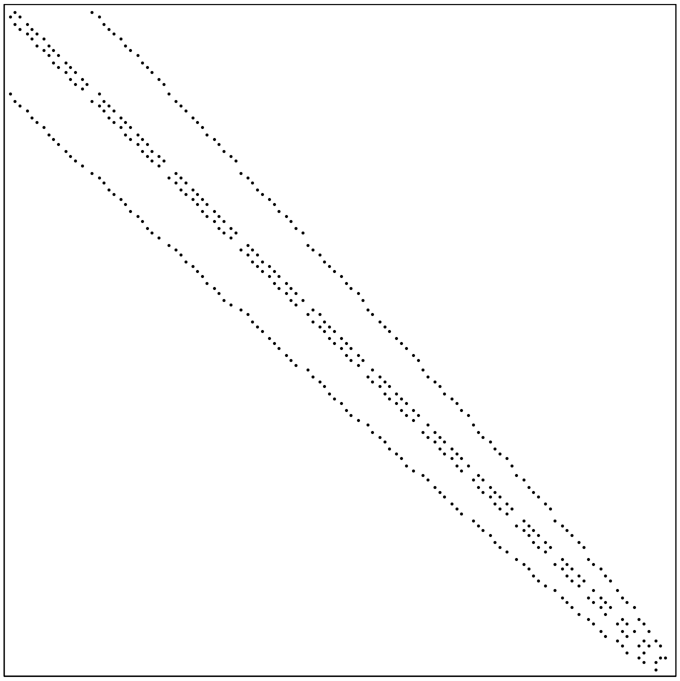


Figure 2.4 Sparsity pattern of the matrix Mark(15).

5.2. Chemical Reactions

The second test example, models concentration waves in reaction and transport interaction of some chemical solutions in a tubular reactor. The concentrations $x(\tau, z), y(\tau, z)$ of two reacting and diffusing components, where $0 \leq z \leq 1$ represents a coordinate along the tube, and τ is the time, are modeled by the system:

$$\frac{\partial x}{\partial \tau} = \frac{D_x}{L^2}\frac{\partial^2 x}{\partial z^2} + f(x, y), \tag{2.3}$$

$$\frac{\partial y}{\partial \tau} = \frac{D_y}{L^2}\frac{\partial^2 y}{\partial z^2} + g(x, y), \tag{2.4}$$

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:

$$x(0,\tau) = x(1,\tau) = \bar{x}$$
$$y(0,\tau) = y(1,\tau) = \bar{y}.$$

The linear stability of the above system is traditionally studied around the steady state solution obtained by setting the partial derivatives of x and y with respect to time to be zero. More precisely, the stability of the system is the same as that of the Jacobian of (2.3) - (2.4) evaluated at the steady state solution. In many problems one is primarily interested in the existence of limit cycles, or equivalently the existence of periodic solutions to (2.3), (2.4). This translates into the problem of determining whether the Jacobian of (2.3), (2.4) evaluated at the steady state solution admits a pair of purely imaginary eigenvalues.

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

$$\begin{aligned} f(x,y) &= A - (B+1)x + x^2 y \\ g(x,y) &= Bx - x^2 y. \end{aligned}$$

Then, the above system admits the trivial stationary solution $\bar{x} = A$, $\bar{y} = B/A$. A stable periodic solution to the system exists if the eigenvalues of largest real parts of the Jacobian of the right-hand side of (2.3), (2.4) is exactly zero. To verify this numerically, we first need to discretize the equations with respect to the variable z and compute the eigenvalues with largest real parts of the resulting discrete Jacobian.

For this example, the exact eigenvalues are known and the problem is analytically solvable. The following set of parameters have been commonly used in previous articles,

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

The bifurcation parameter is L. 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 $n + 1$ points, and define the mesh size $h \equiv 1/n$. The discrete vector is of the form $\binom{x}{y}$ where x and y are n-dimensional vectors. Denoting 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

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

and

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

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 because the steady state solution is a constant with respect to 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. The matrix A has the following structure

$$A = \begin{pmatrix} \alpha T & \beta I \\ \gamma I & \delta T \end{pmatrix},$$

In which $T = \text{tridiag}\ \{1, -2, 1\}$, and α, β, γ, and δ are scalars. 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 T.

5.3. The Harwell-Boeing Collection

This large collection of test matrices has been gathered over several years by I. Duff (Harwell) and R. Grimes and J. Lewis (Boeing) [39]. The number of matrices in the collection at the time of this writing is 292. The matrices have been contributed by researchers and engineers in many different areas. The sizes of the matrices vary from very small, such as counter example matrices, to very large. One drawback of the collection is that it contains few *non-Hermitian* eigenvalue problems. Many of the eigenvalue problems in the collection are from structural engineering, which are generalized eigenvalue problems. One the other hand the collection provides a data structure which constitutes an excellent medium of exchanging matrices.

The matrices are stored as ASCII files with a very specific format consisting of a 4 or 5 line header and then the data containing the matrix stored in CSC format together with any right-hand sides, initial guesses, or exact solutions.

The collection is available for public distribution from the authors.

6. SPARSKIT

SPARSKIT is a package aimed at providing subroutines and utilities for working with general sparse matrices. Its purpose is not as much to solve particular problems involving sparse matrices (linear systems, eigenvalue problems) but rather to make available the little tools to manipulate and performs simple operations with sparse matrices. For example there are tools for exchanging data structures, e.g., passing from the Compressed Sparse Row format to the diagonal format and vice versa. There are various tools for extracting submatrices or performing other similar manipulations. SPARSKIT also provides matrix generation subroutines as well as basic linear algebra routines for sparse matrices (such as addition, multiplication, etc...).

A short description of the contents of SPARSKIT follows. The package is divided up in six modules, each having a different function. To refer to these six parts we will use the names of the subdirectories where they are held in the package in its current version.

FORMATS This module contains essentially two sets of routines. The first set contained in the file formats.f consists of the routines needed to translate data structures. Translations from the basic Compressed Sparse Row format to any of the other formats supported is provided together with a routine for the reverse transformation. This way one can translate from any of the data structures supported to any other one with two transformation at most. The formats currently supported are the following.

DNS Dense format

BND Linpack Banded format

CSR Compressed Sparse Row format

CSC Compressed Sparse Column format

COO Coordinate format

ELL Ellpack-Itpack generalized diagonal format

DIA Diagonal format

BSR Block Sparse Row format

MSR Modified Compressed Sparse Row format

SSK Symmetric Skyline format

NSK Nonsymmetric Skyline format

JAD The Jagged Diagonal scheme

The second set of routines contains a number of routines, currently 27, called 'unary', to perform simple manipulation functions on sparse matrices, such as extracting a particular diagonal or permuting a matrix, or yet for filtering out small elements. For reasons of space we cannot list these routines here.

BLASSM This module contains a number of routines for doing basic linear algebra with sparse matrices. It is comprised of essentially two sets of routines. Basically, the first one consists of matrix-matrix operations (e.g., multiplication of matrices) and the second consists of matrix-vector operations. The first set allows to perform the following operations with sparse matrices, where A, B, C are sparse matrices, D is a diagonal matrix, and σ is a scalar. $C = AB$, $C = A + B$, $C = A + \sigma B$, $C = A \pm B^T$, $C = A + \sigma B^T$, $A := A + \sigma I$, $C = A + D$.

The second set contains various routines for performing matrix by vector products and solving sparse triangular linear systems in different storage formats.

INOUT This module consists of routines to read and write matrices in the Harwell-Boeing format. For more information on this format and the Harwell-Boeing collection see the reference [39]. It also provides routines for plotting the pattern of the matrix or simply dumping it in a nice format.

INFO There is currently only one subroutine in this module. Its purpose is to provide as many statistics as possible on a matrix with little cost. About 33 lines of information are written. For example, the code analyzes diagonal dominance of the matrix (row and column), its degree of symmetry (structural as well as numerical), its block structure, its diagonal structure, etc,...

MATGEN The set of routines in this module allows one to generate test matrices. For now there are generators for 5 different types of matrices.

1. Five-point and seven point matrices on rectangular regions discretizing a general elliptic partial differential equation.

2. Same as above but provides block matrices (several degrees of freedom per grid point in the PDE).

3. Finite elements matrices for the heat condition problem, using various domains (including user provided ones).

4. Test matrices from the paper by Z. Zlatev, K. Schaumburg, and J. Wasniewski, [187].

5. Markov chain matrices arising from a random walk on a triangular grid. See Section 5.1 for details.

UNSUPP As is suggested by its name this module contains various *unsupported* software tools that are not necessarily portable or that do not fit in any of the previous modules. For example software for viewing matrix patterns on some workstations will be found here. For now UNSUPP contains subroutines for visualizing matrices and a preconditioned GMRES package (with a 'robust' preconditioner based on Incomplete LU factorization with controlled fill-in).

Problems

P-2.1 Write a FORTRAN code segment to perform the matrix-vector product for matrices stored in Ellpack-Itpack format.

P-2.2 Write a small subroutine to perform the following operations on a sparse matrix in coordinate format, diagonal format, and in CSR format: a) count the number on nonzero elements in the main diagonal; b) extract the diagonal whose offset is k (which may be negative); c) add a nonzero element in position (i,j) of the matrix (assume that this position may contain a zero or a nonzero element); d) add a given diagonal to the matrix. What is the most convenient storage scheme for each of these operations?

P-2.3 Generate explicitly the matrix Mark(4). Verify that it is a stochastic matrix. Verify that 1 and -1 are eigenvalues.

Notes and References. Two good sources of reading on sparse matrix computations are the books by George and Liu [56] and by Duff, Erisman, and Reid [38]. Also of interest are [111] and [129] and the early survey by Duff [35]. For applications related to eigenvalue problems, see [27] and [8]. For details on Markov Chain modeling see [86, 162]. The SPARSKIT package is part of an ongoing project. Write to the author for information. Some documentation is available in the technical report [151]. Another manipulation package for sparse matrices, similar to SPARSKIT in spirit, is SMMS developed by Alvarado [1]. ♠

Chapter III

Perturbation Theory and Error Analysis

This chapter introduces some elementary spectral theory for linear operators on finite dimensional spaces as well as some elements of perturbation analysis. The main question that perturbation theory addresses is: how does an eigenvalue and its associated eigenvectors, spectral projector, etc.., vary when the original matrix undergoes a small perturbation. This information is important both for theoretical and practical purposes. The spectral theory introduced in this chapter is the main tool used to extend what is known about spectra of matrices to general operators on infinite dimensional spaces. However, it has also some consequences in analyzing the behavior of eigenvalues and eigenvectors of matrices. The material discussed in this chapter is probably the most theoretical of the book. Fortunately, most of it is independent of the rest and may be skipped in a first reading. The notions of condition numbers and some of the results concerning error bounds are crucial in understanding the difficulties that eigenvalue routines may encounter.

