The Rayleigh Quotient Iteration

For Non-Normal Matrices

by

Nai-Fu Chen

Memorandum No. ERL-M-548

2 October 1975

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

THE RAYLEIGH QUOTIENT ITERATION FOR NON-NORMAL MATRICES[†]

Nai-Fu Chen

Department of Mathematics
University of California at Berkeley

Ph.D. Dissertation

December 1975

B.N. Parlett
Professor B.N. Parlett
Thesis Chairman

Abstract

The Rayleigh Quotient Iteration (RQI) is a method for computing eigenvectors and eigenvalues of a square matrix.

The behaviour, both local and global, of RQI with symmetric and normal matrices is almost completely understood. The vector sequence converges for almost all starting vectors.

In this paper, we investigate the global properties of RQI on non-normal matrices. Results on nearly normal matrices with real eigenvalues are obtained, and at the other extreme, results on completely degenerate matrices are also obtained. In particular, the question of global convergence of the vector iteration on a general matrix is reduced to the convergence of the scalar sequence of the Rayleigh quotients. In practice, the vector iteration always converges.

The main difficulty in the degenerate case is that the iteration function is discontinuous near the eigenvector. An example is used to display the sectorial behaviour of the iteration. Further we construct a sequence of numbers that converges to the eigenvalue. Yet if the numbers are used as shifts with inverse iteration, the vector sequence fails to converge.

[†] Research partially sponsored by Office of Naval Research Contract N00014-69-A-0200-1017.

Acknowledgment

The author would like to give his greatest thanks to his advisor, Professor Beresford Parlett, who guided the research with unfailing help and encouragement. He is grateful to the other committee members, to Professor William Kahan for helpful conversations.

The author gratefully acknowledges the support by the Office of Naval Research under Contract Number N00014-69-A-0200-1017.

The author thanks Ruth Suzuki for her immaculate typing and drawing of diagrams.

Table of Contents

Page

Chapter Zero - Introduction 1

0.1 History 1
0.2 RQI and SQR 2
0.3 A Brief Outline of Results 3

Chapter One - Definitions and Known Results 5

1.1 Notation 5
1.2 Definition and Basic Properties of Rayleigh Quotients 5
1.3 The Rayleigh Quotient Iteration and Its Invariant Properties 7
1.4 Local Results 8
1.5 Fixed Points of RQI 9
1.6 Global Results 10

Chapter Two - Rayleigh Quotient Iteration for Semi-Simple Matrices 14

2.1 Synopsis 14
2.2 Anatomy of the Rayleigh Quotient 16
2.3 The 2 × 2 Case 17
2.4 The Hermitian Case 25
2.5 The Nearly Normal Case 35
2.6 Bisectors of Eigenvectors as Limit Vectors of RQI . . 39
2.7 Characterization of Limit Vectors When ρ_k Converges . . 41
2.8 Instability of Case (ii) 43

Chapter Three - The Completely Degenerate Case 45

3.1 An Overview 45
3.2 A 3 × 3 Example - Sectorial Behaviour Near the Eigenvector 49
3.3 Behaviour of ρ_k as $k \to \infty$ 53
3.4 Behaviour of $V^{(k)}$ When ρ_k Converges 59
3.5 A Shift Sequence Which Prevents Convergence 66

Chapter Four - The General Case 74

Bibliography 77

CHAPTER ZERO

Introduction

§0.1 History

The idea of Rayleigh Quotient Iteration (RQI) originated in the nineteenth century. Its earliest function was to improve an approximation to a mode shape in the theory of sound.

With the advent of high speed digital computers in the 1950's the RQI was turned into a way of computing eigenvectors and eigenvalues of Hermitian matrices by successively improving an arbitrary initial starting vector.

In 1958/59 Ostrowski [4] published a series of six articles giving detailed analyses of the local asymptotic behaviour of the RQI and some variants of it. He discussed both the symmetric and the nonsymmetric cases, but not much was said about the global convergence because of the complexity of the behaviour at early stages of the iterative process. In 1968, Parlett and Kahan [5] proved the global convergence, for almost all starting vectors, in the symmetric case. In 1974, Parlett [7] proved the global convergence, for almost all starting vectors, in the normal case.

In this paper, we continue the investigation. The Rayleigh Quotient Iteration is of interest to us for the following reasons:

1) Because of its excellent local convergence rate, which we will discuss in later chapters. This is a very fast way to compute a few eigenvectors, especially if one has a fair approximation to the eigenvector or the eigenvalue.

2) There is an intimate relationship between RQI and the Shifted QR (SQR) method, and thus, understanding RQI would hold the key to unlock the mystery of Shifted QR which is so successful in practice. This will be discussed in more detail in our next section.

3) A generalized form of RQI was found to be a powerful algorithm for finding zeros of a polynomial. The iteration is applied to the Frobenius matrix associated with the polynomial (see Wilkinson [9], p. 349).

§0.2 RQI and SQR

The shifted QR algorithm is currently the champion for computing eigenvalues and the process is stable, but the shifted QR algorithm is so complicated to analyze that any direct approach to its global behavior seems intimidating.

Given a matrix C, the Shifted QR is defined as follows: for $k = 1,2,3,\dots,$ let $f_k(C^{(k)}) = Q^{(k)}R^{(k)}$, then set $C^{(k+1)} = Q^{(k)*}C^{(k)}Q^{(k)}$ where $C^{(1)} = C^*$, $Q^{(k)}$ is an unitary matrix, $R^{(k)}$ is an upper triangular matrix, $f_k(t)$ can be any sequence of polynomials of fixed degree δ. We may define:

$$f_k(t) = \det(tI - E^*C^{(k)}E)$$

where E is the last δ columns of the $n \times n$ identity matrix I. A practical SQR method takes $E = e_n = (0,\dots,0,1)^T$, i.e., the shift is the (n,n) element of $C^{(k)}$.

The relationship between RQI and SQR was discussed by Parlett and Kahan [5], and Wilkinson [9]. It is briefly the following:

If $V^{(k)}$ is the vector at the k^{th} step of RQI with initial vector

e_n, then $V^{(k)}$ = the last column of the matrix $\prod_{i=1}^{k} Q^{(i)}$. Hence if $V^{(k)} \to x$, an eigenvector of C, then

$$C^{(k)} = Q^{(k)*} \cdots Q^{(1)*} C^* Q^{(1)} \cdots Q^{(k)} \to \begin{pmatrix} x & \cdots & x & x \\ \vdots & & \vdots & \vdots \\ x & \cdots & & x \\ 0 & \cdots & 0 & \lambda \end{pmatrix}$$

where $Ax = \lambda x$.

Thus, convergence of SQR can be deduced from the convergence of RQI.

§0.3 A Brief Outline of Results

Here we investigate what RQI brings us when the matrix C and an arbitrary starting vector V is given.

In Chapter 1, we define RQI and state known results.

In Chapter 2, we investigate the semi-simple case. We show that a bisector of a pair of eigenvectors can always be a limit vector of the iteration. We show the almost-always convergence properties of RQI on Hermitian matrices through a different method than that employed by Parlett and Kahan [5], and thus the results can be extended into the non-normal cases for well-conditioned matrices. We also have a complete characterization of limit vectors of RQI on semi-simple matrices when a scalar sequence, called the Rayleigh Quotients (ρ_k) converges.

In Chapter 3, we expose new difficulties that we encounter in the completely degenerate case. There will be a detailed analysis of a 3×3 matrix with sectorial behavior around the eigenvector. Then we shall prove global convergence under a certain weak condition when the Rayleigh Quotients converge, and we construct a sequence of numbers, which, when used as shifts, would force inverse iteration not to converge even though the sequence of numbers converges to the eigenvalue.

In Chapter 4, we summarize the picture into a theorem which makes the only unsettled question about the global behavior of RQI for a general matrix the convergence of ρ_k.

CHAPTER ONE

Definition and Known Results

§1.1 Notation

First we shall explain our notation. Matrices will be represented by capital Roman letters, column vectors by small Roman letters except for $i, j, k, \ell, m, n, p, q$ which are reserved for indices. Greek letters represent scalars. The conjugate transpose of a vector u is denoted by u^*, and unless otherwise specified, $\|u\| = \sqrt{u^*u}$, the Euclidean norm. $I = (e_1, e_2, \ldots, e_n)$ is the identity matrix, and the matrix $C - \rho I$ is often abbreviated as $C - \rho$.

§1.2 Definition and Basic Properties of Rayleigh Quotient

The Rayleigh Quotient ρ is a function defined by

$$\rho\colon \mathbb{C}^n - \{0\} \to \mathbb{C} \ ,$$
$$u \mapsto u^*Cu/u^*u \ , \qquad u \neq 0 \ ,$$

where C is a matrix whose eigenvalues and eigenvectors we seek. So ρ assigns to each non-zero complex vector a scalar value. Also, if it is necessary to emphasize the role of C, we write $\rho(u) \equiv \rho(u,C)$. The following are some basic facts:

Homogeneity: $\rho(\alpha u, \beta C) = \beta\rho(u,C)$; $\alpha \neq 0$.

Translation Invariance: $\rho(u, C - \alpha I) = \rho(u,C) - \alpha$.

Continuity: The function ρ is a continuous function.

Boundedness: $\{\rho(u),\ u \neq 0\}$ is a region (the field of values) in the complex plane for a given matrix C. By homogeneity $\{\rho(u),\ u \neq 0\} = \{\rho(u),\ u^*u = \|u\| = 1\}$, i.e., we only have to consider unit vectors. Since the unit sphere is compact and ρ is continuous, therefore, $\{\rho(u),\ u \neq 0\}$ is compact. $C = R \times R$ is a metric space implies $\{\rho(u),\ u \neq 0\}$ is closed and bounded.

Stationary Values: We say ρ is stationary at n if $\lim[\rho(u+tv) - \rho(u)]/t = 0$ as $t \to 0$ through real values for all v. A straightforward calculation (see Parlett [7]) shows that ρ is stationary at u if and only if $(C-\rho(u))u = 0$ and $u^*(C-\rho(u)) = 0^*$, i.e., u must be an eigenvector of C and C^*. Note that if C is normal $(CC^* = C^*C)$, the eigenvectors of C are the stationary points of ρ.

Minimal Residual: *Given* $u \neq 0$, $\|(C-\mu)u\|$ *is minimal if and only if* $\mu = \rho(u)$.

Proof:
$$\begin{aligned}\|(C-\mu)u\|^2 &= u^*u|\mu|^2 - \bar{\mu}u^*Cu - \mu u^*C^*u + u^*C^*u \\ &= u^*u\{(\bar{\mu}-\overline{\rho(u)})(\mu-\rho(u)) - |\rho(u)|^2 + u^*C^*Cu/u^*u\} \\ &= \|Cu\|^2 - |\rho(u)|^2\|u\|^2 + (\bar{\mu}-\overline{\rho(u)})(\mu-\rho(u))u^*u\ .\end{aligned}$$

Therefore, $\|(C-\mu)u\|^2 \geq \|Cu\|^2 - |\rho(u)|^2\|u\|^2$, with equality if and only if $\mu = \rho(u)$. □

Corollary: $\|u^*(C-\mu)\|^2 \geq \|u^*C\|^2 - |\rho(u)|^2\|u^*\|^2$ *and equality holds if and only if* $\mu = \rho(u)$.

Corollary. u *is orthogonal to* $(C-\rho(u))u$ *in Euclidean space*.

The above fact is equivalent to: Let f be an arbitrary polynomial and compute $\|f(C)u\|$ over all monic polynomials of degree one. Then the polynomial $t - \rho(u)$ is minimal.

§1.3 The Rayleigh Quotient Iteration and Its Invariant Properties

The Rayleigh Quotient Iteration (RQI) is the following scheme: For an arbitrary starting unit vector $V^{(0)}$, and for $k = 0,1,2,\ldots$

(i) Compute $\rho_k = \rho(V^{(k)})$.

(ii) If $C - \rho_k$ is singular, solve $(C-\rho_k)V^{(k+1)} = 0$ for $\underline{V}^{(k+1)} \neq 0$ and stop. Otherwise

(iii) Solve $(C-\rho_k)W^{(k+1)} = V^{(k)}$.

(iv) Normalize $V^{(k+1)} = W^{(k+1)}/\|W^{(k+1)}\|$.

The sequence $\{\rho_k, V^{(k)}\}$ is called the Rayleigh sequence generated by $V^{(0)}$ and C.

If $\underline{V}^{(k)}$ converges to x, an eigenvector of C, then, by continuity of ρ, $\rho_k = \rho(V^{(k)})$ converges to λ, the associated eigenvalue of x. Hence RQI can be regarded as a method to find eigenvectors or eigenvalues or both.

Let $\{\rho_k, V^{(k)}\}$ be the sequence generated by RQI from C and $V^{(0)}$. The following are invariant properties of the iteration:

(i) Scaling: The matrix αC, $\alpha \neq 0$, produces the same sequence as C.

(ii) Translation: The matrix $C-\alpha$ produces the sequence $\{\rho_k-\alpha, V^{(k)}\}$.

(iii) Unitary Similarity: The matrix QCQ^*, Q unitary, produces the sequence $\{\rho_k, QV^{(k)}\}$ if you start with $QV^{(0)}$. This property says that RQI is coordinate free, in contrast with QR.

With these invariant properties in mind, we shall normalize our matrix in the most convenient form in subsequent chapters without loss of generality.

§1.4 Local Results

In 1958/59 Ostrowski [4] published six articles giving a rigorous and detailed analysis of the local asymptotic behaviour of RQI and some variant of it. Ostrowski showed that for a semi-simple, i.e., nondefective non-normal, matrix C, the local convergence rate of RQI is quadratic if C has real eigenvalues.

Theorem (Ostrowski). *If the sequence* ρ_k *tends to* λ *without any of the* ρ_k *becoming equal to one of the other eigenvalues of* C, *if all eigenvalues of* C *are real, and* $\underline{v}^{(k)} \to x$, *the associated eigenvector, then either for a constant* K, $\rho_{k+1} - \lambda = O(K^k \prod_{j=0}^{k} (\rho_k-\lambda)^2)$ *as* $k \to \infty$ *or* $\frac{\rho_{k+1} - \lambda}{(\rho_k-\lambda)^2} \to g \neq 0$ *as* $k \to \infty$. *If some of the eigenvalues of* C *are complex, then*

$$\rho_{k+1} - \lambda = o(\rho_k-\lambda) \quad \underline{\text{as}} \quad k \to \infty ,$$

and for any fixed integer p

$$\rho_{k+1} - \lambda = O((\rho_k-\lambda)(\rho_{k-1}-\lambda)\cdots(\rho_{k-p}-\lambda)) \quad \underline{\text{as}} \quad k \to \infty .$$

In 1974, Parlett [7] showed that for a normal matrix C, the local convergence rate for RQI is cubic (the equivalent result for SQR was known to Wilkenson and Buurma [1]), and this excellent local convergence rate makes the RQI all the more interesting.

Theorem (Parlett). *If* C *is normal and* $V^{(k)} \to x$, *an eigenvector for* λ, *as* $k \to \infty$, *then*

either $$\|V^{(k+1)}-X\|/\|V^{(k)}-X\|^3 \to 1$$

or $0 < \|V^{(k+1)}-X\|/\|V^{(k)}-X\|^3 < 1$ *for all sufficiently large* k.

For defective matrices, the local convergence rate is *at best* linear. In 1970, Kiho Lee Kim [2] computed the local convergence rate for completely degenerate matrices of sizes ranging from three to twenty. But in his analysis, he assumed the iteration function to be differentiable whereas RQI is not continuous in any neighborhood containing the eigenvector when C is completely degenerate. This will be discussed in greater detail in Chapter 3.

§1.5 Fixed Points of RQI

It is easy to see, by step (ii) that the only fixed points of RQI are the eigenvector because: If u is a fixed point, consider $C-\rho(u)$:

Case 1: If $C-\rho(u)$ is singular, by step (ii), we shall get an eigenvector.

Case 2: If $C-\rho(u)$ is nonsingular, then $(C-\rho(u))u = u$ or $(C-\rho(u)-1)u = 0$ implies u is an eigenvector.

This result depends strongly on the definition of RQI as stated in §1.3.

For semi-simple matrices, Ostrowski [4] described a finite neighborhood around each eigenvector. When RQI is started in this region, convergence will occur to the associated eigenvector.

§1.6 Global Results

In 1966, Kahan [5] had a proof that for Hermitian matrices, the RQI converges for almost all starting vectors. In 1974 Parlett [7] proved global convergence of RQI for normal matrices except in an unstable special case:

Theorem. Let the RQI be applied to a normal matrix C with starting vector $v^{(0)}$. As $k \to \infty$

(i) $\rho_k = \rho(v^{(k)})$ converges, and either

(ii) $(\rho_k, v^{(k)})$ converges to an eigenpair (λ, x) or

(iii) ρ_k converges to a point equidistant from m (≥ 2) eigenvalues of C, and the sequence $\{v^{(k)}\}$ cannot converge. It may or may not have a limit cycle.

The main tool in proving the above theorem is the observation that the norm of the residual (to be defined) is monotone decreasing:

Theorem. Let $r^{(k)} = (C-\rho_k)v^{(k)}$ be the residual at the k^{th} step of RQI. If C is normal, then the sequence $\{\|r^{(k)}\| : k = 0,1,\ldots\}$ is monotone decreasing for all starting vectors $v^{(0)}$.

Proof.
$$
\begin{aligned}
\|r^{(k+1)}\| &= \|(C-\rho_{k+1})v^{(k+1)}\| \\
&\leq \|(C-\rho_k)v^{(k+1)}\| , \quad \text{by minimal residual property} \\
&= |v^{(k)*}(C-\rho_k)v^{(k+1)}| , \quad \text{since } (C-\rho_k)v^{(k+1)} \text{ is} \\
&\qquad \text{parallel to } v^{(k)} \\
&\leq \|v^{(k)*}(C-\rho_k)\|\|v^{(k+1)}\| , \quad \text{Cauchy-Schwarz} \\
&= \|v_k^*(C-\rho_k)\| \\
&= \|(C-\rho_k)v^{(k)}\| , \quad \text{since } C \text{ is normal} \\
&= \|r^{(k)}\| .
\end{aligned}
$$

Equality can occur only if $\rho_{k+1} = \rho_k$ and $V^{(k+1)*}$ is parallel to $V^{(k)*}(C-\rho_k)$. From monotonicity of norm of residuals, it can be shown that

$$(1.6.1) \qquad |\rho_k-\rho_{k+1}| \to 0 \text{ as } k \to \infty .$$

Then global convergence of ρ_k for normal matrices can then be proved. Parlett [7] in his paper remarked this fact without going into details. For completeness, we include the proof of the global convergence of ρ_k here:

Proof. I. If $\|r^{(k)}\| \to 0$, we know $V^{(k)} \to x$, an eigenvector, hence $\rho_k \to \rho(x) = \lambda$. Therefore ρ_k converges.