1. Projectors and their Properties

A projector P is a linear transformation from $\mathbb{C}^n$ to itself which is idempotent, i.e., such that

$$P^2 = P.$$

When P is a projector then so is $(I - P)$ and we have $\mathrm{Ker}(P) = \mathrm{Ran}(I - P)$. The two subspaces $\mathrm{Ker}(P)$ and $\mathrm{Ran}(P)$ have only the element zero in common. This is because if a vector x is in $\mathrm{Ran}(P)$ then $Px = x$ and if it is also in $\mathrm{Ker}(P)$ then $Px = 0$ so that $x = 0$ and the intersection of the two subspaces reduces to $\{0\}$. Moreover, every element of $\mathbb{C}^n$ can be written as $x = Px + (I - P)x$. As a result the space $\mathbb{C}^n$ can be decomposed as the direct sum

$$\mathbb{C}^n = \mathrm{Ker}(P) \ \oplus \ \mathrm{Ran}(P).$$

Conversely, every pair of subspaces M and S that form a direct sum of $\mathbb{C}^n$ define a unique projector such that $\mathrm{Ran}(P) = M$ and $\mathrm{Ker}(P) = S$. The corresponding transformation P is the linear mapping that maps any element x of $\mathbb{C}^n$ into the component x_1 where x_1 is the M-component in the unique decomposition $x = x_1 + x_2$ associated with the direct sum. In fact, this association is unique in that a projector is uniquely determined by its kernel and range, two subspaces that form a direct sum of $\mathbb{C}^n$.

1.1. Orthogonal Projectors

An important particular case is when the subspace S is the orthogonal complement of M, i.e., when

$$\mathrm{Ker}(P) = \mathrm{Ran}(P)^{\perp}.$$

In this case the projector P is said to be the *orthogonal projector* onto M. Since $\mathrm{Ran}(P)$ and $\mathrm{Ker}(P)$ from a direct sum of $\mathbb{C}^n$, the decomposition $x = Px + (I - P)x$ is unique and the vector Px is uniquely defined by the set of equations

$$Px \in M \quad \text{and} \quad (I - P)x \perp M \tag{3.1}$$

or equivalently,

$$Px \in M \quad \text{and} \quad ((I - P)x, y) = 0 \quad \forall y \in M \ .$$

Proposition 3.1 *A projector is orthogonal if and only if it is Hermitian.*

Proof. As a consequence of the equality

$$(P^H x, y) = (x, Py) \quad \forall x \ , \ \forall y \tag{3.2}$$

we conclude that

$$\text{Ker}(P^H) = \text{Ran}(P)^\perp \tag{3.3}$$
$$\text{Ker}(P) = \text{Ran}(P^H)^\perp \ . \tag{3.4}$$

By definition an orthogonal projector is one for which $\text{Ker}(P) = \text{Ran}(P)^\perp$. Therefore, by (3.3), if P is Hermitian then it is orthogonal.

To show that the converse is true we first note that P^H is also a projector since $(P^H)^2 = (P^2)^H = P^H$. We then observe that if P is orthogonal then (3.3) implies that $\text{Ker}(P) = \text{Ker}(P^H)$ while (3.4) implies that $\text{Ran}(P) = \text{Ran}(P^H)$. Since P^H is projector this implies that $P = P^H$, because a projector is uniquely determined by its range and its kernel. ∎

Given *any* unitary $n \times m$ matrix V whose columns form an orthonormal basis of $M = \text{Ran}(P)$, we can represent P by the matrix $P = VV^H$. Indeed, in addition to being idempotent, the linear mapping associated with this matrix satisfies the characterization given above, i.e.,

$$VV^H x \ \in M \quad \text{and} \quad (I - VV^H)x \ \in \ M^\perp.$$

It is important to note that this representation of the orthogonal projector P is not unique. In fact any orthonormal basis V will give a different representation of P in the above form. As a consequence for any two orthogonal bases V_1, V_2 of M, we must have $V_1 V_1^H = V_2 V_2^H$, an equality which can also be verified independently, see Exercise P-3.2.

From the above representation it is clear that when P is an orthogonal projector then we have $\|Px\|_2 \le \|x\|_2$ for any x. As a result the maximum of $\|Px\|_2 / \|x\|_2$ for all x in $\mathbb{C}^n$ does not exceed one. On the other hand the value one is reached for any element in $\mathrm{Ran}(P)$ and therefore,

$$\|P\|_2 = 1$$

for any orthogonal projector P.

Recall that the acute angle between two nonzero vectors of $\mathbb{C}^n$ is defined by

$$\cos\theta(x,y) = \frac{|(x,y)|}{\|x\|_2 \|y\|_2} \quad 0 \le \theta(x,y) \le \frac{\pi}{2} \ .$$

We define the acute angle between a vector and a subspace S as the smallest acute angle made between x and all vectors y of S,

$$\theta(x,S) = \min_{y \in S} \ \theta(x,y) \ . \tag{3.5}$$

An optimality property of orthogonal projectors is the following.

Theorem 3.1 *Let P be an orthogonal projector onto the subspace S. Then given any vector x in $\mathbb{C}^n$ we have,*

$$\min_{y \in S} \|x - y\|_2 = \|x - Px\|_2 \ , \tag{3.6}$$

or, equivalently,

$$\theta(x,S) = \theta(x,Px) \ . \tag{3.7}$$

Proof. Let y any vector of S and consider the square of its distance from x. We have,

$$\|x-y\|_2^2 = \|x - Px + (Px - y)\|_2^2 = \|x - Px\|_2^2 + \|(Px - y)\|_2^2 \ ,$$

because $x - Px$ is orthogonal to S to which $Px - y$ belongs. Therefore, $\|x-y\|_2 \geq \|x - Px\|_2$ for all y in S and this establishes the first result by noticing that the minimum is reached for $y = Px$. The second equality is a simple reformulation of the first. ■

It is sometimes important to be able to measure distances between two subspaces. If P_i represents the orthogonal projector onto M_i, for $i = 1, 2$, a natural measure of the distance between M_1 and M_2 is provided by their *gap* defined by:

$$\omega(M_1, M_2) = \max \left\{ \max_{\substack{x \in M_2 \\ \|x\|_2 = 1}} \|x - P_1 x\|_2 \ , \ \max_{\substack{x \in M_1 \\ \|x\|_2 = 1}} \|x - P_2 x\|_2 \right\}$$

We can also redefine $\omega(M_1, M_2)$ as

$$\omega(M_1, M_2) = \max\{\|(I - P_1)P_2\|_2 \ , \ \|(I - P_2)P_1\|_2\}$$

and it can even be shown that

$$\omega(M_1, M_2) = \|P_1 - P_2\|_2. \qquad (3.8)$$

1.2. Oblique Projectors

A projector that is not orthogonal is said to be oblique. It is sometimes useful to have a definition of oblique projectors that resembles that of orthogonal projectors, i.e., a definition similar to (3.1). If we call L the subspace that is the orthogonal complement to $S = \text{Ker}(P)$, it is clear that L will have the same dimension as M. Moreover, to say that $(I-P)x$ belongs to $\text{Ker}(P)$ is equivalent to saying that it is in the orthogonal complement of L. Therefore, from the definitions seen at the beginning of Section 1, the projector P can be characterized by the defining equation

$$Px \in M \quad \text{and} \quad (I - P)x \perp L \ . \qquad (3.9)$$

We say that P is a projector onto M and orthogonal to L or along the orthogonal complement of L. This is illustrated in Figure 3.1.

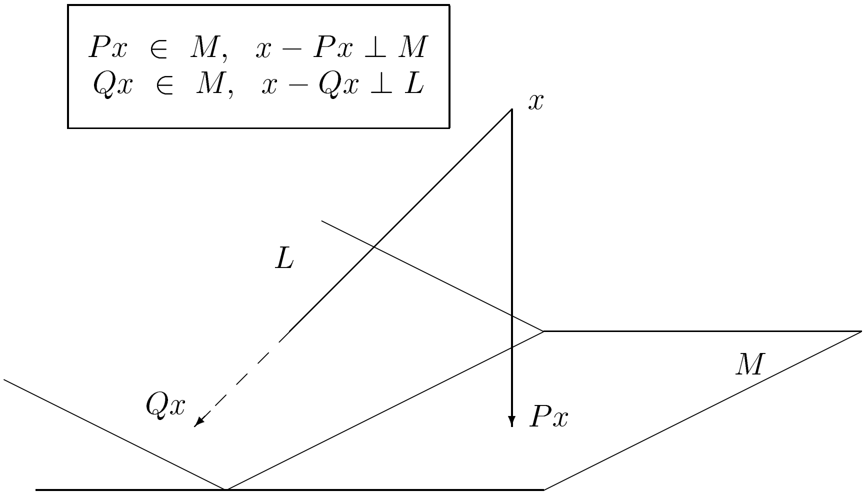

$Px \in M, \quad x - Px \perp M$
$Qx \in M, \quad x - Qx \perp L$
x
L
M
Qx
Px

Figure 3.1 Orthogonal and oblique projectors P and Q.

Matrix representations of oblique projectors require two bases: a basis $V = [v_1, \ldots, v_m]$ of the subspace $M = \mathrm{Ran}(P)$ and the other $W = [w_1, \ldots, w_m]$ for the subspace L, the orthogonal complement of $\mathrm{Ker}(P)$. We will say that these two bases are *biorthogonal* if

$$(v_i, w_j) = \delta_{ij} \tag{3.10}$$

Given any pair of biorthogonal bases V, W the projector P can be represented by

$$P = VW^H \tag{3.11}$$

In contrast with orthogonal projectors, the norm of P is larger than one in general. It can in fact be arbitrarily large, which implies that the norms of $P - Q$, for two oblique projectors P and Q, will not, in general, be a good measure of the distance between the two subspaces $\mathrm{Ran}(P)$ and $\mathrm{Ran}(Q)$. On the other hand, it may give an idea on the difference between their rank as is stated in the next theorem.