II. If $\|r^{(k)}\| \to \tau > 0$, the set $F = \{\rho(u) | \; u \in S^{n-1}\}$ is compact, so let $\rho^{(1)}$ be a limit point of $\{\rho_k\}$, and let the infinite set $K \subseteq \{0,1,2,\ldots\}$ be such that $\lim \rho_k = \rho^{(1)}$ for $k \in K$. Consider $\{V^{(k)} | \; k \in K\}$. It has limit points because S^{n-1} is compact. Let $\Gamma \subseteq K$ be such that $\lim V_\gamma = V^{(1)}$ for $\gamma \in \Gamma$. Then $\rho(V^{(1)}) = \rho^{(1)}$.

From Parlett's theorem, it was proved that

(i) $\|(C-\rho^{(1)})V^{(1)}\| = \tau$,

(ii) $(C-\rho^{(1)})*(C-\rho^{(1)})V^{(1)} = \tau^2 V^{(1)}$,

(iii) $\rho^{(1)} = \lim_{\gamma\in\Gamma} \rho(V_\gamma)$.

From the above three relations, we can derive that $\rho^{(1)}$ is

(a) equidistant from m (≥ 2) distinct eigenvalues of C,

(b) a weighted mean of them.

We can see immediately that $\rho^{(1)}$ can assume at most (2^n-n) values. We shall see later that only a very few of the (2^n-n) values can actually be candidates for a limit point of ρ_k. We have established that there are only finitely many limit points of ρ_k. Let $\rho^{(1)},\dots,\rho^{(q)}$ be the distinct limit points. We proceed to show that if $q \geq 2$ there is a contradiction.

Let $d = \min\{|\rho^{(i)}-\rho^{(j)}|, i \neq j, 1 \leq i,j \leq q\} > 0$. Let $B_i = \{\xi |\ |\xi-\rho^{(i)}| < d/4, \xi \in \mathbb{C}\}$, i.e., open balls in the complex plane. Then $F \setminus \bigcup_{i=1}^{q} B_i$ is still compact, hence there can only be finitely many ρ_k's belonging to $F \setminus \bigcup_{i=1}^{q} B_i$ (otherwise we have another limit point of ρ_k). Let this finite number of ρ_k's be M.

From (1.6.1) we have $|\rho_k-\rho_{k+1}| \to 0$, so there exists N such that $|\rho_k-\rho_{k+1}| < d/2M$ for $k \geq N$.

If $q \geq 2$ pick k_1 and k_2 such that $k_2 > k_1 > N$ and $\rho_{k_1} \in B_1$, $\rho_{k_2} \in B_2$. Then one of the following:

$$\begin{gathered} |\rho_{k_1}-\rho_{k_1+1}| \\ |\rho_{k_1+1}-\rho_{k_1+2}| \\ \vdots \\ |\rho_{k_2-1}-\rho_{k_2}| \end{gathered}$$

must be greater than $d/2M$ for the $\{\rho_k\}$ to travel from B_1 to B_2, which is a contradiction to $|\rho_k-\rho_{k+1}| < d/2M$ for $k \geq N$.

So $q = 1$, in other words, ρ_k converges. □

Remarks. (1) The values that ρ_k can converge to is limited by properties (a) and (b). (a) says that a circle must be able to be drawn through the eigenvalues and ρ is the center. So, not any combinations of eigenvalues can produce a possible limit for ρ_k. (b) says that ρ is a weighted mean of those eigenvalues, so, not any combination of co-cyclic eigenvalues can produce a limit point.

(2) Though ρ_k always converges, $V^{(k)}$ does not and may have infinitely many limit vectors. For example, a 3×3 matrix $C = \text{diag}(1,e^{2i},e^{-2i})$. It is easy to see that there exist α_1, α_2, α_3 such that $\sum|\alpha_i|^2\lambda_i = 0$. If we let $V^{(0)} = \alpha_1 e_1 + \alpha_2 e_2 + \alpha_3 e_3$, then for $k = 0,1,2,\ldots$, $V^{(k)} = \alpha_1 e_1 + \alpha_2 e^{-i2k} e_2 + \alpha_3 e^{i2k} e_3$. The sequence $\{V^{(k)}\}$ is infinite because if $V^{(i)} = V^{(j)}$ with $i \neq j$ would imply $2i \equiv 2j \pmod{2\pi}$ which is impossible. It is also obvious that each $V^{(k)}$ is a limit vector of the set $\{V^{(k)}:\ k = 0,1,2,\ldots\}$.

CHAPTER TWO

Rayleigh Quotient Iteration for Semi-Simple Matrices

§2.1 Synopsis

Here we turn our attention to non-normal matrices. Remember the main tool in the proof of the convergence of RQI in the normal case is the monotonicity of the residual. That no longer holds for non-normal matrices. For example, the matrix

$$\begin{pmatrix} 3 & 2 & 1 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

with an initial vector of $(0,.75,.6614)^T$ would produce a sequence of residuals which is not monotonic (see Table 2.1). And the closer the initial vector is to the boundary that separates regions of convergence to different eigenvectors, the more capricious the behaviour of the residuals is.

In the normal case, the residuals of different limit vectors are the same. For example, for the matrix $A = \text{diag}(3,2,1)$, $(e_1+e_2)/\sqrt{2}$ and $(e_1-e_2)/\sqrt{2}$ are limit vectors belonging to the sequence of vectors generated by $v^{(0)} = (e_1+e_2)/\sqrt{2}$. Both of the residuals equal 0.5, but residuals of different limit vectors may be different in non-normal cases. This will be shown in Section 3 of this chapter.

The residuals have been measured in norms other than the Euclidean norm. First we define X and Y^* as follows: Let X be an $n \times n$ matrix whose columns are unit column-eigenvectors of the matrix C, and Y^* is an $n \times n$ matrix whose rows are unit row-eigenvectors of matrix C, and Y^* is obtained by normalizing rows of X^{-1} to have

Table 2.1

k number of iterations	ρ Rayleigh Quotient	$(V^{(k)})^T$	$\|r^{(k)}\|$ residual
0	2.555	(0,.75,.66)	2.560
1	2.543	(-.97,.20,.03)	.06928
2	2.790	(-.31,-.94,-.040)	2.112
3	2.807	(-.99,-.09,-.002)	.07129
4	3.058	(-1.0,-.002,-.000)	.0308
5	3.003	(-1.0,-0.0,-0.0)	.0015
6	3.000	(-1.0,-0.0,-0.0)	.48E-5
7	3,000	(-1.0,-0.0,-0.0)	.46E-10
8	3.000	(-1.0,-0.0,-0.0)	.98E-20
0	2.540	(0.,.93,.368)	2.302
1	2.523	(-.97,.21,.02)	.081
2	2.503	(.20,.988,.27)	2.132
3	2.500	(-.97,.23,.002)	.114
4	2.411	(.19,.98,.003)	2.126
5	2.401	(-.96,.27,0.0)	.113
6	1.488	(-.70,.71,0.0)	.525
7	1.846	(-.85,.51,0.0)	.093
8	1.981	(-.89,.46,0.0)	.010
9	2.000	(-.89,.44,0.0)	.002
10	2.000	(-.89,.45,0.0)	.58E-7
11	2,000	(-.89,.45,0.0)	.13E-13

unit length. We have tested the following norms: $\|X^{-1}(C-\rho)v\|$, $\|X^{-1}(C-\rho)v\|/\|X^{-1}v\|$, $\|Y^*(C-\rho)v\|/\|Y^*v\|$, and some of these were designed to force residuals of different limit vectors to be the same, but the monotonicity has not been recaptured.

§2.2 Anatomy of the Rayleigh Quotient

To pursue the behaviour of Rayleigh Quotient Iterations, we first analyze the Rayleigh Quotient. From the definition, the Rayleigh Quotient of a non-zero vector v is $\rho = v^*Cv/v^*v$. If C is normal, then X, as defined in the last section, can be taken as unitary. Then

$$\begin{aligned}\rho &= \frac{(Xx)^*C(Xx)}{(Xx)^*(Xx)} \quad \text{when} \quad x = X^{-1}v \\ &= \frac{x^*X^*CXx}{x^*X^*Xx} \\ &= \frac{x^*Dx}{x^*x} \qquad \text{where} \quad D = \operatorname{diag}(\lambda_i) = X^*CX \\ &\qquad\qquad \text{by definition of} \quad X.\end{aligned}$$

So if $x = (\alpha_1,\alpha_2,\ldots,\alpha_n)^T$, i.e., $v = \sum\alpha_i x_i$ where $Cx_i = \lambda_i x_i$, then

$$\rho = \frac{\sum_{i=1}^{n} |\alpha_i|^2\lambda_i}{\sum_{i=1}^{n} |\alpha_i|^2} .$$

Therefore, ρ can be regarded as a weighted mean of the eigenvalues of C. The weights are proportional to the square of the coefficients when v is expressed as a linear combination of eigenvectors.

In the non-normal case, the picture is very different. In this chapter, we shall concern ourselves only with non-defective matrices. We now remind the reader that we are introducing some facts and standard notation about non-normal matrices. Let X and Y^* be defined as in the last section. Then $CX = X\Lambda$, $Y^*C = \Lambda Y$, where $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$, and x_i, y_i, $i = 1,\ldots,n$, are unit column and row vectors of X and Y^* respectively. Then $\gamma_i = y_i^* x_i$ $(0 < \gamma_i \leq 1)$ are reciprocals of the condition numbers of the eigenvalues (see Wilkinson [8]).

Let $v = \sum_{i=1}^{n} \alpha_i x_i$, $v^* = \sum_{i=1}^{\beta} \beta_i y_i^*$. Then

$$
\begin{aligned}
(2.2.1) \qquad \rho &= \frac{v^*Cv}{v^*v} \\
&= \frac{y^*Y^*CXx}{y^*Y^*Xx} \qquad y^* = (\beta_1,\ldots,\beta_n), \quad x = (\alpha_1,\ldots,\alpha_n)^T \\
&= \frac{y^*\Gamma\Lambda x}{y^*\Gamma x} \qquad \text{where} \quad \Lambda = \mathrm{diag}(\lambda_i), \quad \Gamma = \mathrm{diag}(\gamma_i) \\
&= \frac{\sum_{i=1}^{n} \alpha_i \beta_i \gamma_i \lambda_i}{\sum_{i=1}^{n} \alpha_i \beta_i \gamma_i} .
\end{aligned}
$$

Although the Rayleigh Quotient can still be considered as some sort of mean of the eigenvalues, it is no longer a barycentric mean (convex combination) of λ_i's because even in the real case $\gamma_i \alpha_i \beta_i$ may be negative. When some $\alpha_i \beta_i < 0$, it is possible then for some $|\alpha_i \beta_i \gamma_i / \sum(\alpha_i \beta_i \gamma_i)| > 1$ and hence ρ is no longer confined to the convex hull of eigenvalues in the complex plane. (Recall that a matrix is normal if (and only if) its numerical range $\{\rho(v)\colon v \neq 0\}$ is the convex hull of its eigenvalues [Hausdorff] (see Mareus and Minc [3]).

§2.3 The 2 × 2 Case

Having outlined the general picture, we shall now analyze in detail the 2 × 2 case. This will on one hand show the role that the Rayleigh

Quotient plays in the convergence of the iteration, and on the other hand, supply a result needed later in Section 6 which deals with a sufficient condition for a vector to be a limit vector.

By the invariance properties of RQI, it is sufficient to consider

$$C = \begin{pmatrix} 1 & \kappa \\ 0 & 0 \end{pmatrix}, \quad \kappa \geq 0 .$$

And since v and $-v$ will give the same sequence of vectors except for sign, it is sufficient to consider vectors on a unit hemisphere, and in the present case we can have our column and row eigenvectors on a half circle of the unit circle (even in the complex case, with the above normalization, column eigenvectors can be expressed as a real linear combination of y_1 and y_2 (see Section 2 for definition)). We have the following figure:

$$y_1^*C = y_1^* , \quad y_2^*C = 0$$

$$Cx_1 = x_1, \quad Cx_2 = 0$$

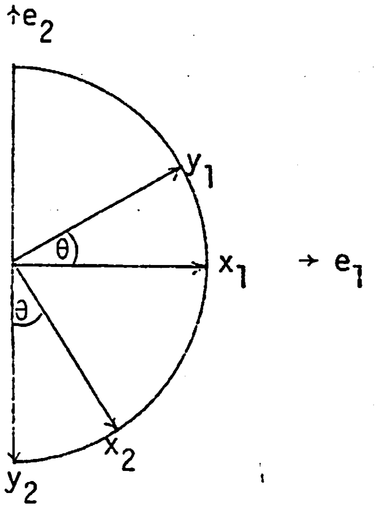


Figure 2.3.1

Note that we choose directions so that the angle between y_i and x_i is acute. If we want to express x_1, x_2, y_1, y_2 in the orthonormal basis $\{e_1,e_2\}$, then $x_1 = (1,0)^T$, $x_2 = (\kappa,-1)^T/\|(\kappa,-1)^T\|$,

$y_1 = (1,\kappa)/\|(1,\kappa)\|, \quad y_2 = (0,-1)^T.$

<u>Lemma 1</u>. $\gamma_1 = \gamma_2$

<u>Proof</u>.
$$y_1^* x_1 = 1/\|(1,\kappa)\| = 1/\sqrt{1+\kappa^2}$$
$$y_2^* x_2 = 1/\|(\kappa,-1)\| = 1/\sqrt{1+\kappa^2}$$

Therefore $\gamma_1 = y_1^* x_1 = y_2^* x_2 = \gamma_2$.

$$\theta_1 = \text{angle between } x_1 \text{ and } y_1 = \cos^{-1}\gamma_1 = \cos^{-1}\gamma_2$$
$$= \text{angle between } x_2 \text{ and } y_2 = \theta_2 \ . \qquad \square$$

<u>Lemma 2</u>. <u>Let</u> $v = \alpha_1 x_1 + \alpha_2 x_2 = \beta_1 y_1 + \beta_2 y_2$. <u>Then</u>

$$|\alpha_1| > |\alpha_2| \ \underline{\text{implies}} \ |\beta_1| > |\beta_2| \ .$$

<u>Proof</u>. In order to show the main idea without involving ourselves in unnecessary detail, we shall proceed on the assumption that α_1, α_2 are real. The proof of the complex case is left to the reader.

Let $\tau = \sqrt{1+\kappa^2}$. Then

$$x_1 = (1,0)^T = \tau(1,\kappa)^T/\tau + \kappa(0,-1)^T$$
$$= \tau y_1 + \kappa y_2$$
$$x_2 = (\kappa,-1)/\tau = \kappa(1,\kappa)/\tau + \tau(0,-1)$$
$$= \kappa y_1 + \tau y_2 \ .$$

So if $v = \alpha_1 x_1 + \alpha_2 x_2 = \beta_1 y_1 + \beta_2 y_2$, then

$$\beta_1 = \tau\alpha_1 + \kappa\alpha_2 \qquad (2.3.1)$$
$$\beta_2 = \kappa\alpha_1 + \tau\alpha_2 \ .$$

Recall that $\tau > \kappa$ and we assume $|\alpha_1| > |\alpha_2|$. So among the four products that appear on the right hand sides, $|\tau\alpha_1| > \{|\kappa\alpha_1|,|\tau\alpha_2|\} > |\kappa\alpha_2|$. If α_1, α_2 differ in sign, then β_1 is the difference between the largest and the smallest products while β_2 is the difference between the two middle ones. Hence $|\beta_1| > |\beta_2|$. If α_1, α_2 agree in sign, let $a = |\tau\alpha_1|$, $b = |\kappa\alpha_1|$, $c = |\tau\alpha_2|$, $d = |\kappa\alpha_2|$, then $ad = bc$ and $a > \{b,c\} > d$.

Claim. $a+d > b+c$

Reason. We may assume $b \geq c$ without loss of generality.

$$(b+(a-b))(c+(-c+d)) = ad$$

$$bc+b(d-c)+c(a-b)+(a-b)(d-c) = ad$$

So

$$b(d-c)+c(a-b)+(a-b)(d-c) = 0\,. \tag{2.3.2}$$

If $c-d \geq a-b$, then $b(c-d) \geq c(a-b)$ because $b > c$. This would imply the left hand side of (2.3.2) is negative with right hand side equal zero. So $a-b > c-d$, or $a+d > b+c$. Here ends the proof of the claim.

Now $|\beta_1| = a+d$, $|\beta_2| = b+c$, so $|\beta_1| > |\beta_2|$. □

Lemma 3. $|\alpha_1| = |\alpha_2|$ implies $|\beta_1| = |\beta_2|$,
$|\alpha_1| < |\alpha_2|$ implies $|\beta_1| < |\beta_2|$.

Proof. First statement is trivial by (2.3.1). Second statement follows from the symmetry of Lemma 2. □

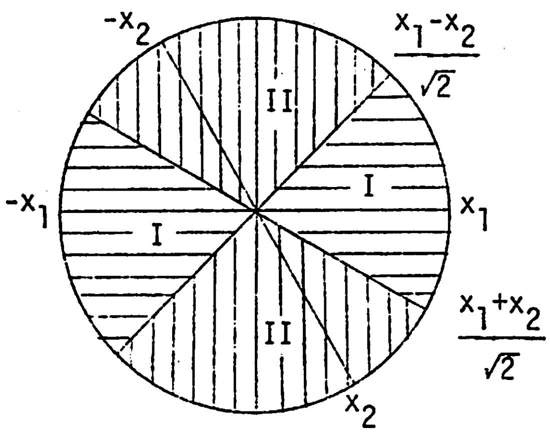


Figure 2.3.2

Theorem. Let C be a 2×2 matrix, and $V^{(0)} = \alpha_1^{(0)}x_1 + \alpha_2^{(0)}x_2$ be the starting vector. Then the Rayleigh sequence converges to an eigenvector if and only if $|\alpha_1^{(0)}| \neq |\alpha_2^{(0)}|$.

Proof. As mentioned at the beginning of this section, it is sufficient to consider $C = \begin{pmatrix}1 & \kappa\\ 0 & 0\end{pmatrix}$ with $\kappa \geq 0$. We can actually separate the half unit circle (see Figure 2.3.2) into invariant regions under the action of RQI:

Region I: $|\alpha_1| > |\alpha_2|$

If $|\alpha_1^{(k)}| > |\alpha_2^{(k)}|$, by Lemma 2, $|\beta_1^{(k)}| > |\beta_2^{(k)}|$.

Dropping the superscript when there is no confusion, we have from (2.2.1)

$$\rho_k = \rho(V^{(k)}) = \frac{\alpha_1\beta_1\gamma_1\lambda_1 + \alpha_2\beta_2\gamma_2\lambda_2}{\alpha_1\beta_1\gamma_1 + \alpha_2\beta_2\gamma_2}$$

and also $\lambda_1 = 1$, $\lambda_2 = 0$, $\gamma_1 = \gamma_2 = \gamma$.