Theorem 3.2 *Let $\|.\|$ be any matrix norm, and assume that two projectors P and Q are such that $\|P-Q\|<1$ then*

$$\text{rank}(P) \;=\; \text{rank}(Q) \tag{3.12}$$

Proof. First let us show that $\text{rank}(Q) \leq \text{rank}(P)$. Given a basis $\{x_i\}_{i=1,\ldots,q}$ of $\text{Ran}(Q)$ we consider the family of vectors $G = \{Px_i\}_{i=1,\ldots,q}$ in $\text{Ran}(P)$ and show that it is linearly independent. Assume that

$$\sum_{i=1}^{q} \alpha_i P x_i = 0.$$

Then the vector $y = \sum_{i=1}^{q} \alpha_i x_i$ is such that $Py = 0$ and therefore $(Q-P)y = Qy = y$ and $\|(Q-P)y\| = \|y\|$. Since $\|Q-P\| < 1$ this implies that $y = 0$. As a result the family G is linearly independent and so $\text{rank}(P) \geq q = \text{rank}(Q)$. It can be shown similarly that $\text{rank}(P) \leq \text{rank}(Q)$. ∎

The above theorem indicates that no norm of $P-Q$ can be less than one if the two subspaces have different dimensions. Moreover, if we have a family of projectors $P(t)$ that depends continuously on t then the rank of $P(t)$ remains constant. In addition, an immediate corollary is that if the gap between two subspaces is less than one then they must have the same dimension.

1.3. Resolvent and Spectral Projector

For any given complex z not in the spectrum of a matrix A we define the resolvent operator of A at z as the linear transformation

$$R(A,z) = (A - zI)^{-1} . \tag{3.13}$$

The notation $R(z)$ is often used instead of $R(A,z)$ if there is no ambiguity. This notion can be defined for operators on infinite dimensional spaces in which case the spectrum is defined as the set of all complex scalars such that the inverse of $(A - zI)$ does not exist, see reference [14, 85] for details.

The resolvent regarded as a function of z admits singularities at the eigenvalues of A. Away from any eigenvalue the resolvent $R(z)$ is analytic with respect to z. Indeed, we can write for any z around an element z_0 not equal to an eigenvalue,

$$\begin{aligned} R(z) \equiv (A - zI)^{-1} &= ((A - z_0 I) - (z - z_0)I)^{-1} \\ &= R(z_0)(I - (z - z_0)R(z_0))^{-1} \end{aligned}$$

The term $(I - (z - z_0)R(z_0))^{-1}$ can be expanded into the Neuman series whenever the spectral radius of $(z - z_0)R(z_0)$ is less than unity. Therefore, the Taylor expansion of $R(z)$ in the open disk $|z - z_0| < 1/\rho(R(z_o))$ exists and takes the form,

$$R(z) = \sum_{k=0}^{\infty} (z - z_0)^k R(z_0)^{k+1}.$$

It is important to determine the nature of the singularity of $R(z)$ at the eigenvalues $\lambda_i, i = 1, \ldots, p$. By a simple application of Cramer's rule it is easy to see that these singularities are not essential. In other words, the Laurent expansion of $R(z)$

$$R(z) = \sum_{k=-\infty}^{+\infty} (z - \lambda_i)^k C_k$$

around each pole λ_i has only a finite number of negative powers. Thus, $R(z)$ is a meromorphic function.

The resolvent satisfies the following immediate properties.

First resolvent equality:

$$R(z_1) - R(z_2) = (z_1 - z_2)R(z_1)R(z_2) \tag{3.14}$$

Second resolvent equality:

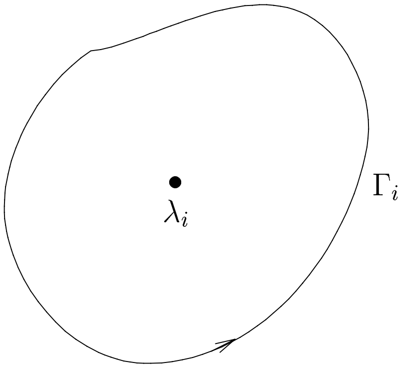
$$R(A_1, z) - R(A_2, z) = R(A_1, z)(A_2 - A_1)R(A_2, z) \tag{3.15}$$

In what follows we will need to integrate the resolvent over Jordan curves in the complex plane. A Jordan curve is a simple closed curve that is piecewise smooth and the integration will always be counter clockwise unless otherwise stated. There is not much difference between integrating complex valued functions with values in $\mathbb{C}$ or in $\mathbb{C}^{n\times n}$. In fact such integrals can be defined over functions taking their values in Banach spaces in the same way.

Consider *any* Jordan curve Γ_i that encloses the eigenvalue λ_i and no other eigenvalue of A, and let

$$P_i = \frac{-1}{2\pi i}\int_{\Gamma_i} R(z)dz \tag{3.16}$$

The above integral is often referred to as the Taylor-Dunford integral.



1.4 Relations with the Jordan form

The purpose of this subsection is to show that the operator P_i defined by (3.16) is identical with the spectral projector defined in Chapter I by using the Jordan canonical form.

Theorem 3.3 *The linear transformations P_i, $i = 1, 2, \ldots, p$, associated with the distinct eigenvalues $\lambda_i, i = 1, \ldots, p$, are such that*

(1) $P_i^2 = P_i$*, i.e., each* P_i *is a projector.*

(2) $P_iP_j = P_jP_i = 0$ *if* $i \neq j$ *.*

(3) $\sum_{i=1}^{p} P_i = I$*.*

Proof. (1) Let Γ and Γ' two curves enclosing λ_i with Γ' enclosing Γ. Then

$$\begin{aligned}(2i\pi)^2 P_i^2 &= \int_\Gamma \int_{\Gamma'} R(z)R(z')dzdz' \\ &= \int_\Gamma \int_{\Gamma'} \frac{1}{z'-z}(R(z') - R(z))dz'dz\end{aligned}$$

because of the first resolvent equality. We observe that

$$\int_\Gamma \frac{dz}{z'-z} = 0 \quad \text{and} \quad \int_{\Gamma'} \frac{dz'}{z'-z} = 2i\pi,$$

so that

$$\int_\Gamma \int_{\Gamma'} \frac{R(z')}{z'-z} dz'dz \;=\; \int_{\Gamma'} R(z') \left(\int_\Gamma \frac{dz}{z'-z} \right) dz' = 0$$

and,

$$\int_\Gamma \int_{\Gamma'} \frac{R(z)}{z'-z} dz'dz \;=\; \int_\Gamma R(z) \left(\int_{\Gamma'} \frac{dz'}{z'-z} \right) dz \;=\; 2i\pi \int_\Gamma R(z)dz$$

from which we get $P_i^2 = P_i$.

(2) The proof is similar to (1) and is left as an exercise.

(3) Consider

$$P = \frac{-1}{2i\pi} \sum_{i=1}^{p} \int_{\Gamma_i} R(z)dz \; .$$

Since there are no poles of $R(z)$ outside of the p Jordan curves, we can replace the sum of the integrals by an integral over any curve that contains all of the eigenvalues of A. If we choose this curve to be a circle C of radius r and center the origin, we get

$$P = \frac{-1}{2i\pi} \int_C R(z) dz \ .$$

Making the change of variables $t = 1/z$ we find that

$$P = \frac{-1}{2i\pi} \int_{C'_-} (A - (1/t)I)^{-1} \left(-\frac{dt}{t^2} \right) \ = \ \frac{-1}{2i\pi} \int_{C'_+} (tA - I)^{-1} \frac{dt}{t}$$

where C'_- (resp. C'_+) is the circle of center the origin, radius $1/r$ run clock-wise (resp. counter-clockwise). Moreover, because r must be larger than $\rho(A)$ we have $\rho(tA) < 1$ and the inverse of $I - tA$ is expandable into its Neuman series, i.e., the series

$$(I - tA)^{-1} = \sum_{k=0}^{\infty} (tA)^k$$

converges and therefore,

$$P = \frac{1}{2i\pi} \int_{C'+} \left[\sum_{k=0}^{k=\infty} t^{k-1} A^k \right] dt = I$$

by the residue theorem. ■

The above theorem shows that the projectors P_i satisfy the same properties as those of the spectral projector defined in the previous chapter, using the Jordan canonical form. However, to show that these projectors are identical we still need to prove that they have the same range. Note that since A and $R(z)$ commute we get by integration that $AP_i = P_i A$ and this implies that the range of P_i is invariant under A. We must show that this invariant subspace is the invariant subspace M_i associated with the eigenvalue λ_i, as defined in Chapter I. The next lemma establishes the desired result.

Lemma 3.1 *Let $\hat{M}_i = \text{Ran}(P_i)$ and let $M_i = \text{Ker}(A - \lambda_i I)^{l_i}$ be the invariant subspace associated with the eigenvalue λ_i. Then we have $M_i = \hat{M}_i$ for $i = 1, 2, \ldots, p$.*

Proof. We first prove that $M_i \subset \hat{M}_i$. This follows from the fact that when $x \in \text{Ker}(A - \lambda_i I)^{l_i}$, we can expand $R(z)x$ as follows:

$$\begin{aligned} R(z)x &= (A - zI)^{-1}x \\ &= [(A - \lambda_i I) - (z - \lambda_i)I]^{-1}x \\ &= -\frac{1}{z - \lambda_i}\left[I - (z - \lambda_i)^{-1}(A - \lambda_i I)\right]^{-1} x \\ &= \frac{-1}{z - \lambda_i}\sum_{j=0}^{l_i}(z - \lambda_i)^{-j}(A - \lambda_i I)^j x \ . \end{aligned}$$

The integral of this over Γ_i is simply $-2i\pi x$ by the residue theorem, hence the result.

We now show that $\hat{M}_i \subset M_i$. From

$$(z - \lambda_i)R(z) = -I + (A - \lambda_i I)R(z) \tag{3.17}$$

it is easy to see that

$$\frac{-1}{2i\pi}\int_\Gamma (z - \lambda_i)R(z)dz = \frac{-1}{2i\pi}(A - \lambda_i I)\int_\Gamma R(z)dz = (A - \lambda_i I)P_i$$

and more generally,

$$\begin{aligned} \frac{-1}{2i\pi}\int_\Gamma (z - \lambda_i)^k R(z)dz &= \frac{-1}{2i\pi}(A - \lambda_i I)^k \int_\Gamma R(z)dz \\ &= (A - \lambda_i I)^k P_i \ . \qquad (3.18) \end{aligned}$$

Notice that the term in the left-hand side of (3.18) is the coefficient A_{-k-1} of the Laurent expansion of $R(z)$ which has no essential singularities. Therefore, there is some integer k after which all the left-hand sides of (3.18) vanish. This proves that for every $x = P_i x$ in $\hat{M}_i$, there exists some l for which $(A - \lambda_i I)^k x = 0, k \geq l$. It follows that x belongs to M_i. ■

This finally establishes that the projectors P_i are identical with those defined with the Jordan canonical form and seen in Chapter I. Each projector P_i is associated with an eigenvalue λ_i. However, it is important to note that more generally one can define a projector associated with a group of eigenvalues, which will be the sum of the individual projectors associated with the different eigenvalues. This can also be defined by an integral similar to (3.16) where Γ is a curve that encloses all the eigenvalues of the group and no other ones. Note that the rank of P thus defined is simply the sum of the algebraic multiplicities of the eigenvalue. In other words, the dimension of the range of such a P would be the sum of the algebraic multiplicities of the distinct eigenvalues enclosed by Γ.

1.5. Linear Perturbations of A

In this section we consider the family of matrices defined by

$$A(t) = A + tH$$

where t belongs to the complex plane. We are interested in the behavior of the eigenelements of $A(t)$ when t varies around the origin. Consider first the 'parameterized' resolvent,

$$R(t, z) = (A + tH - zI)^{-1}.$$

Noting that $R(t,z) = R(z)(I + tR(z)H)^{-1}$ it is clear that if the spectral radius of $tR(z)H$ is less than one then $R(t,z)$ will be analytic with respect to t. More precisely,

Proposition 3.2 *The resolvent $R(t,z)$ is analytic with respect to t in the open disk $|t| < \rho^{-1}(HR(z))$.*

We wish to show by integration over a Jordan curve Γ that a similar result holds for the spectral projector $P(t)$, i.e., that $P(t)$ is analytic for t small enough. The result would be true if the resolvent $R(t, z)$ were analytic with respect to t *for each* z *on* Γ_i. To ensure this we must require that

$$|t| < \inf_{z\in\Gamma} \rho^{-1}(R(z)H)) \ .$$

The question that arises next is whether or not the disk of all t 's defined above is empty. The answer is no as the following proof shows. We have

$$\rho(R(z)H) \leq \|R(z)H\| \leq \|R(z)\|\|H\|.$$

The function $\|R(z)\|$ is continuous with respect to z for $z \in \Gamma$ and therefore it reaches its maximum at some point z_0 of the closed curve Γ and we obtain

$$\rho(R(z)H) \leq \|R(z)H\| \leq \|R(z_0)\|\|H\| \equiv \kappa \ .$$

Hence,

$$\inf_{z\in\Gamma} \rho^{-1}(R(z)H)) \geq \kappa^{-1} \ .$$

Theorem 3.4 *Let* Γ *be a Jordan curve around one or a few eigenvalues of* A *and let*

$$\rho_a = \inf_{z\in\Gamma} [\rho(R(z)H)]^{-1} \ .$$

Then ρ_a>0 *and the spectral projector*

$$P(t) = \frac{-1}{2\pi i} \int_\Gamma R(t, z) dz$$

is analytic in the disk $|t| < \rho_a$.

We have already proved that ρ_a>0. The rest of the proof is straightforward. As an immediate corollary of Theorem 3.4, we know that the rank of $P(t)$ will stay constant as long as t stays in the disk $|t| < \rho_a$.

Corollary 3.1 *The number m of eigenvalues of $A(t)$, counted with their algebraic multiplicities, located inside the curve Γ, is constant provided that $|t| < \rho_a$.*

In fact the condition on t is only a sufficient condition and it may be too restrictive since the real condition required is that $P(t)$ be continuous with respect to t.

While individual eigenvalues may not have an analytic behavior, their average is usually analytic. Consider the average

$$\hat{\lambda}(t) = \frac{1}{m}\sum_{i=1}^{m} \lambda_i(t)$$

of the eigenvalues $\lambda_1(t), \lambda_2(t), \ldots, \lambda_m(t)$ of $A(t)$ that are inside Γ where we assume that the eigenvalues are counted with their multiplicities. Let $B(t)$ be a matrix representation of the restriction of $A(t)$ to the invariant subspace $M(t) = \mathrm{Ran}(P(t))$. Note that since $M(t)$ is invariant under $A(t)$ then $B(t)$ is the matrix representation of the rank m transformation

$$A(t)_{|M(t)} = A(t)P(t)_{|M(t)} = P(t)A(t)_{|M(t)} = P(t)A(t)P(t)_{|M(t)}$$

and we have

$$\begin{aligned}\hat{\lambda}(t) \equiv \frac{1}{m}\mathrm{tr}[B(t)] &= \frac{1}{m}\mathrm{tr}[A(t)P(t)_{|M(t)}] \\ &= \frac{1}{m}\mathrm{tr}[A(t)P(t)] \qquad (3.19)\end{aligned}$$

The last equality in the above equation is due to the fact that for any x not in $M(t)$ we have $P(t)x = 0$ and therefore the extension of $A(t)P(t)$ to the whole space can only bring zero eigenvalues in addition to the eigenvalues $\lambda_i(t), i = 1, \ldots, m$.

Theorem 3.5 *The linear transformation $A(t)P(t)$ and its weighted trace $\hat{\lambda}(t)$ are analytic in the disk $|z| < \rho_a$.*

Proof. That $A(t)P(t)$ is analytic is a consequence of the previous theorem. That $\hat{\lambda}(t)$ is analytic, comes from the equivalent expression (3.19) and the fact that the trace of an operator $X(t)$ that is analytic with respect to t is analytic. ■

Therefore, a simple eigenvalue $\lambda(t)$ of $A(t)$ not only stays simple around a neighborhood of $t = 0$ but it is also analytic with respect to t. Moreover, the vector $u_i(t) = P_i(t)u_i$ is an eigenvector of $A(t)$ associated with this simple eigenvalue, with $u_i = u_i(0)$ being an eigenvector of A associated with the eigenvalue λ_i. Clearly, the eigenvector $u_i(t)$ is analytic with respect to the variable t. However, the situation is more complex for the case of a multiple eigenvalue. If an eigenvalue is of multiplicity m then after a small perturbation, it will split into at most m distinct small branches $\lambda_i(t)$. These branches taken individually are not analytic in general. On the other hand, their *arithmetic average* is analytic. For this reason it is critical, in practice, to try to recognize groups of eigenvalues that are likely to originate from the splitting of a perturbed multiple eigenvalue.

Example 3.1 That an individual branch of the m branches of eigenvalues $\lambda_i(t)$ is not analytic can be easily illustrated by the example

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \; H = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

The matrix $A(t)$ has the eigenvalues $\pm\sqrt{t}$ which degenerate into the double eigenvalue 0 as $t \rightarrow 0$. The individual eigenvalues are not analytic but their average remains constant and equal to zero.

In the above example each of the individual eigenvalues behaves like the square root of t around the origin. One may wonder whether this type of behavior can be generalized. The answer is stated in the next proposition.

Proposition 3.3 *Any eigenvalue $\lambda_i(t)$ of $A(t)$ inside the Jordan curve Γ satisfies*

$$|\lambda_i(t) - \lambda_i| = O\left(|t|^{1/l_i}\right)$$

where l_i is the index of λ_i.

Proof. Let $f(z) = (z - \lambda_i)^{l_i}$. We have seen earlier (proof of Lemma 3.1) that $f(A)P_i = 0$. For an eigenvector $u(t)$ of norm unity associated with the eigenvalue $\lambda_i(t)$ we have

$$\begin{aligned} f(A(t))P(t)u(t) &= f(A(t))u(t) = (A(t) - \lambda_i I)^{l_i} u(t) \\ &= (\lambda(t) - \lambda_i)^{l_i} u(t) \ . \end{aligned}$$

Taking the norms of both members of the above equation and using the fact that $f(A)P_i = 0$ we get

$$\begin{aligned} |\lambda_i(t) - \lambda_i|^{l_i} &= \|f(A(t))P(t)u(t)\| \\ &\leq \|f(A(t))P(t)\| = \|f(A(t))P(t) - f(A)P_i\| \ . \end{aligned}$$

Since $f(A) = f(A(0))$, $P_i = P(0)$ and $P(t), f(A(t))$ are analytic the right-hand-side in the above inequality is $O(t)$ and therefore

$$|\lambda_i(t) - \lambda_i|^{l_i} = O\ (|t|)$$

which shows the result. ∎

Example 3.2 A standard illustration of the above result is provided by taking A to be a Jordan block and H to be the rank one matrix $H = e_n e_1^T$:

$$A = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & 1 & \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix} \qquad H = \begin{pmatrix} 0 & & & & \\ & 0 & & & \\ & & 0 & & \\ & & & 0 & \\ 1 & & & & 0 \end{pmatrix} .$$

The matrix A has nonzero elements only in positions $(i, i+1)$ where they are equal to one. The matrix H has its elements equal to zero except for the element in position $(n, 1)$ which is equal to one. For $t = 0$ the matrix $A + tH$ admits only the eigenvalue $\lambda = 0$. The characteristic polynomial of $A + tH$ is equal to

$$p_t(z) = \det(A + tH - zI) = (-1)^n (z^n - t)$$

and its roots are $\lambda_j(t) = t^{1/n} e^{\frac{2ij\pi}{n}}$ $j = 1, \ldots, n$. Thus, if $n = 20$ then for a perturbation on A of the order of 10^{-16}, a reasonable number if double precision arithmetic is used, the eigenvalue will be perturbed by as much as 0.158..

2. A-Posteriori Error Bounds

In this section we consider the problem of predicting the error made on an eigenvalue/eigenvector pair from some a posteriori knowledge on their approximations. The simplest criterion used to determine the accuracy of an approximate eigenpair $\tilde{\lambda}, \tilde{u}$, is to compute the norm of the so called residual vector

$$r = A\tilde{u} - \tilde{\lambda}\tilde{u}.$$

The aim is to derive error bounds that relate some norm of r, typically its 2-norm, to the errors on the eigenpair. Such error bounds are referred to a posteriori error bounds. Such bounds may help determine how accurate the approximations provided by some algorithm may be. This information can in turn be helpful in choosing a stopping criterion in iterative algorithms, in order to ensure that the answer delivered by the numerical method is within a desired tolerance.

2.1. General Error Bounds

In the non-Hermitian case there does not exist any 'a posteriori' error bounds in the strict sense of the definition. The error bounds that exist are in general weaker and not as easy to use as those known in the Hermitian case. The first error bound which we consider is known as the Bauer-Fike theorem. We recall that the condition number of a matrix X relative to the p-norm is defined by $\text{Cond}_p(X) = \|X\|_p \|X^{-1}\|_p$.