Let $\alpha_1\beta_1\gamma = a+bi$, $\alpha_2\beta_2\gamma = c+di$. Then $b = -d$ because $\alpha_1\beta_1\gamma + \alpha_2\beta_2\gamma = 1$. $|\alpha_1\beta_1| > |\alpha_2\beta_2|$ implies $|a| > |c|$, but $a+c = 1$, so $a > \frac{1}{2}$. Hence

$$\mathrm{Re}(\rho_k) = \mathrm{Re}\left(\frac{\alpha_1\beta_1\gamma}{\alpha_1\beta_1\gamma + \alpha_2\beta_2\gamma}\right) = a > \frac{1}{2} .$$

From the definition of RQI, we have

$$v^{(k+1)} = \alpha_1^{(k+1)}x_1 + \alpha_2^{(k+1)}x_2 = \tau_k(C-\rho_k)^{-1}v^{(k)} .$$

So

$$\alpha_1^{(k+1)} = \tau_k(\lambda-\rho_k)^{-1}\alpha_1^{(k)}$$

and

$$(2.3.3) \qquad \left|\frac{\alpha_1^{(k+1)}}{\alpha_2^{(k+1)}}\right| = \left|\frac{(1-\rho_k)^{-1}\alpha_1^{(k)}}{(0-\rho_k)^{-1}\alpha_2^{(k)}}\right| = \left|\frac{\alpha_1^{(k)}}{\alpha_2^{(k)}}\right|\left|\frac{\rho_k}{1-\rho_k}\right| > \left|\frac{\alpha_1^{(k)}}{\alpha_2^{(k)}}\right| ,$$

since $\mathrm{Re}(\rho_k) > \frac{1}{2}$. Consequently $|\alpha_1^{(k+1)}| > |\alpha_2^{(k+1)}|$ and thus region I is invariant.

We now show that region I has a single attractive fixed point. Let $|\alpha_1^{(0)}| > |\alpha_2^{(0)}|$, then $|\alpha_1^{(k)}| > |\alpha_2^{(k)}|$ and $|\beta_1^{(k)}| > |\beta_2^{(k)}|$ for each k. Then

$$\left|\frac{\alpha_1^{(k)}\beta_1^{(k)}}{\alpha_2^{(k)}\beta_2^{(k)}}\right| > \left|\frac{\alpha_1^{(k)}}{\alpha_2^{(k)}}\right| \qquad \text{(because } |\beta_1^{(k)}| > |\beta_2^{(k)}|\text{)}$$

$$\left.\begin{aligned} &> \left|\frac{\alpha_1^{(k-1)}}{\alpha_2^{(k-1)}}\right| \\ &\quad\vdots \\ &> \left|\frac{\alpha_1^{(0)}}{\alpha_2^{(0)}}\right| . \end{aligned}\right\} \begin{array}{l}\text{by (2.3.3) since region I is invariant} \\ \text{and } |\alpha_1^{(0)}| > |\alpha_2^{(0)}| \text{ by assumption}\end{array}$$

Let $\left|\frac{\alpha_1^{(0)}}{\alpha_2^{(0)}}\right| = \omega > 1$. Let $\gamma\alpha_1^{(k)}\beta_1^{(k)} = a + bi$, $\gamma\alpha_2^{(k)}\beta_2^{(k)} = C - bi$. Now $|\alpha_i|$, $|\beta_i|$ are bounded since $\|v\| = 1$. Therefore $|b|$ is bounded by a constant, say M. By the result above,

$$\left|\frac{\alpha_1^{(k)}\beta_1^{(k)}}{\alpha_2^{(k)}\beta_2^{(k)}}\right| = \sqrt{\frac{a^2+b^2}{c^2+d^2}} > \omega \ .$$

Therefore

$$\frac{a^2+b^2}{c^2+d^2} > \omega^2 \ ,$$

$$a^2+b^2 > \omega^2 c^2 + \omega^2 b^2 \ ,$$

$$\frac{a^2}{c^2} > \omega^2 + (\omega^2-1)b^2 > 1 \ ,$$

$$\frac{a}{c} > \sqrt{\omega^2+(\omega^2-1)b^2} > \omega \ .$$

So

$$\mathrm{Re}(\rho_k) = \mathrm{Re}\left(\frac{\alpha_1^{(k)}\beta_1^{(k)}\gamma}{\gamma\alpha_1^{(k)}\beta_1^{(k)} + \alpha_2^{(k)}\beta_2^{(k)}\gamma}\right) > \frac{\omega}{1+\omega} > \frac{1}{2}$$

for each k. Repeated application of (2.3.3) brings us to

$$\left|\frac{\alpha_1^{(k)}}{\alpha_2^{(k)}}\right| = \left(\prod_{i=0}^{k}\left|\frac{\rho_i}{1-\rho_i}\right|\right)\cdot\left|\frac{\alpha_1^{(0)}}{\alpha_2^{(0)}}\right| \to \infty \quad \text{as} \quad k \to \infty$$

because

$$\left|\frac{\rho_i}{1-\rho_i}\right| = \left|\frac{a+bi}{1-a-bi}\right| = \sqrt{\frac{a^2+b^2}{(1-a)^2+b^2}} > \sqrt{\frac{\left(\frac{\omega}{1+\omega}\right)^2+b^2}{\left(\frac{1}{1+\omega}\right)^2+b^2}} \geq \sqrt{\frac{\left(\frac{\omega}{1+\omega}\right)^2+M^2}{\left(\frac{1}{1+\omega}\right)^2+M^2}} > 1 \ .$$

In other words, $v^{(k)}$ converges to x_1.

Region II: $|\alpha_1| < |\alpha_2|$

By similar arguments, this region is invariant and if $v^{(0)}$ belongs to region II, $v^{(k)}$ converges to x_2.

Region III: $|\alpha_1| = |\alpha_2|$

If $|\alpha_1^{(i)}| = |\alpha_2^{(i)}|$, then $|\beta_1^{(i)}| = |\beta_2^{(i)}|$.

$$\operatorname{Re}(\rho_i) = \operatorname{Re}\left(\frac{\alpha_1\beta_1}{\alpha_1\beta_1 + \alpha_2\beta_2}\right) = \frac{1}{2}$$

because $\gamma(\alpha_1\beta_1 + \alpha_2\beta_2) = 1$.

$$\left|\frac{\alpha_1^{(i+1)}}{\alpha_2^{(i+1)}}\right| = \left|\frac{-\rho_i}{1-\rho_i}\right|\left|\frac{\alpha_1^{(i)}}{\alpha_2^{(i)}}\right| = \left|\frac{\alpha_1^{(i)}}{\alpha_2^{(i)}}\right| .$$

So if $|\alpha_1^{(0)}| = |\alpha_2^{(0)}|$ then $|\alpha_1^{(i)}| = |\alpha_2^{(i)}|$ for all i.

This region is invariant and the vector sequence will not converge. Therefore, the theorem is proved and the sequence converges if and only if $v^{(0)}$ is not a bisector of the eigenvectors. □

Value of Residual Norm for Bisectors

Now, let us compute the residual of the eigenvector bisectors:

Let $u_1 = (x_1 - x_2)/\|x_1 - x_2\|$, $u_2 = (x_1 + x_2)/\|x_1 + x_2\|$. Then

$$u_1 = (1 - \tfrac{\kappa}{\tau}, \tfrac{1}{\tau})^T / \|(1 - \tfrac{\kappa}{\tau}, \tfrac{1}{\tau})\| \quad ;$$

$$u_2 = (1 + \tfrac{\kappa}{\tau}, -\tfrac{1}{\tau})^T / \|(1 + \tfrac{\kappa}{\tau}, \tfrac{1}{\tau})\| \quad .$$

Let $\theta = \tan^{-1}\kappa$. Then

$$u_1 = (1-\sin\theta, \cos\theta)^T/\sqrt{2-2\sin\theta}$$
$$u_2 = (1+\sin\theta, -\cos\theta)^T/\sqrt{2+2\sin\theta}$$

$$(C-\tfrac{1}{2})^{-1} = \begin{pmatrix} \frac{1}{2} & \tan\theta \\ 0 & -\frac{1}{2} \end{pmatrix}^{-1} = \begin{pmatrix} 2 & 4\tan\theta \\ 0 & -2 \end{pmatrix}$$

$$\|r_1\| = \|(C-\tfrac{1}{2})^{-1}u_1\| = \frac{1}{\sqrt{2-2\sin\theta}}\left\|\begin{pmatrix} 2 & 4\tan\theta \\ 0 & -2 \end{pmatrix}\begin{pmatrix} 1-\sin\theta \\ \cos\theta \end{pmatrix}\right\|$$
$$= \frac{1}{\sqrt{2-2\sin\theta}}\left\|\begin{pmatrix} 2+2\sin\theta \\ -2\cos\theta \end{pmatrix}\right\|$$
$$= 2\sqrt{(2+\sin\theta)/(2-\sin\theta)}$$
$$\|r_2\| = \|(C-\tfrac{1}{2})^{-1}u_2\| = \frac{1}{\sqrt{2-2\sin\theta}}\left\|\begin{pmatrix} 2 & 4\tan\theta \\ 0 & -2 \end{pmatrix}\begin{pmatrix} 1+\sin\theta \\ -\cos\theta \end{pmatrix}\right\|$$
$$= 2\sqrt{(2-\sin\theta)/(2+\sin\theta)} .$$

So

$$\|r_1\| = \|r_2\| \quad \text{iff} \quad \sqrt{(2+\sin\theta)/(2-\sin\theta)} = \sqrt{(2-\sin\theta)/(2+\sin\theta)}$$
$$\text{iff} \quad \sin\theta = 0$$
$$\text{iff} \quad \theta = 0 \quad \text{iff} \quad \kappa = 0 \quad \text{iff} \quad C \text{ is normal.}$$

We have just shown that if C is non-normal, then the residual of limit vectors can be different. Thus monotonicity of residuals is lost immediately.

§2.4 The Hermitian Case

W. Kahan and B. Parlett [5] presented a proof of the convergence of RQI in the Hermitian (symmetric) case in 1968 through the use of the monotonicity of the residual. In previous sections, we have seen the irregular behaviour of the residual in non-normal matrices. In this section, we shall present a characterization of limit vectors without the use of the residual, so that we may extend the result further into the nearly normal case in the next section.

Let A be an $n \times n$ matrix such that $A^* = A$. In this case, the eigenvalues λ_i of A are real and $\{x_i = y_i, \ i = 1, \ldots, n\}$ form an orthonormal basis. Further, we arrange the λ_i's so that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Let v be a vector. Then $\rho(v) = v^*Av = \overline{v^*A^*v} = \overline{v^*Av}$ is real.

Let $V^{(0)}$ be the initial vector and $\{V^{(k)} \mid \|V^{(k)}\| = 1, \ k = 1,2,\ldots\}$ be the Rayleigh sequence from A and $V^{(0)}$. Let

$$V^{(k)} = \alpha_1^{(k)} x_1 + \alpha_2^{(k)} x_2 + \cdots + \alpha_n^{(k)} x_n \ .$$

Then from the relation

$$(A - \rho(V^{(k)}))V^{(k+1)} = \tau_k V^{(k)}$$

when $A - \rho(V^{(k)})$ is invertible, we have

$$\alpha_i^{(k+1)} = \tau_k \frac{\alpha_i^{(k)}}{\lambda_i - \rho(V^{(k)})} \tag{2.4.1}$$

so the action of RQI induces an increase in $|\alpha_i|$ which is inversely proportional to the difference between $\rho(V^{(k)})$ and λ_i.

Also, the Rayleigh Quotient $\rho(V^{(k)})$ is a mean of the eigenvalues:

$$\rho(V^{(k)}) = \sum_{i=1}^{n} \lambda_i |\alpha_i^{(k)}|^2 \qquad \text{(see §2.2)} \ .$$

With these in mind, we would like to show that the only limit point of the Rayleigh sequence is either an eigenvector or a bisector of a pair of eigenvectors. Before we proceed with the main theorem, we shall illustrate the main idea through a 3×3 example.

Example. Let $A = \text{diag}(\lambda_1,\lambda_2,\lambda_3)$ where $\lambda_1 < \lambda_2 < \lambda_3$, and $x_i = e_i$, $i = 1,2,3$ are the eigenvectors. Let $V^{(0)} = \alpha_1^{(0)}x_1 + \alpha_2^{(0)}x_2 + \alpha_3^{(0)}x_3$ be the starting vector and $V^{(k)} = \alpha_1^{(k)}x_1 + \alpha_2^{(k)}x_2 + \alpha_3^{(k)}x_3$ be the vector at the k^{th} step and let $V = \alpha_1x_1 + \alpha_2x_2 + \alpha_3x_3$ be a limit vector. The crucial step is to show that at least one of α_1, α_2, α_3 must be zero.

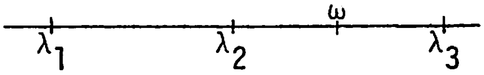


Figure 2.4.1

Suppose $\alpha_1\alpha_2\alpha_3 \neq 0$. Then consider the scalar sequence $\rho_k = \rho(V^{(k)})$, $k = 0,1,2,\ldots$. Let $\omega = (\lambda_2+\lambda_3)/2$. There are three possibilities:

(1) $\rho_k \leq \omega$ for all k but ρ_k converges to ω.

Notice we may assume $\alpha_i^{(0)} \neq 0$, $i = 1,2,3$ because if $\alpha_j^{(0)} = 0$ then $\alpha_j^{(k)} = 0$ for all k by (2.4.1) and thus $\alpha_j = 0$. And for the same reason, assume $\alpha_i^{(k)} \neq 0$ for all finite k and all i. ρ_k converges to ω means there exists a constant N such that if $k \geq N$, $|\rho_k-\omega| < \frac{\omega-\lambda_2}{2}$. Consider

$$\left|\frac{\alpha_2^{(N+k+1)}}{\alpha_1^{(N+k+1)}}\right| = \left|\frac{(\lambda_2-\rho_{k+N})^{-1}\alpha_2^{(N+k)}}{(\lambda_1-\rho_{k+N})^{-1}\alpha_1^{(N+k)}}\right| \quad \text{by (2.4.1)}$$

$$= \left(\prod_{i=0}^{k}\left|\frac{\lambda_1-\rho_{N+i}}{\lambda_2-\rho_{N+i}}\right|\right)\left|\frac{\alpha_2^{(N)}}{\alpha_1^{(N)}}\right| \to \infty \quad \text{as} \quad k \to \infty$$

because each of the terms $|(\lambda_1-\rho_{N+i})/(\lambda_2-\rho_{N+i})| > \frac{\omega-\lambda_1}{\omega-\lambda_2} > 1$, and $\alpha_2^{(N)} \neq 0$. That means $|\alpha_1^{(k)}| \to 0$ as $k \to \infty$, which implies $\alpha_1 = 0$, a contradiction.

(2) $\rho_k \leq \omega$ <u>for all</u> k <u>and</u> ρ_k <u>does not converge to</u> ω.

Then there exists an $\varepsilon > 0$ such that $\rho_k \leq \omega-\varepsilon$ for infinitely many k. Hence

$$\left|\frac{\alpha_2^{(k+1)}}{\alpha_3^{(k+1)}}\right| = \left(\prod_{i=0}^{k}\left|\frac{\lambda_3-\rho_i}{\lambda_2-\rho_i}\right|\right)\left|\frac{\alpha_2^{(0)}}{\alpha_3^{(0)}}\right|$$

$$\left|\frac{\lambda_3-\rho_i}{\lambda_2-\rho_i}\right| \geq 1 \quad \text{because} \quad \rho_i \leq \omega$$

and

$$\left|\frac{\lambda_3-\rho_i}{\lambda_2-\rho_i}\right| \geq \frac{\omega+\varepsilon}{\omega-\varepsilon} > 1 \quad \text{for infinitely many} \quad i \ ,$$

so the ratio tends to infinity, which means $|\alpha_3^{(k)}| \to 0$ as $k \to \infty$. Thus $\alpha_3 = 0$, a contradiction.

(3) <u>There exists a</u> $\rho_k > \omega$.

This inequality is "invariant" under subsequent RQI steps, i.e., $\rho_{k+i} > \omega$ for all i. Actually, we shall show that $\omega < \rho_k \leq \rho_{k+1} \leq \rho_{k+2} \leq \cdots \leq \rho_{k+i} \leq \cdots$. The reason is clear if we consider the following

$$\left|\frac{\alpha_3^{(k+1)}}{\alpha_2^{(k+1)}}\right| = \left|\frac{\lambda_2-\rho_k}{\lambda_3-\rho_k}\right|\left|\frac{\alpha_3^{(k)}}{\alpha_2^{(k)}}\right| > \left|\frac{\alpha_3^{(k)}}{\alpha_2^{(k)}}\right| \quad \text{since} \quad \rho_k > \omega \ .$$

Similarly

$$\left|\frac{\alpha_3^{(k+1)}}{\alpha_1^{(k+1)}}\right| > \left|\frac{\alpha_3^{(k)}}{\alpha_1^{(k)}}\right| .$$

Therefore $|\alpha_3^{(k)}|$ increases with respect to $|\alpha_2^{(k)}|$ and $|\alpha_1^{(k)}|$. We know that for

$$1 = \sum|\alpha_i^{(k)}|^2 = \sum|\alpha_i^{(k+1)}|^2$$

$$\begin{aligned} \rho_k &= \sum\lambda_i|\alpha_i^{(k)}|^2 \\ &\leq \sum\lambda_i|\alpha_i^{(k+1)}|^2 \quad \text{because more weight is put on } \lambda_3 \text{ and } \lambda_3 > \lambda_2 > \lambda_1 \\ &= \rho_{k+1} \,. \end{aligned}$$

So we have $\omega < \rho_k \leq \rho_{k+1} \leq \rho_{k+2} \leq \cdots$ and

$$\left|\frac{\alpha_\ell^{(k+j)}}{\alpha_3^{(k+j)}}\right| = \left(\prod_{i=0}^{j-1}\left|\frac{\lambda_3-\rho_{k+i}}{\lambda_\ell-\rho_{k+i}}\right|\right)\left|\frac{\alpha_\ell^{(k)}}{\alpha_3^{(k)}}\right| \to \infty \quad \text{as} \quad j \to \infty$$

since

$$\left|\frac{\lambda_3-\rho_{k+i}}{\lambda_\ell-\rho_{k+i}}\right| \geq \left|\frac{\lambda_3-\rho_k}{\lambda_\ell-\rho_k}\right| > 1 \,, \qquad \ell = 1,2 \,.$$

Therefore $\alpha_1 = \alpha_2 = 0$, which means $V^{(k)}$ converge to x_3, a contradiction that $\alpha_1\alpha_2\alpha_3 \neq 0$. So if v is a limit vector, then v is a linear combination of at most two eigenvectors. If $v = \alpha_1 x_1 + \alpha_2 x_2$, then $|\alpha_1| = |\alpha_2|$, this will be shown in the last part of the forthcoming theorem. □

When A is $n \times n$, the picture is more complicated, and that is why we prove the theorem separately.

The main trouble is when $V = \sum\alpha_i x_i$ with $\alpha_\ell \neq 0$ and $\alpha_i = 0$, $i > \ell$. Then the existence of a $\rho_k > (\lambda_\ell + \lambda_{\ell-1})/2$ does not guarantee that $\rho_{k+1} \geq \rho_k$ because $\alpha_{\ell+1}^{(k)}$ may decrease at the $(k+1)^{st}$ step and ρ_{k+1} may be less than ρ_k. But if we let k be so large that $V^{(k)}$

is very close to V, then even though ρ_{k+1} may be less than ρ_k, the subsequent ρ_k's are confined to a narrow interval. This is the essence of the following theorem.

Theorem. Let $V^{(0)}$ *be the initial vector and let* V *be a limit vector of the Rayleigh sequence from* $A = A^*$ *and* $V^{(0)}$. *Then either*

1) V *is an eigenvector and* $V^{(k)} \to V = x_i$ *and* $\rho(V^{(k)}) \to \lambda_i$

or 2) V *is the internal or external bisector of two eigenvectors belonging to distinct eigenvalues and the sequence* $V^{(k)}$, $k = 1,2,\ldots$ *oscillates between the bisectors,* $\rho_k \to \rho$ *equals the mean of the eigenvalues*.

Proof. 1) V is an eigenvector. Ostrowski [4] described an eigenvector as an attractive fixed point of the RQI, i.e., if $V^{(k)}$ is close enough to V, then $\rho_k \to \lambda_i$, the eigenvalue associated with V and $V^{(k)}$ converges to a vector in the eigenspace of λ_i.

2) If V is not an eigenvector, then the crucial step is to show that V is a linear combination of only two eigenvectors. So assume $V = \alpha_1 x_1 + \cdots + \alpha_n x_n$ with α_p, α_r, α_ℓ being non-zero and $\alpha_i = 0$ for $n \geq i > \ell$. (If $\ell = n$, then we have essentially the 3×3 example before, and the first three cases listed later will be sufficient and we may skip most of the proof). We proceed to show that if $V^{(k_0)}$ is close enough to V, then there exists an open neighborhood of V such that $V^{(k)}$ is outside of it when k is large and hence contradicts the assumption that V is a limit vector.

Now, V being a limit vector means that there exists k_0 such that $V^{(k_0)}$ is so close to V that $|\alpha_i^{(k_0)} - \alpha_i| < \varepsilon$ where ε satisfies the following conditions:

(i) $\varepsilon < .01|\alpha_\ell|$ and let $\mu = .99|\alpha_\ell|$. Let $s < m < \ell < q$ be indices such that $\alpha_s^{(0)}$, $\alpha_m^{(0)}$, $\alpha_q^{(0)}$ are nonzero and all $\alpha_i^{(0)} = 0$ if i is between any two of them. Then by (2.4.1), we know all those $\alpha_i^{(k)} = 0$ and $\alpha_s^{(k)}$, $\alpha_m^{(k)}$, $\alpha_q^{(k)}$ non-zero for all finite k. Let $\omega = (\lambda_m+\lambda_\ell)/2$ (see Figure 2.4.2).

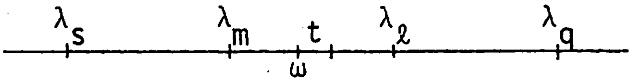


Figure 2.4.2

(ii) ε is chosen so small that

$$\mu^2\left(\frac{\lambda_q-\lambda_\ell}{2}\right) > \sum_{i=q}^{n} \varepsilon^2\left(\lambda_i - \frac{\lambda_q+\lambda_\ell}{2}\right) .$$

(iii) If the scalar t is given by the following relationship

$$\mu^2|\lambda_\ell-\omega| = \mu^2|\lambda_\ell-\omega-t| + \sum_{i=q}^{n} \varepsilon^2|\lambda_i-\omega-t| \tag{2.4.2}$$

then ε is chosen so small that $t < \frac{\lambda_\ell-\omega}{4}$.

(iv) Let ϕ_i, $i = 1,\ldots,s$ be positive numbers defined as

$$\phi_i^2\left(1 - \frac{\omega-\lambda_i}{\lambda_\ell-\omega}\right)(\omega-\lambda_i) = \sum_{j=q}^{n} (\lambda_j-\omega)\varepsilon^2$$

and $\phi_m \geq 0$ is defined as

$$\sum_{i=1}^{m} \phi_i^2\left(\omega + \frac{\lambda_\ell-\omega}{4} - \lambda_i\right) = \mu^2\left(\lambda_\ell - \frac{\lambda_\ell-\omega}{4}\right) .$$

We can see that if $\varepsilon \to 0$, $\phi_i \to 0$ for $i = 1,\ldots,s$ and

$$\phi_m^2 \to \mu^2 \frac{(\lambda_\ell - \frac{\lambda_\ell-\omega}{4})}{(\omega + \frac{\lambda_\ell-\omega}{4} - \lambda_i)} .$$

Let ε be chosen so small such that

$$\phi_m^2(\lambda_m - \tfrac{1}{3}(\lambda_m-\lambda_s)) \geq \sum_{i=1}^{s} \phi_i^2(\lambda_i - \frac{2\lambda_m+\lambda_s}{3}) . \tag{2.4.3}$$

We are now ready to proceed. Consider the scalar sequence $\rho_k = \rho(V^{(k)})$.

<u>Case 1</u>. $\rho_k \leq \omega$ <u>but</u> $\rho_k \to \omega$.

Remember $\alpha_p \neq 0$ in the expression for the limit vector V, but examining (2.4.1) reveals that

$$\left|\frac{\alpha_p^{(k+1)}}{\alpha_m^{(k+1)}}\right| = \left|\frac{\lambda_m-\rho(V^{(k)})}{\lambda_p-\rho(V^{(k)})}\right| \left|\frac{\alpha_p^{(k)}}{\alpha_m^{(k)}}\right| ,$$

and $\rho_k \to \omega$ implies $\left|\frac{\lambda_m-\rho(V^{(k)})}{\lambda_p-\rho(V^{(k)})}\right| \leq \psi < 1$ for $k \geq k_1$ for some k_1. Thus $\left|\frac{\alpha_p^{(k)}}{\alpha_m^{(k)}}\right| \to 0$ as $k \to \infty$. $|\alpha_m^{(k)}| \leq 1$ implies $|\alpha_p^{(k)}| \to 0$, a contradiction to $\alpha_p \neq 0$.

<u>Case 2</u>. $\rho_k \leq \omega$ <u>but there exists an open neighborhood</u> N_ω <u>around</u> ω <u>such that</u> $\rho_k \notin N_\omega$ <u>for infinitely many</u> k's.

Examining (2.4.1) shows that $|\alpha_\ell^{(k)}/\alpha_m^{(k)}| \to 0$ as $k \to \infty$ which in turn implies $|\alpha_\ell^{(k)}| \to 0$, a contradiction to $\alpha_\ell \neq 0$.

So we may assume that there exists ρ_{k_1} such that $\rho_{k_1} > \omega$.

Let

$$t = \begin{cases} 0 & \text{if } \ell = n \\ \text{defined in (2.4.2)} & \text{if } \ell < n . \end{cases}$$

Case 3. Let $\rho_{k_1} \geq \omega + t$.

By (ii) in the choice of ε, we first have $\frac{\lambda_\ell + \lambda_q}{2} > \rho_k$ for all $k \geq k_1$. By the action of RQI, i.e. equation (2.4.1), we know $|\alpha_\ell^{(k)}|$ increases with respect to $|\alpha_i^{(k)}|$ for $i \neq \ell$. The choice of t insures that $\rho_k > \omega$ for all $k \geq k_1$ with $\rho_k \not\to \omega$, $\rho_k \not\to \frac{\lambda_\ell + \lambda_q}{2}$. Therefore, as $k \to \infty$, $|\alpha_i^{(k)}/\alpha_\ell^{(k)}| \to 0$ for all $i \neq \ell$. Thus $V^{(k)} \to x_\ell$, a contradiction to V being a linear combination of eigenvectors of distinct eigenvalues.

Case 4. $\omega + t > \rho_{k_1} > \omega$

Two possibilities exist:

(a) ρ_k is monotonic increasing from this point on. Then $\rho_k \to \lambda_\ell$ and $V^{(k)} \to \lambda_\ell$ as before, a contradiction.

(b) $F = \{k |\ k \geq k_1$ such that $\rho_{k+1} < \rho_k\} \neq \emptyset$. Let $k_2 = \min\{k | k \in F\}$.

If we look closer at the formula $\rho_{k_2} = \sum_{i=1}^{n} \lambda_i |\alpha_i^{(k_2)}|^2$ and regard ρ_k as the center of gravity of a weightless rod that has weights $|\alpha_i^{(k)}|^2$ at positions λ_i, then the hypothesis implies: loss of moment on the right of ρ_k after adjustments due to RQI > loss of moments on the left of ρ_k. Before we calculate how much is lost on each side, we would like to remind the reader that ρ_{k_2} must be in the interval $(\omega;\omega+t)$, hence $|\alpha_i^{(k_2)}|$, $i \neq \ell$ decreases with respect to $|\alpha_\ell^{(k_2)}|$. If we use the normalization that $|\alpha_\ell^{(k_2)}| = \mu$, then the moment lost on the right can at most be

$$\sum_{i=q}^{n} \varepsilon^2(\lambda_i-\omega) \ . \qquad \text{(I)}$$

(Notice $|\alpha_i^{(k)}/\alpha_\ell^{(k)}| > |\varepsilon/\mu|$ for $k \geq k_0$, $i > \ell$ because $\rho_k < \frac{\lambda_\ell+\lambda_q}{2}$). And the moment lost on the left is at least

$$\sum_{i=1}^{s} |\alpha_i^{(k_2)}|^2 (1 - \frac{\omega-\lambda_i}{\lambda_\ell-\omega})(\omega-\lambda_i) \ . \qquad \text{(II)}$$

The hypothesis implies (I) > (II) and by definition of ϕ_i, $i = 1,\dots,s$, we have $\phi_i \geq |\alpha_i^{(k_2)}|$ for $i = 1,\dots,s$. This, together with $\omega < \rho_{k_2} < \omega+t$, implies $|\alpha_m^{(k_2)}| \geq \phi_m$. Hence

$$|\alpha_m^{(k_2)}|^2(\lambda_m - \frac{1}{3}(\lambda_m-\lambda_s)) \geq \sum_{i=1}^{s} |\alpha_i^{(k_2)}|^2(\lambda_i - \frac{2\lambda_m+\lambda_s}{3}) \text{ by (2.4.3) .}$$

This insures that $\rho_k \geq \frac{2\lambda_m+\lambda_s}{3}$ for $k \geq k_2$ and therefore $|\alpha_i^{(k)}/\alpha_m^{(k)}| \to 0$ as $k \to \infty$ for $i = 1,\dots,s$. In particular, $|\alpha_p^{(k)}| \to 0$, a contradiction to $\alpha_p \neq 0$. Here ends the proof of the crucial step that V is a linear combination of only two eigenvectors.

Now $V = \alpha_i x_i + \alpha_j x_j$. If $|\alpha_i| \neq |\alpha_j|$, say $|\alpha_i| > |\alpha_j|$, then the Rayleigh sequence generated from V would converge to x_i by the result of the last section. We know that x_i is an attractive fixed point and there is an open enighborhood N_x around x_i such that if a vector falls into that neighborhood, the subsequent Rayleigh Quotients would converge to λ_i (see Ostrowski [4]). The RQI function is continuous and will map an open neighborhood of V, N_V, into N_x. Therefore if V is a limit point, then there exists $V^{(k)} \in N_V$ and thus $V^{(k)} \to x_i$ as $k \to \infty$, a contradiction to the assumption that V is a limit point. Therefore $|\alpha_i| = |\alpha_j|$ and V is an external or internal bisector and $\rho_k \to \rho(V) = (\lambda_i+\lambda_j)/2$. □

The instability of the case where V is a bisector will be discussed in Section 8.

§2.5 The Nearly Normal Case

We have seen the characterization of limit vectors for a Hermitian matrix without the use of residual in the last section. Here, we shall prove a similar result for real nearly normal matrices with real spectrum.

The proof itself is a modification of the last theorem. Hence we shall not present the same proof twice, but rather we shall discuss what properties are lost when we do not have a normal matrix, and how we can modify the proof accordingly.

By translation invariance property, we may assume $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. From (2.2.1), we have, for $v = \sum_{i=1}^{n} \alpha_i x_i$,

$$\rho(v) = \frac{\sum \alpha_i \beta_i \gamma_i \lambda_i}{\sum \alpha_i \beta_i \gamma_i} .$$

By nearly normal, we mean $\max_i |1-\gamma_i| \leq \delta$ for some small positive δ.

As in the last section, we want to show that the only limit vectors of RQI are the eigenvectors and the internal and external bisectors of two eigenvectors. Once again, let V be a limit vector which is a linear combination of three or more eigenvectors with distinct eigenvalues. We shall draw a contradiction from that assumption if δ is small enough.

The important properties that are lost in non-normal cases are:

(i) In the expression for $\rho(v)$ with $\lambda_i, \gamma_i \geq 0$, the term $\alpha_i\beta_i\lambda_i\gamma_i$ may be negative whereas $\alpha_i\beta_i\gamma_i = |\alpha_i|^2$ in the normal case. We like to know when would $\alpha_i\beta_i$ be negative? If we consider the space spanned by $\{x_i, y_i\}$, and in the same plane draw the $n-1$ dimensional hyperplanes spanned by the other $n-1$ column eigenvectors and $n-1$ row eigenvectors respectively, we have Figure 2.5.2.

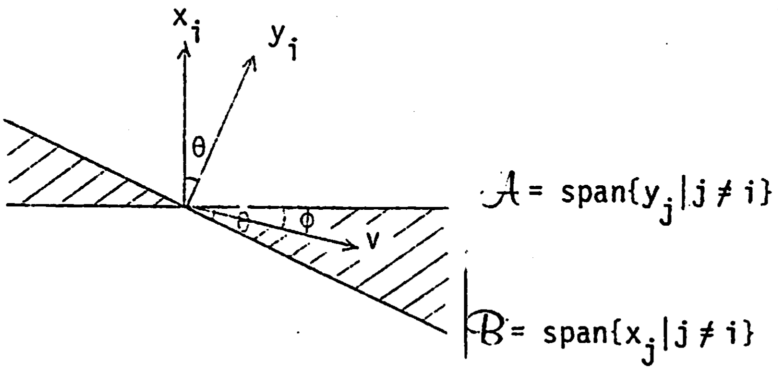
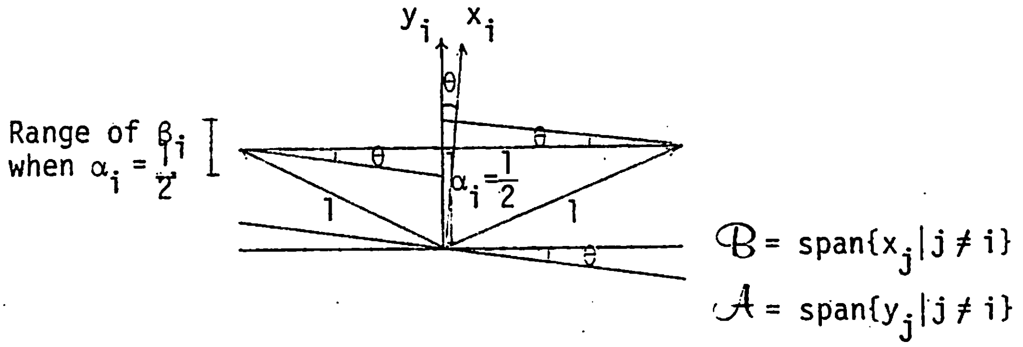

x_i
y_i
θ
θ
ϕ
v
$\mathcal{A} = \text{span}\{y_j | j \neq i\}$
$\mathcal{B} = \text{span}\{x_j | j \neq i\}$

Figure 2.5.2

So the only place where $\alpha_i\beta_i$ is negative is when the projection of v onto the plane spanned by $\{x_i, y_i\}$ is in the shaded area. Let $\measuredangle x_i y_i$ be θ and $\measuredangle v\mathcal{A}$ be $\phi \leq \theta$. Then $|\alpha_i\beta_i| = |\sin\phi \sin(\theta-\phi)/\cos^2\theta|$. So $|\alpha_i\beta_i| \to 0$ faster than $|\tan^2\theta|$ as $\theta \to 0$. We also know $\theta \to 0$ as $\delta = 1 - \cos\theta \to 0$. Therefore, for each $\varepsilon_1 > 0$, we can choose δ so small that $\alpha_i\beta_i\gamma_i < 0$ if and only if $|\alpha_i\beta_i\gamma_i| < \varepsilon_1$, for each i.

(ii) The numerical range of ρ is no longer confined to the interval $[\lambda_1, \lambda_n]$. But $\max\{\alpha_i\beta_i\gamma_i | \ i = 1, \dots, n\} \leq 1 + (n-1)\varepsilon_1$, so if $\lambda_i\varepsilon_1$ is small with respect to $|\lambda_{i+1} - \lambda_i|$ for each i, then ρ cannot be close to λ_n when $\alpha_n = 0$. The effect of this has to be taken in account also.

(iii) Finally, there is the problem in non-normal cases that if $|\alpha_i^{(k)}|$ increases to $|\alpha_i^{(k+1)}|$, $|\beta_i^{(k)}|$ may or may not increase. So we like to have some control over the behaviour of $\beta_i^{(k)}$ when $\alpha_i^{(k)}$ is not small, say $|\alpha_i^{(k)}| \geq \frac{1}{2}$. (See Figure 2.5.3.)



y_i
x_i
θ
θ
θ
θ
Range of β_i when $\alpha_i = \frac{1}{2}$
$\alpha_i = \frac{1}{2}$
1
1
$\mathcal{B} = \text{span}\{x_j | j \neq i\}$
$\mathcal{A} = \text{span}\{y_j | j \neq i\}$

Figure 2.5.3

$$(2.5.1) \quad \alpha_i \cos\theta - \tan\theta\sqrt{1-(\alpha_i\cos\theta)^2} \leq \beta_i \leq \alpha_i\cos\theta + \tan\theta\sqrt{1-(\alpha_i\cos\theta)^2} .$$

Therefore $|\beta_i^{(k)}-\alpha_i^{(k)}| \leq \varepsilon_2|\alpha_i^{(k)}|$ where $\varepsilon_2 \leq 4\delta$ for $|\alpha_i^{(k)}| \geq \frac{1}{2}$.