Theorem 3.6 (Bauer-Fike) *Let $\tilde{\lambda}, \tilde{u}$ be an approximate eigenpair of A with residual vector $r = A\tilde{u} - \tilde{\lambda}\tilde{u}$, where $\tilde{u}$ is of 2-norm* *unity. Moreover, assume that the matrix A is diagonalizable and let X be the matrix that transforms it into diagonal form. Then, there exists an eigenvalue λ of A such that*

$$|\lambda - \tilde{\lambda}| \leq \text{Cond}_2(X)\|r\|_2 \ .$$

Proof. If $\tilde{\lambda} \in \sigma(A)$ the result is true. Assume that $\tilde{\lambda}$ is not an eigenvalue. From $A = XDX^{-1}$, where D is the diagonal of eigenvalues and since we assume that $\lambda \notin \sigma(A)$, we can write

$$\tilde{u} = (A - \tilde{\lambda}I)^{-1}r = X(D - \tilde{\lambda}I)^{-1}X^{-1}r$$

and hence

$$\begin{aligned} 1 &= \|X(D - \tilde{\lambda}I)^{-1}X^{-1}r\|_2 \\ &\leq \|X\|_2\|X^{-1}\|_2\|(D - \tilde{\lambda}I)^{-1}\|_2 \ \|r\|_2 \ . \end{aligned} \tag{3.20}$$

The matrix $(D - \tilde{\lambda}I)^{-1}$ is a diagonal matrix and as a result its 2-norm is the maximum of the absolute values of its diagonal entries. Therefore,

$$1 \leq \text{Cond}_2(X)\|r\|_2 \max_{\lambda_i \in \sigma(A)} |\lambda_i - \tilde{\lambda}|^{-1}$$

from which the result follows. ∎

In case the matrix is not diagonalizable then the previous result can be generalized as follows.

Theorem 3.7 *Let $\tilde{\lambda}, \tilde{u}$ an approximate eigenpair with residual vector $r = A\tilde{u} - \tilde{\lambda}\tilde{u}$, where $\tilde{u}$ is of 2-norm unity. Let X be the matrix that transforms A into its Jordan canonical form, $A = XJX^{-1}$. Then, there exists an eigenvalue λ of A such that*

$$\frac{|\lambda - \tilde{\lambda}|^l}{1 + |\lambda - \tilde{\lambda}| + \cdots + |\lambda - \tilde{\lambda}|^{l-1}} \leq \text{Cond}_2(X)\|r\|_2$$

where l is the index of λ.

Proof. The proof starts as in the previous case but here the diagonal matrix D is replaced by the Jordan matrix J. Because the matrix $(J-\tilde{\lambda}I)$ is block diagonal its 2-norm is the maximum of the 2-norms of each block (a consequence of the alternative formulation for 2-norms seen in Chapter I). For each of these blocks we have

$$(J_i - \tilde{\lambda}I)^{-1} = ((\lambda_i - \tilde{\lambda})I + E)^{-1}$$

where E is the nilpotent matrix having ones in positions $(i, i+1)$ and zeros elsewhere. Therefore,

$$(J_i - \tilde{\lambda}I)^{-1} = \sum_{j=1}^{l_i} (\lambda_i - \tilde{\lambda})^{-j} (-E)^{j-1}$$

and as a result, setting $\delta_i = |\lambda_i - \tilde{\lambda}|$ and noting that $\|E\|_2 = 1$, we get

$$\|(J_i - \tilde{\lambda}I)^{-1}\|_2 \ \le \ \sum_{j=1}^{l_i} |\lambda_i - \tilde{\lambda}|^{-j} \|E\|_2^{j-1} \ = \ \sum_{j=1}^{l_i} \delta_i^{-j} \ = \ \delta_i^{-l_i} \sum_{j=0}^{l_i - 1} \delta_i^j .$$

The analogue of (3.20) is

$$1 \le \mathrm{Cond}_2(X) \|(J - \tilde{\lambda}I)^{-1}\|_2 \|r\|_2 . \tag{3.21}$$

Since,

$$\|(J - \tilde{\lambda}I)^{-1}\|_2 \ = \max_{i=1,\ldots,p} \|(J_i - \tilde{\lambda}I)^{-1}\|_2 \ \le \max_{i=1,\ldots,p} \delta_i^{-l} \sum_{j=0}^{l_i - 1} \delta_i^j$$

we get

$$\min_{i=1,\ldots,p} \left\{ \frac{\delta_i^{l_i}}{\sum_{j=0}^{l_i-1} \delta_i^j} \right\} \le \ \mathrm{Cond}_2(X) \|r\|_2$$

which is essentially the desired result. ■

Corollary 3.2 (Kahan, Parlett, and Jiang, 1980). *Under the same assumptions as those of theorem 3.7, there exists an eigenvalue λ of A such that*

$$\frac{|\lambda - \tilde{\lambda}|^l}{(1 + |\lambda - \tilde{\lambda}|)^{l-1}} \;\leq\; \mathrm{Cond}_2(X) \|r\|_2$$

where l is the index of λ.

Proof. Follows immediately from the previous theorem and the inequality,

$$\sum_{j=0}^{l-1} \delta_i^j \;\leq\; (1 + \delta_i)^{l-1}.$$

∎

For an alternative proof see [82]. Unfortunately, the bounds of the type shown in the previous two theorems are not practical because of the presence of the condition number of X. The second result even requires the knowledge of the index of λ_i, which is not numerically viable. The situation is much improved in the particular case where A is Hermitian because in this case $\mathrm{Cond}_2(X) = 1$. This is taken up next.

2.2. The Hermitian Case

In the Hermitian case, Theorem 3.6 leads to the following corollary.

Corollary 3.3 *Let $\tilde{\lambda}, \tilde{u}$ be an approximate eigenpair of a Hermitian matrix A, with $\|u\|_2 = 1$ and let r be the corresponding residual vector. Then there exists an eigenvalue of A such that*

$$|\lambda - \tilde{\lambda}| \;\leq\; \|r\|_2 \,. \tag{3.22}$$

This is a remarkable result because it constitutes a simple yet general error bound. On the other hand it is not sharp as the next a posteriori error bound, due to Kato and Temple [84, 175], shows. We start by proving a lemma that will be used to prove Kato-Temple's theorem. In the next results it is assumed that the approximate eigenvalue $\tilde{\lambda}$ is the Rayleigh quotient of the approximate eigenvector.

Lemma 3.2 *Let $\tilde{u}$ be an approximate eigenvector of norm unity of A, and $\tilde{\lambda} = (A\tilde{u}, \tilde{u})$. Let (α, β) be an interval that contains $\tilde{\lambda}$ and no eigenvalue of A. Then*

$$(\beta - \tilde{\lambda})(\tilde{\lambda} - \alpha) \ \leq \ \|r\|_2^2.$$

Proof. This lemma uses the observation that the residual vector r is orthogonal to $\tilde{u}$. Then we have

$$\begin{aligned} &((A - \alpha I)\tilde{u}, (A - \beta I)\tilde{u}) \\ &\qquad = ((A - \tilde{\lambda} I)\tilde{u} + (\tilde{\lambda} - \alpha I)\tilde{u}, ((A - \tilde{\lambda} I)\tilde{u} + (\tilde{\lambda} - \beta I)\tilde{u}) \\ &\qquad = \|r\|_2^2 + (\tilde{\lambda} - \alpha I)(\tilde{\lambda} - \beta I), \end{aligned}$$

because of the orthogonality property mentioned above. On the other hand, one can expand $\tilde{u}$ in the orthogonal eigenbasis of A as

$$\tilde{u} = \xi_1 u_1 + \xi_2 u_2 + \cdots + \xi_n u_n$$

to transform the left hand side of the expression into

$$((A - \alpha I)\tilde{u}, (A - \beta I)\tilde{u}) = \sum_{i=1}^{n} |\xi_i|^2 \ (\lambda_i - \alpha)(\lambda_i - \beta) \ .$$

Each term in the above sum is nonnegative because of the assumptions on α and β. Therefore $\|r\|_2^2 + (\beta - \tilde{\lambda})(\tilde{\lambda} - \alpha) \geq 0$ which is the desired result. ■

Theorem 3.8 (Kato and Temple [84, 175]) *Let $\tilde{u}$ be an approximate eigenvector of norm unity of A, and $\tilde{\lambda} = (A\tilde{u}, \tilde{u})$. Assume that we know an interval (a, b) that contains $\tilde{\lambda}$ and one and only one eigenvalue λ of A. Then*

$$-\frac{\|r\|_2^2}{\tilde{\lambda} - a} \leq \tilde{\lambda} - \lambda \leq \frac{\|r\|_2^2}{b - \tilde{\lambda}} .$$

Proof. Let λ be the closest eigenvalue to $\tilde{\lambda}$. In the case where λ is located at left of $\tilde{\lambda}$ then take $\alpha = \lambda$ and $\beta = b$ in the lemma to get

$$0 \leq \tilde{\lambda} - \lambda \leq \frac{\|r\|_2^2}{b - \tilde{\lambda}} .$$

In the opposite case where $\lambda > \tilde{\lambda}$, use $\alpha = a$ and $\beta = \lambda$ to get

$$0 \leq \lambda - \tilde{\lambda} \leq \frac{\|r\|_2^2}{\tilde{\lambda} - a} .$$

This completes the proof. ■

A simplification of Kato-Temple's theorem consists of using a particular interval that is symmetric about the approximation $\tilde{\lambda}$, as is stated in the next corollary.

Corollary 3.4 *Let $\tilde{u}$ be an approximate eigenvector of norm unity of A, and $\tilde{\lambda} = (A\tilde{u}, \tilde{u})$. Let λ be the eigenvalue closest to $\tilde{\lambda}$ and δ the distance from $\tilde{\lambda}$ to the rest of the spectrum, i.e.,*

$$\delta = \min_i \{|\lambda_i - \tilde{\lambda}|, \ \lambda_i \neq \lambda\}.$$

Then,

$$|\tilde{\lambda} - \lambda| \leq \frac{\|r\|_2^2}{\delta} . \tag{3.23}$$

Proof. This is a particular case of the previous theorem with $a = \tilde{\lambda} - \delta$ and $b = \tilde{\lambda} + \delta$. ■

It is also possible to show a similar result for the angle between the exact and approximate eigenvectors.