Now, let $v = \sum\alpha_i x_i$, of which at least three or more α_i's corresponding to distinct λ_i's are non-zero, be a limit vector of the Rayleigh sequence. Choose $V^{(k_0)}$ so close to V such that $|\alpha_i^{(k_0)}-\alpha_i| < \varepsilon$. If δ is sufficiently small and ε is picked carefully, we have the same proof as before. Case 1 and Case 2 need no modification. In Case 3, the definition of t is a little more complicated than before, because we want to make sure if $\rho_{k_1} \geq \omega + t$ for some k_1, then $V^{(k_1)}$ is rich in x_ℓ, i.e., $\rho(V^{(k_1)})$ is closer to λ_ℓ than any other λ_i not because some $\alpha_i\beta_i\gamma_i$, $i \neq \ell$ is much greater than one (insured by having δ small) nor because $\alpha_i\beta_i\gamma_i$, $i > \ell$ is large (insured by $|\alpha_i\beta_i\gamma_i| < n\varepsilon$). The fact that δ is small also insures that at subsequent steps of RQI, when $\alpha_\ell^{(k)}$ increases, $\beta_\ell^{(k)}$

cannot drop off (because (2.5.1)) so that $\rho_k > \omega$ for $k \geq k_1$. Thus $v^{(k)}$ converge to x_ℓ as before. As for Case 4, one extra consideration must be taken: δ be small (and thus ε_1 small) so that if $\alpha_i^{(k)}\beta_i^{(k)}\gamma_i$ is negative (which would be like a weight of $|\alpha_i^{(k)}\beta_i^{(k)}\gamma_i|$ at $-\lambda_i$) the effect of $\alpha_i^{(k)}\beta_i^{(k)}\gamma_i\lambda_i$ in the expression of ρ_k is negligible.

Thus the convergence result proved in the Hermitian case holds here, a non-normal case, also.

Remarks. We have attempted above to present an idea on how to prove the nearly normal case through the extension of a proof of the Hermitian case. The proof is greatly simplified at the expense of choosing an extremely small δ. If we are willing to do some more detailed analysis, e.g., choose the translation of matrices such that $\lambda_1 = -\lambda_n$, obtain better estimates in conjunction with the bound on $|\alpha_i\beta_i\gamma_i|$, etc., we can come up with a larger δ. We like to emphasize here that the main goal in this section is to present a sketch of the global behaviour of RQI for a non-normal matrix, but not to obtain a theorem as powerful as we could. In fact, there are reasons to believe the "nearly normal" condition can be replaced by "well-conditioned", because for the conclusion of the theorem to be false, ρ_k's have to jump around and lie frequently close to each eigenvalue whose associated eigenvector has a non-zero component in the expression for a limit vector. But for a well-conditioned matrix, it is not hard, only tedious, to trace the locus of ρ_k generated by RQI. Therefore, it leads us to conjecture that the same conclusion is true for well-conditioned matrices. Nonetheless, we shall not pursue this matter along this line because in the last two sections of this chapter, we shall look at RQI from a different perspective and reduce the question of global convergence of the vector iteration to that of the convergence of the scalar quantity ρ_k which lies in a compact space. And we believe that this is a more simple and elegant way to look at RQI.

§2.6 Bisectors of Eigenvectors as Limit Vectors of RQI

In previous sections, we have just shown that for sufficiently well-conditioned real matrices with real eigenvalues the necessary condition for vectors to be limit vectors of RQI is that they either be an eigenvector or bisectors of two eigenvectors. In this section, we show that for any non-defective matrix, the same condition is sufficient provided, in the case of bisector, that the mean of their associated eigenvalues is not an eigenvalue of the matrix in question.

Let x_1 and x_2 be the two eigenvectors for C, and by various invariant properties of RQI, we can have $\lambda_1 = 1$, $\lambda_2 = 0$, and we consider, without loss of generality, the action to take place in the coordinate system where $e_1 = x_1$, and x_2 a <u>real</u> linear combination of e_1 and e_2. Then, the first two components of y_1 and y_2 are real because $y_i^* x_j = \delta_{ij}\gamma_i \geq 0$, $\{i,j\} \subseteq \{1,2\}$. Let $\sum$ be the plane spanned by x_1 and x_2. Let u_1, u_2 be real unit vectors on $\sum$ that are orthogonal to x_2, x_1 respectively with the angles between u_1, x_1 and u_2, x_2 being acute. Let P be a function that takes a vector and projects it on $\sum$, i.e., P(w) is the orthogonal projection of w onto $\sum$. Then $u_1 = P(y_1)/\|P(y_1)\|$, $u_2 = P(y_2)/\|P(y_2)\|$ because y_1, y_2 lie in hyperplanes that are orthogonal to x_2, x_1 respectively. It is also obvious that $u_1^* x_1 = u_2^* x_2$, i.e., the angles are the same, call it θ.

Lemma. $(y_1^*u_1)(x_1^*u_1) = x_1^*y_1$, $(y_2^*u_2)(x_2^*u_2) = x_2^*y_2$.

Proof. Let $y_1 = \eta_1 x_1 + \eta_2 e_2 + w$ where $w^*x_1 = w^*e_2 = 0$. Then

Then $$u_1 = (\eta_1 x_1 + \eta_2 e_2)/\sqrt{\eta_1^2+\eta_2^2} \ .$$

So
$$\begin{aligned} y_1^*u_1 &= (\eta_1^2 + \eta_2^2)/\sqrt{\eta_1^2+\eta_2^2} \\ x_1^*u_1 &= \eta_1/\sqrt{\eta_1^2+\eta_2^2} \\ x_1^*y_1 &= \eta_1 \end{aligned}$$

$$(y_1^*u_1)(x_1^*u_1) = ((\eta_1^2+\eta_2^2)/\sqrt{\eta_1^2+\eta_2^2})\eta_1/\sqrt{\eta_1^2+\eta_2^2} = \eta_1 = x_1^*y_1 \ .$$

Similarly $(y_2^*u_2)(x_2^*u_2) = x_2^*y_2$. □

Now we are ready to translate our problem into one that deals with vectors on $\sum$, so that we can use results of Section 3.

Theorem. Let $Cx_1 = \lambda_1 x_1$, $Cx_2 = \lambda_2 x_2$. Then the bisectors of x_1 and x_2 are limit vectors of RQI if and only if $(\lambda_1+\lambda_2)/2$ is not an eigenvalue of C.

Proof. Let $v = \alpha_1 x_1 + \alpha_2 x_2$, $v^* = \beta_1 y_1^* + \beta_2 y_2^* + \cdots + \beta_n y_n^*$. First notice $P(y_j) = 0$ for $j > 2$ because $y_j^*x_1 = y_j^*x_2 = 0$. So $v = P(v) = P(\bar\beta_1 y_1) + P(\bar\beta_2 y_2) = \bar\beta_1(y_1^*u_1)u_1 + \bar\beta_2(y_2^*u_2)u_2$. From the lemma, we have

$$y_1^*u_1 = \frac{x_1^*y_1}{x_1^*u_1} = \frac{\gamma_1}{\cos\theta}, \qquad y_2^*u_2 = \frac{\gamma_2}{\cos\theta} \ .$$

So

$$v = (\bar\beta_1\gamma_1/\cos\theta)u_1 + (\bar\beta_2\gamma_2/\cos\theta)u_2 = \bar\zeta_1 u_1 + \bar\zeta_2 u_2$$

where $\zeta_j = \beta_j\gamma_j/\cos\theta$, $j = 1,2$. Then

$$\rho(v) = \frac{\alpha_1(\beta_1\gamma_1)\lambda_1 + \alpha_2(\beta_2\gamma_2)\lambda_2 + 0 + \cdots + 0}{\alpha_1\beta_1\gamma_1 + \alpha_2\beta_2\gamma_2 + 0 + \cdots + 0}$$

$$= \frac{\alpha_1\zeta_1 \cos\theta\ \lambda_1 + \alpha_2\zeta_2 \cos\theta\ \lambda_2}{\alpha_1\zeta_1 \cos\theta + \alpha_2\zeta_2 \cos\theta}$$

which brings us back to the situation of the 2×2 case with $\beta_i = \zeta_i$, $\gamma_i = \cos\theta$ (see §2.3). And the results of Section 3 apply.

So if $(\lambda_1+\lambda_2)/2$ is not an eigenvalue, the bisectors of x_1 and x_2 are limit vectors, but if $(\lambda_1+\lambda_2)/2$ is an eigenvalue, then by definition of RQI, it will give us the eigenvector associated with $(\lambda_1+\lambda_2)/2$. □

We know also that bisectors as limit points are unstable in the sense that if there is a slight perturbation in $\sum$ of v, the RQI will give us either x_1 or x_2, as in the 2×2 case.

<u>Note</u>. Bisector here means $v = \alpha_1 x_1 + \alpha_2 x_2$ where $|\alpha_1| = |\alpha_2|$.

§2.7 Characterization of Limit Vectors When ρ_k Converges

We know that $\rho_k = \rho(V^{(k)})$ is a sequence of numbers in a compact metric space (real or complex). In this section, we shall investigate what happens to $V^{(k)}$ if ρ_k converges.

<u>Definition</u>. If $v = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$, then we say v is deficient in x_i if $\alpha_i = 0$ (see Parlett and Poole [6]).

Theorem. If ρ_k *converges, then either*

(i) ρ_k *converges to* λ_i *for some* i *and* $V^{(k)}$ *converges to* x_i *provided* $V^{(0)}$ *is not deficient in* x_i,

or (ii) (a) ρ_k *converges to* $\rho' = \lambda_i$ *and* $V^{(0)}$ *is deficient in* x_i,

(b) ρ_k *converges to* ρ' *which is not an eigenvalue.*

In either case,

$$\rho' = \sum_{i=1}^{m} \alpha_{j_i} \beta_{j_i} \gamma_{j_i} \lambda_{j_i} / \sum_{i=1}^{m} \alpha_{j_i} \beta_{j_i} \gamma_{j_i}$$

where $|\lambda_{j_1} - \rho'| = |\lambda_{j_2} - \rho'| = \cdots = |\lambda_{j_m} - \rho'|$. *In other words,* $\lambda_{j_1}, \ldots, \lambda_{j_m}$ *must be co-cyclic with center* ρ'. *In this case,* $V^{(k)}$ *may not converge. If* v *is a limit vector of the Rayleigh sequence* $\{V^{(k)} | k = 1,2,\ldots\}$, *then* $v = \sum_{i=1}^{m} \alpha_{j_i} x_{j_i}$.

Proof. From the definition of RQI, $V^{(k+1)} = \tau_k (A - \rho_k)^{-1} V^{(k)}$ so we have $\alpha_i^{(k+1)} = \dfrac{\tau_k}{\lambda_i - \rho_k} \alpha_i^{(k)}$ (as in (2.2.1)). If $\alpha_i^{(0)} \neq 0$, then

$$\left| \frac{\alpha_j^{(k+1)}}{\alpha_i^{(k+1)}} \right| = \left| \frac{\lambda_i - \rho_k}{\lambda_j - \rho_k} \right| \left| \frac{\alpha_j^{(k)}}{\alpha_i^{(i)}} \right| .$$

Hence in the case ρ_k converges to λ_i, a careful examination of the above formula reveals that $|\alpha_j^{(k)} / \alpha_i^{(k)}| \to 0$ as $k \to \infty$ for $j \neq i$. Therefore $V^{(k)}$ converges to x_i. In Case (ii), let $v = \sum_{i=1}^{m} \alpha_{j_i} x_{j_i}$, $m \leq n$, $\alpha_{j_i} \neq 0$ be a limit vector. Let $\omega = \min_{1 \leq i \leq m} \{|\lambda_{j_i} - \rho'|\}$. Then $\omega \neq 0$ because otherwise $\rho_k \to \lambda_{j_i}$ with $\alpha_{j_i} \neq 0$, which is the first case, a contradiction.

Let $\omega = |\lambda_{j_i} - \rho'|$ for some i. If there exists ℓ such that $|\lambda_{j_\ell} - \rho'| - \omega = \delta > 0$, then there exists K such that for $k \geq K$

$||\lambda_{j_i}-\rho_k|-\omega| < \frac{\delta}{3}$. Then

$$\left|\frac{\alpha_{j_\ell}^{(k+1)}}{\alpha_{j_i}^{(k+1)}}\right| = \left|\frac{\lambda_{j_i}-\rho_k}{\lambda_{j_\ell}-\rho_k}\right|\left|\frac{\alpha_{j_\ell}^{(k)}}{\alpha_{j_i}^{(k)}}\right| \le \frac{\omega+\delta/3}{\omega+2\delta/3}\left|\frac{\alpha_{j_\ell}^{(k)}}{\alpha_{j_i}^{(k)}}\right| .$$

Thus as $k \to \infty$, $\alpha_{j_\ell}^{(k)} \to 0$ implies $\alpha_{j_\ell} = 0$, a contradiction.

Therefore, $|\lambda_{j_1}-\rho'| = |\lambda_{j_2}-\rho'| = \cdots = |\lambda_{j_m}-\rho'| = \omega$,

$$\rho' = \rho(v) = \sum_{i=1}^{m} \alpha_{j_i}\beta_{j_i}\gamma_{j_i}\lambda_{j_i} / \sum_{i=1}^{m} \alpha_{j_i}\beta_{j_i}\gamma_{j_i} \ . \qquad \square$$

Definition. $\Gamma = \{j_1, j_2, \ldots, j_m\}$, Δ be the subspace spanned by $\{x_j, j \in \Gamma\}$ and $\lambda_j = \rho' + \omega e^{i\theta_j}$.

Corollary. *In the real case*, $\rho' = (\lambda_p+\lambda_q)/2$ *and the only limit vectors are* $(x_p \pm x_q)/\sqrt{2}$.

Proof. ρ' can be equidistant from only two distinct real eigenvalues, hence $\rho' = (\lambda_p+\lambda_q)/2$. The results from the previous section tell us that the limit vectors must be the internal or external bisectors of x_p and x_q. λ_p and λ_q could be multiple, but x_p and x_q are the unique eigendirections in the plane defined by the projection of $v^{(0)}$ onto Δ.

§2.8 Instability of Case (ii)

Let z be any unit vector in the invariant subspace Δ defined above. Let $\phi_i = \alpha_i\beta_i\gamma_i$. Thus with the notion of Section 7:

$$z = \sum_{j=1}^{n} \alpha_j x_j \ , \quad z^* = \sum_{j=1}^{n} \beta_j y_j^*$$

with $$1 = \sum\alpha_j\beta_j\gamma_j = \sum\phi_j$$

and $$\rho(z) = [\rho'\sum\phi_j + \omega\sum e^{i\theta_j}\phi_j]/\sum\phi_j \ .$$

Differentiating, we find

$$\frac{\partial\rho(z)}{\partial\phi_j} = [\rho' + \omega e^{i\theta_j} - \rho(z)]$$

$$\left.\frac{\partial\rho}{\partial\phi_j}\right|_v = \omega e^{i\theta_j} \neq 0 \quad \text{for each} \quad j \ .$$

So an increase in ϕ_j pushes the Rayleigh Quotient from ρ' towards $\lambda_j = \rho' + \omega e^{i\theta_j}$. Almost all perturbation in Δ of a limit point v generates Rayleigh sequences which converge toward an eigenvector.

CHAPTER THREE

The Completely Degenerate Case

§3.1 An Overview

We now focus our attention on the completely degenerate case. We say that a matrix C is completely degenerate if C is similar to a single Jordan block, i.e., there exist X invertible such that

$$X^{-1}CX = J = \begin{pmatrix} \alpha & 1 & 0 & & \bigcirc \\ & \alpha & 1 & 0 & \\ & & \alpha & \ddots & \\ & & & \ddots & 0 \\ & \bigcirc & & \ddots & 1 \\ & & & & \alpha \end{pmatrix} .$$

Because of the translation invariant properties of the RQI, it is equivalent to consider $C - \alpha I$, i.e., a matrix C with zero as its only eigenvalue and has only one eigenvector. Note that it is not sufficient to consider the canonical Jordan block

$$\begin{pmatrix} 0 & 1 & 0 & & \bigcirc \\ & 0 & 1 & & \\ & & 0 & \ddots & \\ & & & \ddots & \\ \bigcirc & & & \ddots & 1 \\ & & & & 0 \end{pmatrix} = N \quad \text{(N for nilpotent)}$$

because not every completely degenerate matrix C with zero eigenvalues is unitarily similar to N, and we know RQI is only unitarily invariant, not invariant under similarity transformation.

We adopt the following notation throughout this chapter. Let x_i be the eigenvector of C of i^{th} grade, i.e., $Cx_1 = 0$, $\|x_1\| = 1$, $Cx_{i+1} = x_i$ for $1 \leq i \leq n-1$. Thus $C^i x_i = 0$, $C^{i-1}x_i \neq 0$. With standard calculation, we have x_1 orthogonal to x_i, $i \neq 1$. (Notice

x_i, $i \neq 1$ may be of length other than unity).

If $v = \sum\alpha_i x_i$, then $\rho(v) = \sum\bar{\alpha}_i\alpha_{i+1}/(\sum|\alpha_i|^2)^{1/2}$ provided $\{x_i\}$ forms an orthonormal set (i.e,, the original matrix is unitarily similar to N), and $\rho(v)$ is more complicated if $\{x_i\}$ is not orthonormal:

$$\rho(v) = \frac{v^*Cv}{\|v\|}$$

$$= \frac{a^*X^*CXa}{\|v\|} \quad \text{where} \quad v = Xa, \quad X = (x_1,x_2,\ldots,x_n), \quad a = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}$$

$$= a^*Ga/\|v\| \quad \text{where} \quad G = (g_{ij}), \quad g_{ij} = x_i^*x_{j-1},$$

$$x_0 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$= (\sum_i\bar{\alpha}_i \sum_j g_{ij}\alpha_j)/\|v\|$$

$$(3.1.1) \qquad = (\sum_i \sum_j g_{ij}\bar{\alpha}_i\alpha_j)/\|v\| .$$

Note that in the above expression, $g_{1j} = 0$ for $j \neq 2$ and $g_{12} = 1$ because $x_1^*x_j = 0$ by choice and $x_1^*x_1 = 1$.

Now, we want to study what one step of RQI does to the vector v. Let $v' = \sum\alpha_i'x_i$ be the resultant vector. Then if $C - \rho(v)$ is nonsingular,

$$w' = (C - \rho(v))^{-1}v$$

$$v' = \frac{w'}{\|w'\|} .$$

$(C-\rho)^{-1}$ expressed in the basis of generalized eigenvectors would be

$$(C-\rho)^{-1} = \begin{pmatrix} -\rho & 1 & & \bigcirc \\ & -\rho & 1 & \\ & & -\rho & \ddots \\ \bigcirc & & \ddots & 1 \\ & & & -\rho \end{pmatrix}^{-1} = \begin{pmatrix} -\frac{1}{\rho} & -\frac{1}{\rho^2} & \cdots & -\frac{1}{\rho^n} \\ & -\frac{1}{\rho} & \cdots & -\frac{1}{\rho^{n-1}} \\ & & & \vdots \\ \bigcirc & & \ddots & -\frac{1}{\rho} \end{pmatrix}$$

so

$$w' = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} = (C-\rho)^{-1}v = \begin{pmatrix} -\frac{1}{\rho}\alpha_1 - \frac{1}{\rho^2}\alpha_2 - \cdots - \frac{1}{\rho^n}\alpha_n \\ -\frac{1}{\rho}\alpha_2 - \cdots - \frac{1}{\rho^{n-1}}\alpha_n \\ \bigcirc \quad \ddots \quad \vdots \\ -\frac{1}{\rho}\alpha_n \end{pmatrix}$$

$$(3.1.2) \qquad = \begin{pmatrix} -\frac{1}{\rho}(\alpha_1+\beta_2) \\ \vdots \\ -\frac{1}{\rho}(\alpha_i+\beta_{i+1}) \\ \vdots \\ -\frac{1}{\rho}(\alpha_{n-1}+\beta_n) \\ -\frac{1}{\rho}\alpha_n \end{pmatrix}$$

$$= -\frac{1}{\rho}(v+Nw') \ .$$

Studying these equations carefully reveals the main difficulty: $\beta_1 = (-1/\rho)\alpha_1 - (1/\rho^2)\alpha_2 - \cdots - (1/\rho^n)\alpha_n$. When v is close to x_1, ρ is close to zero. Therefore, $1/\rho^n$ is very large and those "arbitrary" small coefficients of x_i, $i \neq 1$ cannot be ignored. Also, there may be some unfortunate cancellations that make β_1 extremely small as compared to β_2.

Wilkinson [9] studied the effect of Inverse iteration method as applied to ill-conditioned matrices and degenerate matrices and came to the conclusion that you should not do inverse iteration more than once if you have a very good approximation to the eigenvalues. Now RQI is nothing but inverse iteration with a shift which is "optimal" in a certain sense, and the shift tends to the eigenvalue. So we can expect the same irregular behaviour near an eigenvector here. We shall demonstrate the complexity of the local picture through an example in the next section.

The local behaviour and convergence rate of RQI for a degenerate matrix was investigated by Kiho Lee Kim [2] in 1970. He derived the equation $V^{(i+1)} = r + H(V^{(i)} - r) + g(V^{(i)})$ where H is the Jacobian of RQI at r, the fixed point of the iteration, and g is a function that satisfies $\|g(x)\| \leq M\|x-r\|^2$ for some norm $\|\cdot\|$ and some constant M. Kim showed that the convergence rate should be the spectral radius of H. But his conclusion depends on the assumption that the iterative function has second derivative in U, an open set for which r belongs to. In the example of the next section, we can see that the RQI function may have second derivatives in <u>some sector</u> of a neighborhood of the eigenvector (and hence afford a Taylor expansion there), but the RQI function is not even continuous at the eigenvector. Our Lemma below will illustrate this point.

<u>Lemma</u>. <u>Let</u> f <u>be the</u> RQI <u>mapping</u>, i.e., $f(V^{(k)}) = V^{(k+1)}$. <u>Let</u> x_1 <u>be the eigenvector</u>. <u>Then for every</u> $\delta_0 > 0$, <u>there exists</u> v, $\|v\| = 1$, <u>such that</u> $\|x_1 - v\| < \delta_0$ <u>and</u> $f(v)$ <u>is orthogonal to</u> x_1.

<u>Proof</u>. By invariant properties of RQI, it is equivalent to consider the matrix C which is similar to N. The set of generalized eigenvectors has the property that x_1 is orthogonal to x_i, $i \neq 1$.

From (3.1.2) we know if $v = \sum\alpha_i x_i$. Then

$$(3.1.3) \qquad \beta_1 = -(1/\rho)\alpha_1 - (1/\rho^2)\alpha_2 - \cdots - (1/\rho^n)\alpha_n .$$

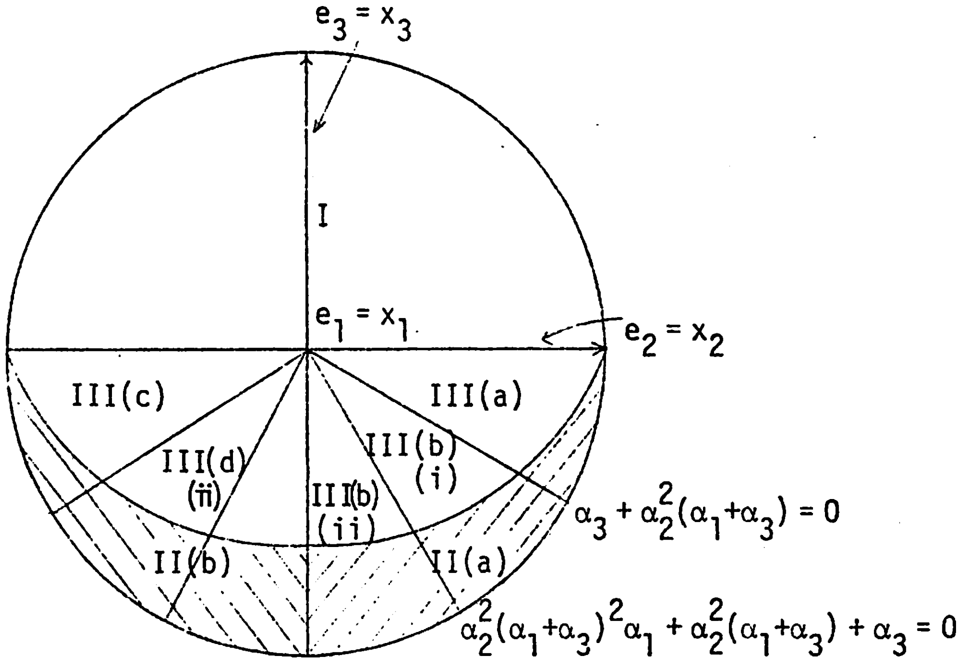
Let $\mu := \max\{\|x_i\|, i = 1,\dots,n\}$ (see the redefinition of x_i at the beginning of the section). Let $\delta = \min(\delta_0/10n, (10n)^{-5})$. Pick v such that $\alpha_i = (\frac{\delta}{\mu n})^{i-1}$, $i = 2,\dots,n-1$ and $0 \geq \alpha_n \geq -n\delta^{n-2}$ and $\alpha_1 \geq 0$ such that $\|v\| = 1$. First notice $\|v-x_1\| < \delta_0$ because $\|v-x_1\| \leq 2\sum_{i=2}^{n}|\alpha_i| \leq 2\sum_{i=1}^{n-2}(\frac{\delta}{\mu n})^i + 2n\delta^{n-2} < \delta_0$ by choice of δ. Then $\rho(v) = \alpha_1\frac{\delta}{\mu n} + O(\delta^2)$ by (3.1.1). If $\alpha_n = 0$ then $\beta_1 < 0$ because each term is positive in (3.1.3). If $\alpha_n = -n\delta^{n-2}$ then $\beta_1 > 0$ because each term in (3.1.3) is $O(\frac{\mu n}{\delta})$ except the last term which is $O(\mu^n n^n/\delta^2)$. Therefore, the last term dominates and $\beta_1 > 0$.

Since v' depends continuously on v when the range of $\rho(v)$ is bounded away from zero, we must have a t, $0 \leq t \leq 1$, such that when $\alpha_n = -n\delta^{n-2}t$, then $\beta_1 = 0$. This implies $f(v) = \sum_{i=2}^{n}\alpha_i' x_i$ that lies in the hyperplane orthogonal to x_1. □

§3.2 A 3 × 3 Example - Sectorial Behaviour Near the Eigenvector

From the lemma of the last section, it is obvious that one cannot isolate a small open neighborhood of the eigenvector to study the local behaviour of RQI, because no matter how small you take the neighborhood to be, there are points and regions around the points that can throw you out of that neighborhood. Thus, we divide the region about the eigenvector into sectors of attraction and repulsion. We shall give an example of the sectorial behaviour of the RQI.

<u>Example</u>. Let $C = N = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$. Then for $v = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3$, $\rho(v) = (\alpha_1\alpha_2 + \alpha_2\alpha_3)/(\alpha_1^2 + \alpha_2^2 + \alpha_3^2)^{1/2}$, with α_i's real, so if $\|v\| = 1$, $\rho(v) = \alpha_1\alpha_2 + \alpha_2\alpha_3$. We first show that in this case, the RQI converges for all starting vectors, and in the course of doing so, demonstrate the sectorial behaviour of RQI. (See Figure 3.2.1.)


$e_3 = x_3$
I
$e_1 = x_1$
$e_2 = x_2$
III(c)
III(a)
III(b)
(i)
III(d)
(ii)
III(b)
(ii)
II(b)
II(a)
$\alpha_3 + \alpha_2^2(\alpha_1+\alpha_3) = 0$
$\alpha_2^2(\alpha_1+\alpha_3)^2\alpha_1 + \alpha_2^2(\alpha_1+\alpha_3) + \alpha_3 = 0$

e_1 is coming out of the plane, i.e., $\alpha_1 \geq 0$ in the figure

Hemisphere containing x_1

Figure 3.2.1

Let $w' = (A-\rho(v))^{-1}v = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ where $v = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3$. Then by (3.1.2)

$$\begin{aligned} \beta_3 &= -\alpha_3/\alpha_2 b \\ \beta_2 &= -1/b - \alpha_3/(\alpha_2^2 b^2) \qquad (3.2.1) \\ \beta_1 &= -\alpha_1/\alpha_2 b - 1/\alpha_2 b^2 - \alpha_3/\alpha_2^3 b^3 \end{aligned}$$

where $b = \alpha_1 + \alpha_3$ and $\rho(v) = \alpha_2 b$. We can normalize, without loss of generality, $\alpha_1 \geq 0$ for all our vectors. We shall separate the region in Figure 3.2.1 into the following:

I. $\alpha_3 \geq 0$. This is an invariant region under RQI because in (3.2.1), $\text{sign}(\beta_1) = \text{sign}(\beta_2)$ can be deduced from the fact that each term in the expression for β_1 and β_2 has the same sign as α_2. The convergence is monotonic in this region in the following sense:

$$\left|\frac{\beta_1}{\beta_2}\right| = \left|\frac{\alpha_1}{\alpha_2}\right| \left|\frac{-1/\alpha_2 b - 1/\alpha_1\alpha_2 b^2 - \alpha_3/\alpha_1\alpha_2^2 b^3}{-1/\alpha_2 b - \alpha_3/\alpha_2^3 b^3}\right|$$

$$= \left|\frac{\alpha_1}{\alpha_2}\right| \left|\frac{|1/\alpha_2 b| + |1/\alpha_1\alpha_2 b^2| + |\alpha_3/\alpha_1\alpha_2^3 b^3|}{|1/\alpha_2 b| + |\alpha_3/\alpha_2^3 b^3|}\right.$$

(because each term has the same sign as α_2)

$$> \left|\frac{\alpha_1}{\alpha_2}\right| \quad \text{(because } |\alpha_1| < 1\text{, in the last term of numerator)}$$

and it is obvious that the factor does not tend to one.

$$\left|\frac{\beta_1}{\beta_3}\right| > \left|\frac{\alpha_1}{\alpha_3}\right|$$

by similar arguments. Thus the vector iteration converges to x_1, the eigenvector.

Now we consider the region $\alpha_3 < 0$.

II(a) $|\alpha_3| \geq |\alpha_1|$, $\alpha_2 > 0$. Then $\rho(v) < 0$, $b = \alpha_1 + \alpha_2 < 0$ and (3.2.1) gives $\beta_3 < 0$. $\beta_1 < 0$ because the second term dominates the first term, and the third term has the same sign as the second term, which is negative. We get thrown into region I and get convergence.

II(b) $|\alpha_3| \geq |\alpha_1|$, $\alpha_2 < 0$. Then $\rho(v) > 0$, $b = \alpha_1 + \alpha_3 < 0$. (3.2.1) gives $\beta_3 > 0$, $\beta_1 > 0$ for the same reason as above. We get thrown in region I and get convergence.

Now let $|\alpha_1| > |\alpha_3|$. Then $b = \alpha_1 + \alpha_3 > 0$.

III(a) $\alpha_2 > 0$, $|\alpha_2^2 b| > |\alpha_3|$. Then $\beta_3 > 0$, $\beta_2 < 0$. If $\beta_1 \geq 0$, we are in region I and converge. If $\beta_1 < 0$, a standard calculation shows

$$\left|\frac{\beta_3}{\beta_2}\right| = \left|\frac{\alpha_3/\alpha_2 b}{1/b + \alpha_3/\alpha_2^2 b^2}\right| = \left|\frac{\alpha_3}{\alpha_2}\right| \left|\frac{1/\alpha_2 b}{1/\alpha_2 b + \alpha_3/\alpha_2^3 b^2}\right|$$

$$> \left|\frac{\alpha_3}{\alpha_2}\right| \quad \text{because the two terms in the denominator differ in sign and are dominated by the first.}$$

So at each step, the ratio $\left|\frac{\alpha_3}{\alpha_2}\right|$ increases and we stay in III(a) until $|\alpha_2^2 b| \leq |\alpha_3|$.

III(b) $\alpha_2 > 0$, $|\alpha_2^2 b| \leq |\alpha_3|$.

(i) $\alpha_2^2 b^2 \alpha_1 + \alpha_2^2 b + \alpha_3 > 0$. Then (3.2.1) gives $\beta_1 < 0$, $\beta_2 > 0$, $\beta_3 > 0$ and a standard calculation shows that $\left|\frac{\beta_3}{\beta_1}\right| > \left|\frac{\alpha_3}{\alpha_1}\right|$, $\left|\frac{\beta_3}{\beta_2}\right| > \left|\frac{\alpha_3}{\alpha_2}\right|$. If $|\beta_3| \geq |\beta_1|$, we are in region II(b) and thus have convergence. If $|\beta_3| < |\beta_1|$, we have III(d) which we consider later.

(ii) $\alpha_2^2 b^2 \alpha_1 + \alpha_2^2 b + \alpha_3 \leq 0$. Then (3.2.1) gives $\beta_1 > 0$, $\beta_2 > 0$, $\beta_3 > 0$ and we are in region I, thus have convergence.

III(c) $\alpha_2 < 0$, $|\alpha_2^2 b| > |\alpha_3|$. This is a mirror image of region III(a) and has the same properties. The vector in this region is either thrown into region I or region III(d) below.

III(d) $\alpha_2 < 0$, $|\alpha_2^2 b| \leq |\alpha_3|$. This is a mirror image of region III(b) and has the same properties.

(i) $\alpha_2^2 b^2 \alpha_1 + \alpha_2^2 b + \alpha_3 > 0$. Then (3.2.1) gives $\beta_1 < 0$, $\beta_2 > 0$, $\beta_3 > 0$ and calculation shows that $\left|\frac{\beta_3}{\beta_1}\right| > \left|\frac{\alpha_3}{\alpha_1}\right|$, $\left|\frac{\beta_3}{\beta_2}\right| > \left|\frac{\alpha_3}{\alpha_2}\right|$. If $|\beta_3| \geq |\beta_1|$, we are in region II(a) and thus have convergence. If

$|\beta_3| < |\beta_1|$, we shall be back at region III(b) (hence a possible cycle).

(ii) $\alpha_2^2 b^2 \alpha_1 + \alpha_2^2 b + \alpha_3 \leq 0$. Then (3.2.1) gives $\beta_1 > 0$, $\beta_2 > 0$, $\beta_3 > 0$ and we are in region I and thus have convergence.

From all of the above regions whose union is the hemisphere, we either have convergence or are thrown into convergence regions ultimately except that there is a possible cycle to go from III(b)(i) to III(d)(i) back and forth. This is fortunately not an infinite loop because at each step, the ratio $|\frac{\alpha_3}{\alpha_2}|$, $|\frac{\alpha_3}{\alpha_1}|$ both increase and thus force $\alpha_2^2 b^2 \alpha_1 + \alpha_2^2 b + \alpha_3 < 0$ ultimately and we shall be thrown into III(b)(ii) or III(d)(ii), then region I and have convergence.

We have omitted the case $\alpha_2 = 0$ because this would make $\rho(v) = 0$ and by definition of RQI, we have convergence in one step.

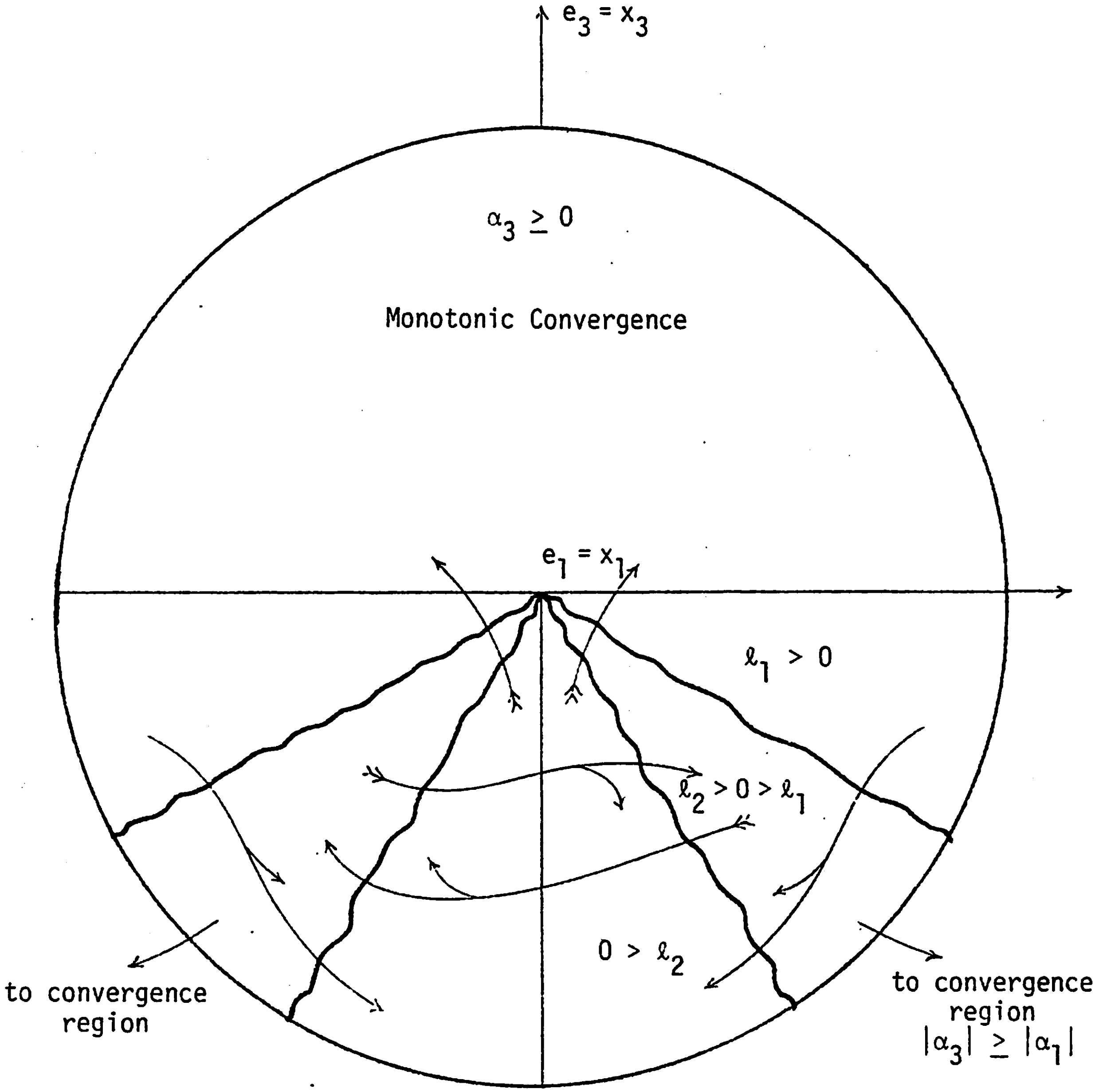
Now that we have shown global convergence of the example, it would seem instructive to draw the graph of a neighborhood of x_1 to illustrate the different regions and their possible route to convergence (see Figure 3.2.2).