Theorem 3.9 *Let $\tilde{u}$ be an approximate eigenvector of norm unity of A, $\tilde{\lambda} = (A\tilde{u}, \tilde{u})$ and $r = (A - \tilde{\lambda} I)\tilde{u}$. Let λ be the eigenvalue closest to $\tilde{\lambda}$ and δ the distance from $\tilde{\lambda}$ to the rest of the spectrum, i.e., $\delta = \min_i \{|\lambda_i - \tilde{\lambda}|, \lambda_i \neq \lambda\}$. Then, if u is an eigenvector of A associated with λ we have*

$$\sin \theta(\tilde{u}, u) \ \le \ \frac{\|r\|_2}{\delta} \ . \tag{3.24}$$

Proof. Let us write the approximate eigenvector $\tilde{u}$ as $\tilde{u} = u \cos \theta + z \sin \theta$ where z is a vector orthogonal to u. We have

$$\begin{aligned} (A - \tilde{\lambda} I)\tilde{u} &= \cos\theta \ (A - \tilde{\lambda} I)u + \sin\theta \ (A - \tilde{\lambda} I)z \\ &= \cos\theta \ (\lambda - \tilde{\lambda} I)u + \sin\theta \ (A - \tilde{\lambda} I)z \ . \end{aligned}$$

The two vectors on the right hand side are orthogonal to each other because,

$$(u, (A - \tilde{\lambda} I)z) = ((A - \tilde{\lambda} I)u, z) = (\lambda - \tilde{\lambda})(u, z) = 0 \ .$$

Therefore,

$$\|r\|_2^2 = \|(A - \tilde{\lambda} I)\tilde{u}\|^2 = \sin^2\theta \ \|(A - \tilde{\lambda} I)z\|_2^2 + \cos^2\theta \ |\lambda - \tilde{\lambda}|^2 \ .$$

Hence,

$$\sin^2\theta \ \|(A - \tilde{\lambda} I)z\|_2^2 \le \|r\|_2^2 \ .$$

The proof follows by observing that since z is orthogonal to u then $\|(A - \tilde{\lambda} I)z\|_2$ is larger than the smallest eigenvalue of $A - \tilde{\lambda} I$ restricted to the subspace orthogonal to u, which is precisely δ. ∎

Although the above bounds for the Hermitian case are sharp they are still not computable since δ involves a distance from the 'next closest' eigenvalue of A to $\tilde{\lambda}$ which is not readily available.

In order to be able to use these bounds in practical situations one must provide a lower bound for the distance δ. One might simply approximate δ by $\tilde{\lambda} - \tilde{\lambda}_j$ where $\tilde{\lambda}_j$ is some approximation to the next closest eigenvalue to $\tilde{\lambda}$. The result would no longer be an actual upper bound on the error but rather an 'estimate' of the error. This may not be safe however. To ensure that the computed error bound used is rigorous it is preferable to exploit the simpler inequality provided by Corollary 3.3 in order to find a lower bound for the distance δ, for example

$$\begin{aligned} \delta = |\tilde{\lambda} - \lambda_j| &\geq |(\tilde{\lambda} - \tilde{\lambda}_j) + (\lambda_j - \tilde{\lambda}_j)| \\ &\geq |\tilde{\lambda} - \tilde{\lambda}_j| - |\lambda_j - \tilde{\lambda}_j| \\ &\geq |\tilde{\lambda} - \tilde{\lambda}_j| - \|r_j\|_2. \end{aligned}$$

where $\|r_j\|_2$ is the residual norm associated with the eigenvalue λ_j. Now the above lower bound of δ is computable. In order for the resulting error bound to have a meaning, $\|r_j\|_2$ must be small enough to ensure that there are no other potential eigenvalues λ_k that might be closer to λ than is λ_j. The above error bounds when used cautiously can be quite useful.

Example 3.3 Let

$$A = \begin{pmatrix} 1.0 & 2.0 & & & \\ 2.0 & 1.0 & 2.0 & & \\ & 2.0 & 1.0 & 2.0 & \\ & & 2.0 & 1.0 & 2.0 \\ & & & 2.0 & 1.0 \end{pmatrix} .$$

The eigenvalues of A are $\{3, -1, 1, 1 - 2\sqrt{3}, 1 + 2\sqrt{3}\}$

An eigenvector associated with the eigenvalue $\lambda = 3$ is

$$u = \begin{pmatrix} -0.5 \\ -0.5 \\ 0.0 \\ 0.5 \\ 0.5 \end{pmatrix} .$$

Consider the vector

$$\tilde{u} = \begin{pmatrix} -0.49 \\ -0.5 \\ 0.0 \\ 0.5 \\ 0.5 \end{pmatrix} .$$

The Rayleigh quotient of $\tilde{u}$ with respect to A is $\tilde{\lambda} = 2.9998...$ The closest eigenvalue is $\lambda = 3.0$ and the corresponding actual error is 2.02×10^{-4}. The residual norm is found to be

$$\|(A - \tilde{\lambda} I)\tilde{u}\|_2 \approx 0.0284 \ .$$

The distance δ here is

$$\delta = |2.9998 - 4.464101...| \approx 1.46643 \ .$$

So the error bound for the eigenvalue 2.9998 found is

$$\frac{(0.0284..)^2}{1.4643} \approx 5.5177 \times 10^{-4}.$$

For the eigenvector, the angle between the exact and approximate eigenvector is such that $\cos\theta = 0.999962$, giving an angle $\theta \approx 0.0087$ and the sine of the angle is approximately $\sin\theta \approx 0.0087$. The error as estimated by (3.9) is

$$\sin\theta \le \frac{0.0284}{1.4643} \approx 0.01939$$

which is about twice as large as the actual error.

We now consider a slightly more realistic situation. There are instances in which the off-diagonal elements of a matrix are small. Then the diagonal elements can be considered approximations to the eigenvalues of A and the question is how good an accuracy can one expect? We illustrate this with an example.

Example 3.4 Let

$$A = \begin{pmatrix} 1.00 & 0.0055 & 0.10 & 0.10 & 0.00 \\ 0.0055 & 2.00 & -0.05 & 0.00 & -0.10 \\ 0.10 & -0.05 & 3.00 & 0.10 & 0.05 \\ 0.10 & 0.00 & 0.10 & 4.00 & 0.00 \\ 0.00 & -0.10 & 0.05 & 0.00 & 5.00 \end{pmatrix} .$$

The eigenvalues of A rounded to 6 digits are

$$\sigma(A) = \{0.99195, 1.99443, 2.99507, 4.01386, 5.00466\} \ .$$

A natural question is how accurate is each of the diagonal elements of A as an approximate eigenvalue? We assume that we know nothing about the exact spectrum. We can take as approximate eigenvectors the e_i's, $i = 1, \cdots, 5$ and the corresponding residual norms are

$$0.141528 \ ; 0.1119386 \ ; 0.1581139 \ ; 0.1414214 \ ; 0.1118034$$

respectively. The simplest residual bound (3.22) tells us that

$$\begin{array}{ll} |\lambda - 1.0| \le 0.141528; & |\lambda - 2.0| \le 0.111939; \\ |\lambda - 3.0| \le 0.158114; & |\lambda - 4.0| \le 0.141421; \\ |\lambda - 5.0| \le 0.111803. & \end{array}$$

The intervals defined above are all disjoint. As a result, we can get a reasonable idea of δ_i the distance of each of the approximations from the eigenvalues not in the interval. For example,

$$\delta_1 \equiv |a_{11} - \lambda_2| \ge |1 - (2.0 - 0.1119386)| \approx 0.88806$$

and

$$\begin{array}{rcl} \delta_2 & = & \min\{|a_{22} - \lambda_3|, |a_{22} - \lambda_1|\} \\ & \ge & \min\{|2.0 - (3.0 - 0.15811)|, |2.0 - (1.0 + 0.14153)|\} \\ & = & 0.8419... \end{array}$$

We find similarly $\delta_3 \ge 0.8585$, $\delta_4 \ge 0.8419$, and $\delta_5 \ge 0.8586$.

We now get from the bounds (3.23) the following inequalities,

$$\begin{array}{ll} |\lambda - 1.0| \le 0.0226; & |\lambda - 2.0| \le 0.0149; \\ |\lambda - 3.0| \le 0.0291; & |\lambda - 4.0| \le 0.0238; \\ |\lambda - 5.0| \le 0.0146. & \end{array}$$

whereas the actual errors are

$$\begin{array}{ll} |\lambda - 1.0| \approx 0.0080; & |\lambda - 2.0| \approx 0.0056; \\ |\lambda - 3.0| \approx 0.0049; & |\lambda - 4.0| \approx 0.0139; \\ |\lambda - 5.0| \approx 0.0047. & \end{array}$$

2.3. The Kahan-Parlett-Jiang Theorem

We now return to the general non-Hermitian case. The results seen for the Hermitian case in the previous section can be very useful in practical situations. For example they can help develop efficient stopping criteria in iterative algorithms. In contrast, those seen in Section 2.1 for the general non-Hermitian case are not too easy to exploit in practice. The question that one might ask is whether or not any residual bounds can be established that will provide information similar to that provided in the Hermitian case. There does not seem to exist any such result in the literature. A result established by Kahan, Parlett and Jiang [82], which we now discuss, seems to be the best compromise between generality and sharpness. However, the theorem is of a different type. It does not guarantee the existence of, say, an eigenvalue in a given interval whose size depends on the residual norm. It only gives us the size of the smallest perturbation that must be applied to the original data (the matrix), in order to transform the approximate eigenpair into an exact one (for the perturbed problem).

To explain the nature of the theorem we begin with a very simple result which can be regarded as a one-sided version of the one proved by Kahan, Jiang and Parlett, in that it only considers the right eigenvalue – eigenvector pair instead of the eigen-triplet consisting of the eingenvalue and the right and left eigenvectors.

Proposition 3.4 *Let a square matrix A and a unit vector u be given. For any scalar γ define the residual vector,*

$$r = Au - \gamma u,$$

and let $\mathcal{E} = \{E : (A - E)u = \gamma u\}$. Then

$$\min_{E\in\mathcal{E}} \|E\|_2 = \|r\|_2 \ . \tag{3.25}$$

Proof. From the assumptions we see that each E is in $\mathcal{E}$ if and only if it satisfies the equality

$$Eu = r \ . \tag{3.26}$$

Since $\|u\|_2 = 1$ the above equation implies that for any such E

$$\|E\|_2 \geq \|r\|_2,$$

which in turn implies that

$$\min_{E \in \mathcal{E}} \|E\|_2 \geq \|r\|_2. \tag{3.27}$$

Now consider the matrix $E_0 = ru^H$ which is a member of $\mathcal{E}$ since it satisfies (3.26). The 2-norm of E_0 is such that

$$\|E_0\|_2^2 = \sigma_{max}\{ru^H ur^H\} = \sigma_{max}\{rr^H\} = \|r\|_2^2.$$

As a result the minimum in the left hand side of (3.27) is reached for $E = E_0$ and the value of the minimum is equal to $\|r\|_2$. ■

We now state a simple version of the Kahan-Parlett-Jiang theorem [82].

Theorem 3.10 (Kahan, Parlett, and Jiang) *Let a square matrix A and two unit vectors u, w with $(u, w) \neq 0$ be given. For any scalar γ define the residual vectors,*

$$r = Au - \gamma u \qquad s = A^H w - \bar{\gamma} w$$

and let $\mathcal{E} = \{E : (A - E)u = \gamma u; (A - E)^H w = \bar{\gamma} w\}$. Then

$$\min_{E \in \mathcal{E}} \|E\|_2 = \max \ \{\|r\|_2, \|s\|_2\} \ . \tag{3.28}$$

Proof. We proceed in the same way as for the proof of the simpler result of the previous proposition. The two conditions that a matrix E must satisfy in order to belong to $\mathcal{E}$ translate into

$$Eu = r \quad \text{and} \quad E^H w = s. \tag{3.29}$$

By the same argument used in the proof of Proposition 2.4, any such E satisfies

$$\|E\|_2 \geq \|r\|_2 \quad \text{and} \quad \|E\|_2 \geq \|s\|_2. \tag{3.30}$$

which proves the inequality

$$\min_{E \in \mathcal{E}} \|E\|_2 \geq \max\{\|r\|_2, \|s\|_2\}. \tag{3.31}$$

We now define,

$$\begin{aligned} \delta &= s^H u = w^H r \qquad (3.32) \\ x &= r - \delta\ w \\ y &= s - \bar{\delta}\ u \end{aligned}$$

and consider the particular set of matrices of the form

$$E(\beta) = ru^H + ws^H - \delta\ wu^H - \beta\ xy^H \tag{3.33}$$

where β is a parameter. It is easy to verify that these matrices satisfy the constraints (3.29) for any β.

We distinguish two different cases depending on whether $\|s\|_2$ is larger or smaller than $\|r\|_2$. When $\|s\|_2 > \|r\|_2$ we rewrite $E(\beta)$ in the form

$$E(\beta) = x(u - \beta\ y)^H + ws^H \tag{3.34}$$

and select β in such a way that

$$s^H(u - \beta\ y) = 0 \tag{3.35}$$

which leads to

$$\beta = \frac{\delta}{\|s\|_2^2 - |\delta|^2}.$$

We note that the above expression is not valid when $\|s\|_2 = |\delta|$, which occurs only when $y = 0$. In this situation $E(\beta) = ru^H$ for any β, and the following special treatment is necessary. As in the proof of the previous proposition $E(\beta) = \|r\|_2$. On the other hand we have

$$\|s\|_2 = |\delta| = |w^H r| \le \|r\|_2$$

which shows that $\max\{\|r\|_2, \|s\|_2\} = \|r\|_2$ and establishes the result that the minimum in the theorem is reached for $E(\beta)$ in this very special case.

Going back to the general case where $\|s\|_2 \neq |\delta|$, with the above choice of β the two vectors x and w in the range of $E(\beta)$ as defined by (3.34) are orthogonal and similarly, the vectors $u - \beta y$ and s are also orthogonal. In this situation the norm of $E(\beta)$ is equal to [See problem P-2.14]:

$$\|E(\beta)\|_2 = \max\{\|s\|_2, \|x\|_2\|\|u^H - \beta\ y\|_2\}.$$

Because of the orthogonality of x and w, we have

$$\|x\|_2^2 = \|r\|_2^2 - |\delta|^2\ .$$

Similarly, exploiting the orthogonality of the pair u, y, and using the definition of β we get

$$\begin{aligned}\|u - \beta\ y\|_2^2 &= 1 + \beta^2\|y\|_2^2 \\ &= 1 + \beta^2[\|s\|_2^2 - |\delta|^2] \\ &= \frac{\|s\|_2^2}{\|s\|_2^2 - |\delta|^2}\ .\end{aligned}$$

The above results yield

$$\|E(\beta)\|_2^2 = \max\ \left\{\|s\|_2^2, \|s\|_2^2\frac{\|r\|_2^2 - |\delta|^2}{\|s\|_2^2 - |\delta|^2}\right\} = \|s\|_2^2.$$

This shows from (3.31) that the equality (3.28) is satisfied for the case when $\|s\|_2 > \|r\|_2$.

To prove the result for the case $\|s\|_2 < \|r\|_2$, we proceed in the same manner, writing this time $E(\beta)$ as

$$E(\beta) = ru^H + (\alpha w - \beta\ x)y^H$$

and choosing β such that $u^H(w - \beta\ x) = 0$. A special treatment will also be necessary for the case where $\|r\|_2 = |\delta|$ which only occurs when $x = 0$. ■

The actual result proved by Kahan, Parlett and Jiang is essentially a block version of the above theorem and includes results with other norms, such as the Frobenius norm.

Example 3.5 Consider the matrix,

$$A = \begin{pmatrix} 1.0 & 2.1 & & & \\ 1.9 & 1.0 & 2.1 & & \\ & 1.9 & 1.0 & 2.1 & \\ & & 1.9 & 1.0 & 2.1 \\ & & & 1.9 & 1.0 \end{pmatrix}.$$

which is obtained by perturbing the symmetric tridiagonal matrix of Example 3.3. Consider the pair

$$\gamma = 3.0, \qquad v = \begin{pmatrix} -0.5 \\ -0.5 \\ 0.0 \\ 0.5 \\ 0.5 \end{pmatrix}.$$

Then we have

$$\|r\|_2 = \|(A - \gamma I)u\|_2 \approx 0.1414,$$

which tells us, using the one-sided result (Proposition 3.4), that we need to perturb A by a matrix E of norm 0.1414 to make the pair γ, v an exact eigenpair of A.

Consider now v as defined above and

$$w = \alpha\ \ (0.6, 0.6, 0.0, 0.4, 0.4)^T\ ,$$

where α is chosen to normalize w to so that its 2-norm is unity. Then, still with $\gamma = 3$, we find

$$\|r\|_2 \approx 0.1414 \ , \quad \|s\|_2 \approx 0.5004 \ .$$

As a result of the theorem, we now need a perturbation E whose 2-norm is roughly 0.5004 to make the triplet γ, v, w an exact eigentriplet of A, a much stricter requirement than with the one-sided result.

The outcome of the above example was to be expected. If one of the left of right approximate eigen-pair, for example the left pair (γ, v), is a poor approximation, then it will take a larger perturbation on A to make the triplet γ, v, w exact, than it would to make the pair γ, u exact. Whether one needs to use the one-sided or the two-sided result depends on whether one is interested in the left and right eigenvectors simultaneously or in the right (or left) eigenvector only.

3. Conditioning of Eigen-problems

When solving a linear system $Ax = b$, an important question that arises is how sensitive is the solution x to small variations of the initial data, namely to the matrix A and the right-hand side b. A measure of this sensitivity is called the condition number of A defined by

$$\text{Cond}(A) = \|A\| \|A^{-1}\|$$

relative to some norm.

For the eigenvalue problem we raise a similar question but we must now define similar measures for the eigenvalues as well as for the eigenvectors and the invariant subspaces.

3.1. Conditioning of Eigenvalues

Let us assume that λ is a simple eigenvalue and consider the family of matrices $A(t) = A + tE$. We know from the previous sections that there exists a branch of eigenvalues $\lambda(t)$ of $A(t)$ that is analytic with respect to t, when t belongs to a small enough disk centered at the origin. It is natural to call conditioning of the eigenvalue λ of A relative to the perturbation E the modulus of the derivative of $\lambda(t)$ at the origin $t = 0$. Let us write

$$A(t)u(t) = \lambda(t)u(t) \tag{3.36}$$

and take the inner product of both members with a left eigenvector w of A associated with λ to get

$$((A + tE)u(t), w) = \lambda(t)(u(t), w)$$

or,

$$\begin{aligned} \lambda(t)(u(t), w) &= (Au(t), w) + t(Eu(t), w) \\ &= (u(t), A^H w) + t(Eu(t), w) \\ &= \lambda(u(t), w) + t(Eu(t), w). \end{aligned}$$

Hence,

$$\frac{\lambda(t) - \lambda}{t}(u(t), w) \ = (Eu(t), w)$$

and therefore by taking the limit at $t = 0$,

$$\lambda'(0) = \frac{(Eu, w)}{(u, w)}$$

Here we should recall that the left and right eigenvectors associated with a simple eigenvalue cannot be orthogonal to each other. The actual conditioning of an eigenvalue, given a perturbation "in the direction of E"is the modulus of the above quantity. In practical situations, one often does not know the actual perturbation E but only its magnitude, e.g., as measured by some matrix norm $\|E\|$. Using the Cauchy-Schwartz inequality and the 2-norm, we can derive the following upper bound,

$$|\lambda'(0)| \le \frac{\|Eu\|_2 \|w\|_2}{|(u, w)|} \le \|E\|_2 \frac{\|u\|_2 \|w\|_2}{|(u, w)|}$$

In other words the actual condition number of the eigenvalue λ is bounded from above by the norm of E divided by the cosine of the acute angle between the left and the right eigenvectors associated with λ. Hence the following definition.

Definition 3.1 *The condition number of a simple eigenvalue λ of an arbitrary matrix A is defined by*

$$\mathrm{Cond}(\lambda) = \frac{1}{\cos\theta(u,w)}$$

in which u and w are the right and left eigenvectors, respectively, associated with λ.

Example 3.6 Consider the matrix

$$A = \begin{pmatrix} -149 & -50 & -154 \\ 537 & 180 & 546 \\ -27 & -9 & -25 \end{pmatrix}$$

The eigenvalues of A are $\{1, 2, 3\}$. The right and left eigenvectors of A associated with the eigenvalue $\lambda_1 = 1$ are approximately

$$u = \begin{pmatrix} 0.3162 \\ -0.9487 \\ 0.0 \end{pmatrix} \quad \text{and} \quad w = \begin{pmatrix} 0.6810 \\ 0.2253 \\ 0.6967 \end{pmatrix} \tag{3.37}$$

and the corresponding condition number is approximately

$$\mathrm{Cond}(\lambda_1) \approx 603.64$$

A perturbation of order 0.01 may cause perturbations of magnitude up to 6. Perturbing a_{11} to -149.01 yields the spectrum:

$$\{0.2287, 3.2878, 2.4735\}.$$

For Hermitian, or more generally normal, matrices every simple eigenvalue is well-conditioned, since $\mathrm{Cond}(\lambda) = 1$. On the other hand the condition number of a non-normal matrix can be excessively high, in fact arbitrarily high.

Example 3.7 As an example simply consider the matrix

$$\begin{pmatrix} \lambda_1 & -1 & & & \\ & \lambda_2 & -1 & & \\ & & . & . & \\ & & & . & -1 \\ & & & & \lambda_n \end{pmatrix} \tag{3.38}$$

with $\lambda_1 = 0$ and $\lambda_i = 1/(i-1)$ for $i > 1$. A right eigenvector associated with the eigenvalue λ_1 is the vector e_1. A left eigenvector is the vector w whose i-th component is equal to $(i-1)!$ for $i = 1, \ldots, n$. A little calculation shows that the condition number of λ_1 satisfies

$$(n-1)! \le \text{Cond}(\lambda_1) \le (n-1)! \ \sqrt{n}.$$

Thus, this example shows that the condition number can be quite large even for modestly sized matrices.