§3.3 Behaviour of ρ_k as $k \to \infty$

The main objective in this section is to show if ρ_k converges, then ρ_k converges to λ, the single point in the spectrum of the operator. With the normalization as in §3.1, we have $\lambda = 0$.

Before we proceed with the main theorem, it may be illuminating to prove the following lemma that presents some analysis that will recur throughout this chapter.

Lemma. If the shift ρ_k is constant, i.e., $\rho_k = \rho$ a fixed number, then $v^{(k)} = \tau_k(C-\rho)^{-k}v^{(0)}$ converges to x_1.



where

$$\ell_1 \equiv \alpha_3 + \alpha_2^2(\alpha_1+\alpha_3)$$
$$\ell_2 \equiv \alpha_2^2(\alpha_1+\alpha_3)^2\alpha_1 + \ell_1$$

Figure 3.2.2

Proof: In this case, RQI is reduced to inverse iteration with the shift ρ. The vector sequence converges because of the special form of $(C-\rho)^{-k}$. So we consider $-\rho(C-\rho)^{-1}$ in the basis of generalized eigenvectors:

$$-(\tfrac{1}{\rho})^{-1}\begin{pmatrix} -\rho & 1 & & \bigcirc \\ & -\rho & 1 & \\ & & -\rho & \ddots \\ \bigcirc & & \ddots & 1 \\ & & & -\rho \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -\frac{1}{\rho} & & \bigcirc \\ & 1 & -\frac{1}{\rho} & \\ & & \ddots & \ddots \\ \bigcirc & & \ddots & -\frac{1}{\rho} \\ & & & 1 \end{pmatrix}^{-1} = (I-\tfrac{1}{\rho}N)^{-1}$$

where $N = \begin{pmatrix} 0 & 1_1 & \bigcirc \\ & \ddots & \ddots \\ \bigcirc & \ddots & 1 \\ & & 0 \end{pmatrix}$.

Let $\xi = 1/\rho$. Then $v^{(k)} = \tau_k'(I-\xi N)^{-k}v^{(0)}$. But

$$\begin{aligned}(I-\xi N)^{-k} &= [(I-\xi N)^{-1}]^k \\ &= (I+\xi N+\xi^2N^2+\cdots+\xi^{n-1}N^{n-1})^k \\ &= I+\tbinom{k}{1}\xi N+(\tbinom{k}{2}+\tbinom{k}{1})\xi^2N^2 \\ &\quad +(\tbinom{k}{3}+2\tbinom{k}{2}+\tbinom{k}{1})\xi^3N^3 \\ &\quad +(\tbinom{k}{4}+\cdots)\xi^4N^4 \\ &\quad +\cdots+(\tbinom{k}{n-1}+\cdots)\xi^{n-1}N^{n-1} \qquad \text{for} \quad k \geq n \\ &= \sum_{j=0}^{n-1}\tbinom{k+j-1}{j}\xi^jN^j\end{aligned}$$

So $(I-\xi N)$ is a unit upper triangular Toeplitz matrix (see Marcus and Minc [3]). If we write ${}_{k+j-1}C_j$ to denote $\binom{k+j-1}{j}$, then it is clear that ${}_{k+j-1}C_j$ is of $O(k^j)$ as $k \to \infty$.

$$(I-\xi N)^{-k} = \begin{pmatrix} 1 & \xi\,{}_kC_1 & \xi^2\,{}_{k+1}C_2 & \cdots & \xi^{k-1}\,{}_{k+n-2}C_{n-1} \\ & 1 & \xi\,{}_kC_1 & & \xi^{k-2}\,{}_{k+n-3}C_{n-2} \\ & & 1 & & \vdots \\ & \bigcirc & & \ddots & \xi\,{}_kC_1 \\ & & & & 1 \end{pmatrix}$$

Assume $V^{(0)} = (\eta_1,\eta_2,\ldots,\eta_n)^T$ (in the basis $x_1,\ldots,x_n$). Then $V^{(k)} = \tau_k'(I-\xi N)^{-k}V^{(0)}$ is a linear combination of columns of $(I-\xi N)^{-k}$ with coefficients $\eta_1,\eta_2,\ldots,\eta_n$. Let η_ℓ be the last nonzero element in $V^{(0)}$. Then as $k \to \infty$, the fact that ${}_{k+j-1}C_j$ is of order k^j implies the components of the ℓ^{th} column vector would dominate in the expression for $V^{(k)}$ written as a linear combination of column vectors. So when k is large

$$V^{(k)} \to \begin{pmatrix} \xi^{\ell-1}\,{}_{k+\ell-2}C_{\ell-1} \\ \xi^{\ell-2}\,{}_{k+\ell-3}C_{\ell-2} \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

But here again $|{}_{k+j-2}C_{j-1}/{}_{k+j-1}C_j| \to 0$ as $k \to \infty$. Therefore $V^{(k)} \to e_1 = x_1$ as $k \to \infty$. □

The same conclusion is still valid if $\rho_k = \rho(V^{(k)})$ converges to a scalar other than the eigenvalue:

Theorem. If ρ_k converges to $\omega = re^{i\theta} \neq \lambda$, then $V^{(k)}$ converges to x_1.

Proof. Recall we have the normalization that $\lambda = 0$. The assumption that $\rho_k \to \omega$ implies that there exists K such that for $k \geq K$, then $|r-r_k| < \varepsilon_1$ and $|\theta-\theta_k| < \varepsilon_2$ where $\rho_k = r_k e^{i\theta_k}$, $\varepsilon_1 = \frac{|r|}{2} > 0$ and $\varepsilon_2 < \frac{\pi}{4n}$. Let $V^{(K)} = (\eta_1,\eta_2,\ldots,\eta_n)$ (in the basis of $x_1,\ldots,x_n$). Then

$$V^{(K+k+1)} = \tau_k(C-\rho_{K+k})^{-1}(C-\rho_{K+k-1})^{-1}\cdots(C-\rho_{K+1})^{-1}V^{(K)} \quad .$$

Let $\xi_k = 1/\rho_k$. Then

$$\begin{aligned} V^{(K+k+1)} &= \tau'_k(I-\xi_{K+k}N)^{-1}\cdots(I-\xi_{K+1})^{-1}V^{(k)} \\ &= \tau'_k\big(I + d_1(k)N + d_2(k)N^2 + \cdots + d_{n-1}(k)N^{n-1}\big)V^{(K)} \end{aligned}$$

where $d_i(k)$ = sum of ${}_{k+i-1}C_i$ terms of the form $(\xi_{j_1}\xi_{j_2}\cdots\xi_{j_i})$.

Lemma. *There exists* M_i, m_i *independent of* k *such that* $M_i({}_{k+i-1}C_i) \geq d_i(k) \geq m_i({}_{k+i-1}C_i)$.

Proof. Since $\frac{1}{r-\varepsilon_1} > |\xi_j| > \frac{1}{r+\varepsilon_1}$

$$\left(\frac{1}{r-\varepsilon_1}\right)^i > |\xi_{j_1}\cdots\xi_{j_i}| > \left(\frac{1}{r+\varepsilon_1}\right)^i \quad .$$

So we can take $M_i = \left(\frac{1}{r-\varepsilon_1}\right)^i$. As for m_1 we have to make use of the condition

$$|\theta-\theta_k| < \varepsilon_2 < \frac{\pi}{4n} \; .$$

If $\xi_{j_1}\xi_{j_2}\cdots\xi_{j_i} = se^{i\theta'}$ then

$$|-\theta-\theta'| < \frac{i\pi}{4n} \leq \frac{\pi}{4} \; .$$

The length of the orthogonal projection of $\xi_{j_1}\cdots\xi_{j_i}$ on the line $-\theta$ in the complex plane is at least $s\cos\pi/4 = s/\sqrt{2}$, and the orientation is the same for each of the ${}_{k+i-1}C_i$ of them. Hence m_i can be taken as $(\frac{1}{r+\varepsilon_1})^i/\sqrt{2}$. Here ends the proof of the lemma. □

The rest of the proof of the theorem is similar to that of the case where ρ_k is constant. $V^{(K+k)}$ is a linear combination of the columns of

$$(I-\xi N)^{-k} = \begin{pmatrix} 1 & d_1(k) & d_2(k) & \cdots & d_{n-1}(k) \\ & 1 & d_1(k) & \cdots & d_{n-2}(k) \\ & & 1 & & \vdots \\ & \bigcirc & & \ddots & d_1(k) \\ & & & & 1 \end{pmatrix}.$$

Let η_ℓ be the last nonzero component of $V^{(K)}$. Then as $k \to \infty$, the components of the ℓ^{th} column vector would dominate, so

$$V^{(k)} \to \begin{pmatrix} d_{\ell-1}(k) \\ d_{\ell-2}(k) \\ \vdots \\ d_1(k) \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tau''_k \to \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Therefore, $V^{(k)} \to e_1 = x_1$. □

Corollary. *Let* C *be a completely degenerate matrix with* λ *as its only eigenvalue, and* ρ_k *be the Rayleigh Quotients sequence*. *If* ρ_k *converges, then* ρ_k *converges to* λ.

Proof. Directly from the last theorem, if $\rho_k \to \omega \neq \lambda$, then $V^{(k)} \to x_1$, but ρ as a function from C^n to C is continuous. Therefore $\rho(V^{(k)}) \to \rho(x_1) = \lambda$, a contradiction. Therefore, ρ_k converges to the eigenvalue if the sequence converges. □

§3.4 Behaviour of $V^{(k)}$ When ρ_k Converges

In this section, we want to show that if ρ_k converges and x_2 is not a limit vector of the Rayleigh sequence, then the vector iteration converges to x_1.

Without loss of generality we assume that ρ_k converges to $\lambda = 0$ and unless otherwise specified, a vector is expressed in the basis of generalized eigenvectors.

Define $\Gamma = \{v|\ v$ is a unit limit vector of the Rayleigh sequence$\}$. From Γ we pick out a vector whose last nonzero element has a maximal index, ℓ say. Thus $u = (\eta_1,\eta_2,\ldots,\eta_n)^T \in \Gamma$ satisfies

(1) $\eta_\ell = \omega \neq 0$ and $\eta_j = 0$ for $j > \ell$,

(2) $u' = (\eta_1',\eta_2',\ldots,\eta_n') \in \Gamma$ implies $\eta_j' = 0$ for $j > \ell$.

Now u being a limit vector means that there exists a set $K_0 \subseteq \{1,2,\ldots\}$ such that $V^{(k)}$, $k \in K_0$ converges to u. We shall first investigate what $V^{(k-1)}$, $k \in K_0$, looks like when ρ_{k-1} is sufficiently close to zero.

Since $\rho_k \to 0$, there exists a number N_0 such than whenever

$$k \geq N_0\ , \qquad |\rho_k| < \delta \tag{3.4.1}$$

where δ is so small that

$$(3.4.2) \qquad |\omega/\delta| > 3\kappa$$

where $\kappa = \sup_i\{|\xi_i|\,|\, w=(\xi_1,\ldots,\xi_n),\ w \text{ is a unit vector}\}$. (Recall x_i, $i > 1$ may be of length other than unity because we normalize C in the convenient form which is a Jordan block).

It is more convenient to change the basic RQI equation $(C-\rho_k)V^{(k+1)} = \tau_k V^{(k)}$ into the following form:

$$(I-\xi_k N)V^{(k+1)} = \xi_k\tau_k V^{(k)}$$

where $\xi_k = 1/\rho_k$. In detail, we have, for $V^{(k)} = (\alpha_1^{(k)},\ldots,\alpha_n^{(k)})^T$,

$$(3.4.3) \qquad \begin{aligned} \alpha_n^{(k)} &= \xi_{k-1}\tau_{k-1}\alpha_n^{(k-1)} \\ \alpha_i^{(k)} &= \xi_{k-1}\tau_{k-1}\alpha_i^{(k-1)} + \xi_{k-1}\alpha_{i+1}^{(k)}, \quad 1 \le i \le n-1 . \end{aligned}$$

When $k \in K_0$ is large enough, $\alpha_\ell^{(k)} \doteq \omega$. Studying these relations leads us to a key result.

Lemma 1. Assume $\rho_k \to 0$. Then

(1) For k large enough, there exists constants $d_2 \ge d_1 \ge 0$ such that $d_1 \le \tau_{k-1} \le d_2$, $k \in K_0$.

(2) $\alpha_\ell^{(k-1)} \to 0$ as $k \to \infty$ for $k \in K_0$ (K_0 is defined at the beginning of the section).

(3) There exists constants h_1, h_2 such that $0 < h_1 \le |\alpha_{\ell-1}^{(k)}| \le h_2$ for k large enough, $k \in K_0$.

Recall that for normal matrices $\tau_{k-1} \ge d_1 > 0$ implies that $V^{(k)}$ does not converge.

Proof. (1) We want to show that τ_k is bounded away from zero and infinity. We know there exists a number N_1 such that when $k \geq N_1$, $k \in K_0$, then $|\alpha_\ell^{(k)}-\omega| < \delta$. So for $k \geq N_2 = \max(N_0,N_1)$ where N_0 is given in (3.4.1)

$$\begin{aligned} |\tau_{k-1}\alpha_{\ell-1}^{(k-1)} + \alpha_\ell^{(k)}| &\geq |\tau_{k-1}\alpha_{\ell-1}^{(k-1)} + \omega| - \delta \\ &\geq |\omega| - (|\tau_{k-1}\alpha_{\ell-1}^{(k-1)}| + \delta) . \end{aligned}$$

From (3.4.3)

$$\begin{aligned} |\alpha_{\ell-1}^{(k)}| &= |\xi_{k-1}||\tau_{k-1}\alpha_{\ell-1}^{(k-1)} + \alpha_\ell^{(k)}| \\ &\geq |\xi_{k-1}|(|\omega| - \tau_{k-1}\alpha_{\ell-1}^{(k-1)} - \delta) . \end{aligned}$$

If there exists k such that $|\tau_{k-1}| \leq \delta$, then

$$\begin{aligned} |\alpha_{\ell-1}^{(k)}| &\geq |\xi_{k-1}|(|\omega|-2\kappa\delta) \\ &\geq |\xi_{k-1}|(3\kappa\delta-2\kappa\delta) \quad (\text{because } |\omega| \geq 3\kappa\delta) \\ &> \kappa \quad (\text{since } |\xi_{k-1}| > 1/\delta \text{ and } \kappa \geq 1) , \end{aligned}$$

a contradiction to the definition of κ in (3.4.2). Therefore, there exists $d_1 \geq \delta > 0$ such that $\tau_{k-1} > d_1$ for $k \geq N_2$. Now from the defining equation of RQI, we have $(C-\rho_{k-1})V^{(k)} = \tau_{k-1}V^{(k-1)}$, and thus $|\tau_{k-1}| \leq \|C\| + |\rho_k| \leq d_2$ for some d_2 (because ρ_k is bounded). So (1) is proved.

(2) We want to show $\alpha_\ell^{(k-1)} \to 0$ as $k \to \infty$ and $k \in K_0$, despite that $\alpha_\ell^{(k)} \to \omega$. We assume here that $\ell < n$, and the case $\ell = n$ is treated later in Lemma 2. Notice that $\alpha_{\ell+1}^{(k)} \to 0$ as $k \to \infty$ for $k \in K_0$ because $\eta_j = 0$ for $j > \ell$, and we have $|\alpha_\ell^{(k)}-\omega| < \delta$ for $k \geq N_2$, $k \in K_0$. From (3.4.3)

$$\begin{aligned}\alpha_\ell^{(k)} &= \xi_{k-1}\tau_{k-1}\alpha_\ell^{(k-1)} + \xi_{k-1}\alpha_{\ell+1}^{(k)} \\ &= \xi_{k-1}(\tau_k\alpha_\ell^{(k-1)} + \alpha_{\ell+1}^{(k)}) .\end{aligned}$$

As $k \to \infty$ for $k \in K_0$, $\xi_{k-1} \to \infty$. Therefore $(\tau_k\alpha_\ell^{(k-1)} + \alpha_{\ell+1}^{(k)}) \to 0$. But $\alpha_{\ell+1}^{(k)} \to 0$ as noted above and $\tau_k \geq d_1 > 0$ from (1). Hence, $\alpha_\ell^{(k-1)} \to 0$ as $k \to \infty$ for $k \in K_0$.

(3) For $k \geq N_2$, $k \in K_0$

$$\alpha_{\ell-1}^{(k)} = \xi_{k-1}(\tau_k\alpha_{\ell-1}^{(k-1)} + \omega - \omega + \alpha_\ell^{(k)})$$

but $0 \leq |\alpha_{\ell-1}^{(k)}| \leq \kappa$, so

$$|\tau_k\alpha_{\ell-1}^{(k-1)} + \omega - \omega + \alpha_\ell^{(k)}| \leq \kappa\delta .$$

Recall $|\omega-\alpha_\ell^{(k)}| < \delta$ for $k \geq N_2 \geq N_0$, so

$$\begin{aligned}||\tau_k\alpha_{\ell-1}^{(k-1)}| - |\omega|| &\leq |\tau_k\alpha_{\ell-1}^{(k-1)} + \omega| \\ &\leq |\tau_k\alpha_{\ell-1}^{(k-1)} + \omega - (\omega-\alpha_\ell^{(k)})| + |\omega-\alpha_\ell^{(k)}| \\ &\leq (\kappa+1)\delta ,\end{aligned}$$

i.e.

$$|\omega| - (\kappa+1)\delta \leq \tau_k\alpha_{\ell-1}^{(k-1)} \leq |\omega| + (\kappa+1)\delta$$

$$0 < h_1 = \frac{|\omega|-(\kappa+1)\delta}{d_2} \leq |\alpha_{\ell-1}^{(k-1)}| \leq \frac{|\omega|+(\kappa+1)\delta}{d_1} = h_2$$

because $d_1 \leq \tau_k \leq d_2$. □

We still need the following lemma before we proceed to the main theorem. The lemma is true for the complex case, but for simplicity, we shall only prove it for the real case.

Lemma 2. *Assume* $\rho_k \to 0$. *Then* $\alpha_n^{(k)} \to 0$ *as* $k \to \infty$.

Proof. Let all quantities be real and proceed by contradiction. Suppose $\alpha_n^{(k)}$ does not tend to zero. Then there exists ε such that $|\alpha_n^{(k)}| \geq \varepsilon$ for infinitely many k. Let N_3 be an integer so large that whenever $k \geq N_3$, $|\rho_k| < \delta$, where $\varepsilon/\sqrt{\delta_1} > n\kappa$ and $1/\sqrt{\delta_1} >> 1$. Now consider $V^{(k+1)}$ where $|\alpha_n^{(k)}| \geq \varepsilon$:

$$\alpha_n^{(k+1)} = \xi_k \tau_k \alpha_n^{(k)}$$

$$\text{(3.4.4)} \qquad \alpha_i^{(k+1)} = \xi_k \tau_k \sum_{j=i}^{n} \alpha_j^{(k)} \xi_k^{j-i}$$

$$= \tau_k \xi_k^{n-i+1} \alpha_n^{(k)} + O(\xi_k^{n-i}) \qquad (\text{where } \xi_k = 1/\rho_k)$$

by backward solving $(I-\xi_k N)V^{(k+1)} = \xi_k \tau_k V^{(k)}$ (see (3.4.3)).