An important comment should be made concerning the above example. The eigenvalues of A are explicitly known in terms of the diagonal entries of the matrix, whenever the structure of A stays the same. One may wonder whether it is sensible to discuss the concept of condition number in such cases. For example, if we perturb the (1,1) elements by 0.1 we know exactly that the eigenvalue λ_1 will be perturbed likewise. Is the notion of condition number useless in such situations? The answer is no. First, the argument is only true if perturbations are applied in specific positions of the matrix, namely its upper triangular part. If perturbations take place elsewhere then some or all of the eigenvalues of the perturbed matrix may not be explicitly known. Second, one can think of applying an orthogonal similarity transformation to A. If Q is orthogonal then the eigenvalues of the matrix $B = Q^H AQ$ have the same condition number as those of the original matrix A, (see Problem P-3.15). The resulting matrix B may be dense and the dependence of its eigenvalues with respect to its entries is no longer explicit.

3.2. Conditioning of Eigenvectors

To properly define the condition number of an eigenvector we need to use the notion of *reduced resolvent*. Although the resolvent operator $R(z)$ has a singularity at an eigenvalue λ it can still be defined on the restriction to the invariant subspace $\text{Ker}(P)$. More precisely, consider the restriction of the mapping $A - \lambda I$ to the subspace $(I - P)\mathbb{C}^n = \text{Ker}(P)$, where P is the spectral projector associated with the eigenvalue λ. This mapping is invertible because if x is an element of $\text{Ker}(P)$ then $(A - \lambda I)x = 0$, i.e., x is in $\text{Ker}(A - \lambda I)$ which is included in $\text{Ran}(P)$ and this is only possible when $x = 0$. We will call reduced resolvent at λ the inverse of this linear mapping and we will denote it by $S(\lambda)$. Thus,

$$S(\lambda) = \left[(A - \lambda I)_{|\text{Ker}(P)}\right]^{-1} .$$

The reduced resolvent satisfies the relation,

$$S(\lambda)(A-\lambda I)x = S(\lambda)(A-\lambda I)(I-P)x = (I-P)x \quad \forall \ x \tag{3.39}$$

which can be viewed as an alternative definition of $S(\lambda)$.

We now consider a simple eigenvalue λ of a matrix A with an associated eigenvector u, and write that a pair $\lambda(t), u(t)$ is an eigenpair of the matrix $A + tE$,

$$(A + tE)u(t) = \lambda(t)u(t) \ . \tag{3.40}$$

Subtracting $Au = \lambda u$ from both sides we have,

$$A(u(t)-u)+tEu(t) = \lambda(t)u(t)-\lambda u = \lambda(u(t)-u)+(\lambda(t)-\lambda)u(t)$$

or,

$$(A - \lambda I)(u(t) - u) + tEu(t) = (\lambda(t) - \lambda)u(t) \ .$$

We then multiply both sides by the projector $I - P$ to obtain

$$\begin{aligned}(I-P)(A-\lambda I)(u(t)-u) \ &+ \ t(I-P)Eu(t) \\ &= \ (\lambda(t)-\lambda)(I-P)u(t) \\ &= \ (\lambda(t)-\lambda)(I-P)(u(t)-u)\end{aligned}$$

The last equality holds because $(I-P)u = 0$ since u is in $\mathrm{Ran}(P)$. Hence,

$$\begin{aligned}&(A-\lambda I)(I-P)(u(t)-u) = \\ &\quad (I-P)\left[-tEu(t) + (\lambda(t)-\lambda)(u(t)-u)\right].\end{aligned}$$

We now multiply both sides by $S(\lambda)$ and use (3.39) to get

$$\begin{aligned}&(I-P)(u(t)-u) = \\ &\quad S(\lambda)(I-P)\left[-tEu(t) + (\lambda(t)-\lambda)(u(t)-u)\right]\end{aligned} \tag{3.41}$$

In the above development we have not scaled $u(t)$ in any way. We now do so by requiring that its projection onto the eigenvector u be exactly u, i.e., $Pu(t) = u$ for all t. With this scaling, we have

$$(I-P)(u(t)-u) = u(t)-u.$$

As a result, equality (3.42) becomes

$$u(t)-u = S(\lambda)\left[-t(I-P)Eu(t) + (\lambda(t)-\lambda)(u(t)-u),\right]$$

from which we finally get, after dividing by t and taking the limit,

$$u'(0) = -S(\lambda)(I-P)Eu\ . \tag{3.42}$$

Using the same argument as before, we arrive at the following general definition of the condition number of an eigenvector.

Definition 3.2 *The condition number of an eigenvector u associated with an eigenvalue λ of an arbitrary matrix A is defined by*

$$\mathrm{Cond}(u) = \|S(\lambda)(I-P)\|_2. \tag{3.43}$$

in which $S(\lambda)$ is the reduced resolvent of A at λ.

In the case where the matrix A is Hermitian it is easy to verify that the condition number simplifies to the following

$$\mathrm{Cond}(u) = \frac{1}{dist[\lambda, \sigma(A) - \{\lambda\}]}\ . \tag{3.44}$$

In the general non-Hermitian case, it is difficult to assess the size of Cond(u).

To better understand the nature of the operator $S(\lambda)(I-P)$, consider its spectral expansion in the particular case where A is diagonalizable and the eigenvalue λ_i of interest is simple.

$$S(\lambda_i)(I-P_i)=\sum_{\substack{j=1\\ j\neq i}}^{p}\frac{1}{\lambda_j-\lambda_i}P_j$$

Since we can write each projector as a sum of outer product matrices $P_j=\sum_{k=1}^{\mu_i}u_k w_k^H$ where the left and right eigenvectors u_k and w_k are normalized such that $(u_j,w_j)=1$, the expression (2.9) can be rewritten as

$$u'(0)=\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{1}{\lambda_j-\lambda_i}u_j w_j^H E u_i \;=\; \sum_{\substack{j=1\\ j\neq i}}^{n}\frac{w_j^H E u_i}{\lambda_j-\lambda_i}u_j$$

which is the standard expression developed in Wilkinson's book [183].

What the above expression reveals is that when eigenvalues get close to one another then the eigenvectors are not too well defined. This is predictable since a multiple eigenvalue has typically several independent eigenvectors associated with it, and we can rotate the eigenvector arbitrarily in the eigenspace while keeping it an eigenvector of A. As an eigenvalue gets close to being multiple, the condition number for its associated eigenvector deteriorates. In fact one question that follows naturally is whether or not one can define the notion of condition number for eigenvectors associated with multiple eigenvalues. The above observation suggests that a more realistic alternative is to try to analyze the sensitivity of the invariant subspace. This is taken up in the next section.

Example 3.8 Consider the matrix seen in example 3.6

$$A=\begin{pmatrix} -149 & -50 & -154\\ 537 & 180 & 546\\ -27 & -9 & -25 \end{pmatrix}.$$

The matrix is diagonalizable since it has three distinct eigenvalues and

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

One way to compute the reduced resolvent associated with $\lambda_1 = 1$ is to replace in the above equality the diagonal matrix D by the 'inverse' of $D - \lambda_1 I$ obtained by inverting the nonzero entries $(2,2)$ and $(3,3)$ and placing a zero in entry $(1,1)$, i.e.,

$$S(\lambda_1) = X \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix} X^{-1} = \begin{pmatrix} -118.5 & -39.5 & -122.5 \\ 316.5 & 105.5 & 325.5 \\ 13.5 & 4.5 & 14.5 \end{pmatrix}$$

We find that the 2-norm of $\|S(\lambda_1)\|_2$ is $\|S(\lambda_1)\|_2 = 498.27$. Thus, a perturbation of order 0.01 may cause changes of magnitude up to 4.98 on the eigenvector. This turns out to be a pessimistic overestimate. If we perturb a_{11} from -149.00 to -149.01 the eigenvector u_1 associated with λ_1 is perturbed from $u_1 = (-1/3, 1, 0)^T$ to $\tilde{u}_1 = (-0.3170, 1, -0.0174)^T$. A clue as to why we have a poor estimate is provided by looking at the norms of X and X^{-1}.

$$\|X\|_2 = 1.709 \quad \text{and} \quad \|X^{-1}\|_2 = 754.100 ,$$

which reveals that the eigenvectors are poorly conditioned.

3.3. Conditioning of Invariant Subspaces

Often one is interested in the invariant subspace rather than the individual eigenvectors associated with a given eigenvalue. In these situations the condition number for eigenvectors as defined before is not sufficient. We would like to have an idea on how the whole subspace behaves under a given perturbation.

We start with the simple case where the multiplicity of the eigenvalue under consideration is one, and we define some notation. Referring to (3.40), let $Q(t)$ be the orthogonal projector onto the invariant subspace associated with the simple eigenvalue

$\lambda(t)$ and $Q(0) \equiv Q$ be the orthogonal projector onto the invariant subspace of A associated with λ. The orthogonal projector Q onto the invariant subspace associated with λ has different properties from those of the spectral projector. For example A and Q do not commute. All we can say is that

$$AQ = QAQ \quad \text{or} \quad (I-Q)AQ = 0 \ ,$$

leading to

$$\begin{aligned} (I-Q)A &= (I-Q)A(I-Q) \\ (I-Q)(A-\lambda I) &= (I-Q)(A-\lambda I)(I-Q) \end{aligned} \tag{3.45}$$

Note that the linear operator $(A-\lambda I)$ when restricted to the range of $I-Q$ is invertible. This is because if $(A-\lambda I)x = 0$ then x belongs to $\text{Ran}(Q)$ whose intersection with $\text{Ran}(I-Q)$ is reduced to $\{0\}$. We denote by $S^+(\lambda)$ the inverse of $(A-\lambda I)$ restricted to $\text{Ran}(I-Q)$. Note that although both $S(\lambda)$ and $S^+(\lambda)$ are inverses of $(A-\lambda I)$ restricted to complements of $\text{Ker}(A-\lambda I)$, these inverses are quite different.

Starting from (3.40), we subtract λu from each side to get,

$$(A-\lambda I)u(t) = -tEu(t) + (\lambda(t)-\lambda)u(t)$$

Now multiply both sides by the orthogonal projector $I-Q$,

$$(I-Q)(A-\lambda I)u(t) = -t(I-Q)Eu(t) + (\lambda(t)-\lambda)(I-Q)u(t)$$

to obtain from (3.45),

$$\begin{aligned} &[(I-Q)(A-\lambda I)(I-Q)](I-Q)u(t) \\ &\qquad = -t(I-Q)Eu(t) + (\lambda(t)-\lambda)(I-Q)u(t). \end{aligned}$$

Therefore,

$$(I-Q)u(t) = S^+(\lambda)\left[-t(I-Q)Eu(t) + (\lambda(t)-\lambda)(I-Q)u(t)\right].$$