Therefore, when we normalize $\alpha_1^{(k+1)} = 1$, then

$$|\alpha_i^{(k+1)}| = O(\xi_k^{n-i+1}/\xi_k^n)$$

$$= O(\xi_k^{1-i})$$

because $|\alpha_n^{(k)}| \geq \varepsilon$ and $\varepsilon/\sqrt{\delta_1} > n\kappa$.

From (3.4.4), it is clear that either all α_i's have the same sign as α_n or their signs alternate. We have the following two cases when we normalize $\text{sign}(\alpha_1^{(k+1)})$ to be positive.

Case 1: $\text{sign}(\alpha_i^{(k+1)})$ is positive for all i. This characteristic is invariant under RQI for subsequent steps because ρ_{k+1} is then positive and $\rho_{k+1} \doteq \xi_k$. Then (3.1.2) tells us that $\alpha_i^{(k+2)} \geq 0$ for all i.

$$\left|\frac{\alpha_i^{(k+2)}}{\alpha_{i+1}^{(k+2)}}\right| = \left|\frac{\alpha_i^{(k+1)} + \alpha_{i+1}^{(k+2)}/\rho_{k+1}}{\alpha_{i+1}^{(k+2)}}\right| = \left|\frac{\alpha_i^{(k+1)}}{\alpha_{i+1}^{(k+2)}}\right| + \left|\frac{1}{\rho_{k+1}}\right|$$

$$\to \infty \text{ as } \rho_{k+1} \to 0$$

because the two terms in the numerator are both positive. Thus $V^{(k)} \to x_1$ and so $\alpha_n^{(k)} \to 0$ contradicting our initial assumption.

<u>Case 2</u>: $\mathrm{sign}(\alpha_i^{(k+1)}) = (-1)^{i+1}$. This characteristic is also invariant under RQI for subsequent steps and $\rho_{k+1} \doteq -\delta_1$.

$$\left|\frac{\alpha_i^{(k+2)}}{\alpha_{i+1}^{(k+2)}}\right| = \left|\frac{\alpha_i^{(k+1)} + \alpha_{i+1}^{(k+2)}/\rho_{k+1}}{\alpha_{i+1}^{(k+2)}}\right| = \frac{|\alpha_i^{(k+1)}| + |\alpha_{i+1}^{(k+2)}/\rho_{k+1}|}{|\alpha_{i+1}^{(k+2)}|}$$

$$\to \infty \quad \text{as} \quad \rho_{k+1} \to 0$$

because ρ_{k+1} is negative and $\mathrm{sign}(\alpha_{i+1}^{(k+2)}) \neq \mathrm{sign}(\alpha_{i+1}^{(k+1)})$.

Thus $V^{(k)} \to x_1$ and $\alpha_n^{(k)} \to 0$, a contradiction. □

Let us summarize the picture so far: For k large enough and $k \in K_0$, we have

$$\begin{array}{rc}
 & V^{(k-1)} \\
 & \left(\begin{array}{c} x \\ \vdots \\ x \\ \hat{p} \\ a_\ell \\ a_{\ell+1} \\ \vdots \\ a_n \end{array}\right) \\
\end{array}
\begin{array}{c} \\ \\ \\ \ell-1 \to \\ \ell \to \\ \\ \\ \\ \end{array}
\dashrightarrow
\begin{array}{cl}
V^{(k)} & \\
\left(\begin{array}{c} x \\ \vdots \\ x \\ q \\ b_{\ell+1} \\ \vdots \\ b_n \end{array}\right) & \begin{array}{l} \\ \\ \\ \leftarrow \ell \\ \leftarrow \ell+1 \\ \\ \\ \end{array}
\end{array}$$

$q \doteq \omega$, $0 < h_1 \le |\hat{p}| \le h_2$, b's tend to zero (because $\eta_{\ell+1} = \cdots = \eta_n = 0$), a_n tends to zero by Lemma 2, a_ℓ tends to zero by Lemma 1 and $a_{\ell+1},\ldots,a_{n-1}$ tend to zero for the same reason as a_ℓ (by considering (3.4.2)).

Now we can proceed inductively. Consider $A = \{V^{(k-1)} \mid k \in K_0\}$. Let u_1 be a limit vector of A and let $K_1 \subseteq K_0$ be a subset for which $V^{(k)}$, $k \in K_1 \to u_1$ as $k \to \infty$. By the remarks just above, $u_1 = (x,\dots,x,p,0,\dots,0)$ where p is the $(\ell-1)^{st}$ component and $|p| \geq h_1 > 0$. By repeating the same arguments as before, we find that for $k \in K_1$ and k large enough

$$\begin{array}{rcccl}
 & V^{(k-2)} & & V^{(k-1)} & \\
\begin{array}{r} \\ \\ \\ \\ \ell-2 \rightarrow \\ \ell-1 \rightarrow \\ \\ \\ \\ \end{array} &
\left(\begin{array}{c} x \\ \vdots \\ x \\ r \\ g_{\ell-1} \\ \vdots \\ g_n \end{array}\right) &
\dashrightarrow &
\left(\begin{array}{c} x \\ \vdots \\ x \\ s \\ a_{\ell} \\ \vdots \\ a_n \end{array}\right) &
\begin{array}{l} \\ \\ \\ \\ \leftarrow \ell-1 \\ \leftarrow \ell \\ \\ \\ \\ \end{array}
\end{array}$$

where g's and a's tend to zero, and $|r| \geq h_3 > 0$ and so on. Finally we have, $k \in K_{\ell-2} \subseteq K_{\ell-1} \subseteq \cdots \subseteq K_1 \subseteq K_0$, k large enough

$$\begin{array}{ccccccc}
V^{(k-\ell+2)} & \dashrightarrow & V^{(k-\ell+1)} & \dashrightarrow & \cdots & \dashrightarrow & V^{(k)} \\
\left(\begin{array}{c} x \\ t \\ c_3 \\ \vdots \\ \vdots \\ c_n \end{array}\right) & \dashrightarrow & & \cdots & & \dashrightarrow & \left(\begin{array}{c} x \\ \vdots \\ x \\ q \\ b_{\ell+1} \\ \vdots \\ b_n \end{array}\right)
\end{array}$$

where c's tend to zero as $k \to \infty$, $k \in K_{\ell-2}$ and $|t| \geq h_4 > 0$. Consider the set $B = \{V^{(k-\ell+2)} | \, k \in K_{\ell-2}\}$. Let w be a limit vector of B. Then $w = (\zeta_1,\zeta_2,0,\dots,0)^T$ with $|\zeta_2| \geq h_4 > 0$ (by arguments similar to that of Lemma 1(3)). Now $\rho(w) = \zeta_1\zeta_2 = 0$, since $\rho_k \to 0$, and $\zeta_2 \neq 0$ so $\zeta_1 = 0$, when normalized $|\zeta_2| = 1$ and so $w = x_2$.

Hence we have the following situation: Whenever $\ell \geq 2$, then x_2 is a limit vector. Therefore if x_2 is not a limit vector then $\ell < 2$, which means the vector iteration converges to x_1.

On the other hand, if x_2 is a limit vector, we know by the same argument as presented earlier in this section that x_1 must also be a limit vector, and so the vector sequence must cycle. We suspect that this cycling is impossible because otherwise it would mean that, over and over again, the vector sequence approaches x_1 and gets thrown out to a region that is close to x_2, which is orthogonal to x_1. This seems hard to realize, but using the condition $\rho_k \to 0$ alone is not sufficient to prove that the cycling is impossible. This fact will be shown in our next section.

We have, therefore, shown how to prove the following result:

Theorem. If ρ_k converges, then the vector iteration converges if and only if x_2 is not a limit vector.

Note. We merely say "ρ_k converges" because by the corollary of the theorem in the last section, whenever ρ_k converges, ρ_k converges to λ.

§3.5 A Shift Sequence Which Prevents Convergence

In this section, we show the surprising result that $\rho_k \to \lambda$ does not imply $v^{(k)} \to x_1$. But with a very weak hypothesis, the generalized eigenvector of second grade cannot be a limit vector and thus by the result of the last section, the vector iteration converges.

We shall first derive those weak conditions. Let C be normalized such that $\lambda = 0$, and suppose x_2, the generalized eigenvector of second grade, is a limit vector of RQI. Let $V^{(0)} = (\eta_1,\eta_2,\ldots,\eta_n)^T$ be the starting vector. Then by definition, there exists an infinite set $K \subseteq \{1,2,3,\ldots\}$ such that $V^{(k)} \to x_2$ for $k \in K$. Let this sequence of $V^{(k)}$ be denoted by W_i, $i = 1,2,\ldots$. By the definition of RQI:

$$\begin{aligned}
&\hat{\tau}_1 \Big(\prod_{j=1}^{k_1} (I-\xi_j N)\Big)^{-1} V = W_1 \\
(3.5.1) \qquad &\hat{\tau}_2 \Big(\prod_{j=1}^{k_2} (I-\xi_j N)\Big)^{-1} V = W_2 \\
&\hat{\tau}_3 \Big(\prod_{j=1}^{k_3} (I-\xi_j N)\Big)^{-1} V = W_3 \\
&\qquad\vdots \\
&\qquad\text{etc.}
\end{aligned}$$

where $\hat{\tau}_i$ is the normalizing factor so that $\|W_i\| = \|V\| = 1$ and $\xi_j = 1/\rho_j$. Denote $\Big(\prod_{j=1}^{k_m} (I-\xi_j N)\Big)^{-1}$ by $B_m = (b_{ij}^{(m)})$, $m = 1,2,\ldots$,

$$B_m = \begin{pmatrix} 1 & b_{12}^{(m)} & b_{13}^{(m)} & \cdots & b_{1n}^{(m)} \\ & 1 & b_{23}^{(m)} & \cdots & b_{2n}^{(m)} \\ & & 1 & & \vdots \\ & \bigcirc & & \ddots & \\ & & & & 1 \end{pmatrix} .$$

B_m is an upper unit triangular Toeplitz matrix.

<u>Case 1</u>: $\eta_n \neq 0$. The fact that $b_{nn}^{(i)} = 1$, B_m upper triangular and $W_i = \tau_i B_i V \to x_2$, whose last component is zero, implies that $\hat{\tau}_i \to 0$, and consequently some entries of the second row of B_i must tend to infinity.

Lemma. $\left|\frac{b^{(i)}_{23}}{b^{(i)}_{2n}}\right|, \left|\frac{b^{(i)}_{24}}{b^{(i)}_{2n}}\right|, \ldots, \left|\frac{b^{(i)}_{2,n-1}}{b^{(i)}_{2,n}}\right|$ all tend to zero as $i \to \infty$, i.e., $b^{(i)}_{2n}$ tends to infinity faster than any other entry in row 2 of B_i.

Proof. The proof exploits the Toeplitz matrix structure of B_i. So suppose that one of the ratios does not tend to zero. Let $\Gamma \subseteq \{1,2,\ldots,n\}$ be such that when $j \in \Gamma$, $\left|\frac{b^{(i)}_{2j}}{b^{(i)}_{2n}}\right|$ does not tend to zero.

Case 1: There exists M, a constant, such that $\left|\frac{b^{(i)}_{2j}}{b^{(i)}_{2n}}\right| \leq M$ for all $j \in \Gamma$ and all i. Let $m = \min\{j | j \in \Gamma\}$. Then there exists an infinite set $K_1 \subseteq \{1,2,\ldots\}$ such that when $i \in K_1$, $\left|\frac{b^{(i)}_{2m}}{b^{(i)}_{2n}}\right| \geq \varepsilon > 0$ for some ε. Since $b^{(i)}_{2m} = b^{(i)}_{n-m+3n}$ (they are on the same diagonal), the ratio

$$\left|\frac{(n-m+2)^{\text{th}} \text{ component of } W_i}{2^{\text{nd}} \text{ component of } W_i}\right| \geq \frac{\varepsilon}{\varepsilon + (|\Gamma|-1)M + 1} \quad \text{as} \quad i \to \infty, \quad i \in K_1 .$$

Therefore W_i does not converge to x_2, a contradiction.

Case 2: There exists $\Gamma_1 \subseteq \{1,2,\ldots,n\}$ such that for $j \in \Gamma_1$ $\left|\frac{b^{(i)}_{2j}}{b^{(i)}_{2n}}\right|$ is greater than any constant infinitely many times since $b^{(i)}_{2j} = b^{(i)}_{n-j+2,n}$, at least one $j \in \Gamma_1$ such that

$$\left|\frac{(n-j+2)^{\text{th}} \text{ component of } W_i}{2^{\text{nd}} \text{ component of } W_i}\right| \geq \frac{1}{|\Gamma_1| + 1}$$

for infinitely many i. Therefore W_i cannot converge to x_2, a contradiction. Here ends the proof of the lemma. □

Let $B_i' = B_i/b_{2,n}^{(i)}$ ($b_{2,n}^{(i)} \neq 0$ for large enough i because it tends to infinity). Then as $i \to \infty$,

$$B_i' \to \begin{pmatrix} 0 & \cdots & 0 & 1 & b_{1n}^{(i)}/b_{2n}^{(i)} \\ & & & 0 & 1 \\ & & & & 0 \\ & \bigcirc & & & \vdots \\ & & & & 0 \end{pmatrix} .$$

Therefore, x_2 is a limit vector if and only if $b_{1n}^{(i)}/b_{2n}^{(i)}$ tends to $-\eta_{n-1}/\eta_n$.

Case 2: There exists ℓ such that $\eta_\ell \neq 0$, $\eta_j = 0$ for $j > \ell$.

If $\ell > 2$, this just reduces the problem to the consideration of an $\ell \times \ell$ matrix. So, exactly as before, x_2 is a limit vector if and only if $b_{1\ell}^{(i)}/b_{2\ell}^{(i)}$ tends to $-\eta_{\ell-1}/\eta_\ell$.

If $\ell = 2$, the τ_i may not tend to zero, but this case behaves as if the matrix is 2×2, and the argument is easy when C is 2×2. If we let $V^{(0)} = (\eta_1, \eta_2)^T$, then x_2 is not a limit vector if and only if $-\eta_1/\eta_2$ is not a limit point of the sequence $s_i = \sum_{k=1}^{i} \xi_k = \sum_{k=1}^{i} \frac{1}{\rho_k}$.

Now we can summarize what we have done: Let $F_k = \left(\prod_{j=1}^{k} (I - \xi_j N)\right)^{-1}$, $F_k = (f_{ij}^{(k)})$, where

$f_{12}^{(k)}$ = sum of ξ_i , $O(k)$ of them as $k \to \infty$

$f_{13}^{(k)}$ = sum of terms like $\xi_{i_1}\xi_{i_2}$, $O(k^2)$ of them as $k \to \infty$

$\vdots$

$f_{1\ell}^{(k)}$ = sum of terms like $\xi_{i_1} \cdots \xi_{i_{\ell-1}}$, $O(k^{\ell-1})$ of them as $k \to \infty$.

We have shown the following theorem:

Theorem. *Let* C *be a completely degenerate matrix with* 0 *as its eigenvalue*. *Let* $V^{(0)} = (\eta_1, \eta_2, \ldots, \eta_n)^T$ *be the initial vector with* $\eta_\ell \neq 0$, $\eta_i = 0$ *for* $i > \ell$. *Suppose* $\rho_k \to 0$. *Then* RQI *will converge to the eigenvector if and only if*

(3.5.2) $-\eta_{\ell-1}/\eta_\ell$ *is not a limit point of the sequence* $s_i = f_{1\ell}^{(i)}/f_{i,\ell-1}^{(i)}$.

Remark. The last condition is weak and highly technical.

In all of our experience with RQI, s_i always diverges. Even if we can construct a sequence of ρ_k that tends to zero with s_i having countably many limit points, the choice of initial vectors that could force RQI not to converge is a union of $n-1$ dimensional vector spaces, and hence of measure zero in the n-dimensional space.

In the rest of this section, we shall present a sequence of numbers that tend to zero, which, if they are used as shifts, would force the vector iteration not to converge, when an arbitrary but fixed initial vector $V^{(0)} = (\eta_1, \eta_2, \ldots, \eta_n)^T$ is given.

Without loss of generality, we may assume $\eta_n \neq 0$. Suppose

$$-\eta_{n-1}/\eta_n = \phi \ .$$

We shall construct the following sequence:

$$\begin{aligned}
&\psi_1 = q_1^{(1)} \geq q_2^{(1)} \geq q_3^{(1)} \geq q_4^{(1)} \geq \cdots \geq q_{m(1)}^{(1)}, \\
&|\psi_2| \geq q_{m(2)}^{(2)} \geq q_{m(2)+1}^{(2)} \geq q_{m(2)+2}^{(2)} \geq \cdots \geq q_{m(2)}^{(2)}, \\
(3.5.3)\qquad &|\psi_3| \geq q_{m(3)}^{(3)} \geq q_{m(3)+1}^{(3)} \geq \cdots \geq q_{m(3)}^{(3)}, \\
&\qquad\qquad\vdots
\end{aligned}$$

where $q_j^{(i)} = 1/j$ for all i, and ψ_i are to be determined such that $|\psi_i| \leq 1/i$. If the last condition is satisfied, then the sequence constructed by putting ψ_2 right after $q_{m(1)}^{(1)}$'s, ψ_3 right after $q_{m(2)}^{(2)}$'s, etc. would tend to zero.

The ψ_i are inserted at strategic points to force $f_{1\ell}^{(k)}/f_{1\ell-1}^{(k)}$, as defined earlier, to tend to ϕ, and thus force the vector iterations not to converge. The ψ_i's are obtained from the following steps:

Step 1: $\psi_1 = 1$. If ψ_i is known, pick $m(i)$ an integer so large that $|\psi_i| \geq 1/m(i)$.

Step 2: $\psi_i, q_{m(i)}^{(i)}, q_{m(i)+1}^{(i)}, q_{m(i)+2}^{(i)}, \ldots$ is a sequence tending to zero. Using these as shifts, either x_2 is a limit vector, and thus our goal is reached and no further work has to be done, or x_1 is the only limit vector (by the theorem in Section 4).

Step 3: Solve for ψ_{i+1}: Recall that we take the shifts $\rho_j = q_{m(i)+j-1}$ and

$$V^{(k)} = \hat{\tau}_k \Big(\prod_{j=1}^{k} (I-(1/\rho_j)N) \Big)^{-1} V^{(0)} \to x_1 \ .$$

Let us define $B_k \equiv \Big(\prod_{j=1}^{k} (I-(1/\rho_j)N) \Big)^{-1}$ and then, normalizing it,

$$B'_k = B_k/((1,n)\text{-element of } B_k) \ .$$

Then by the same arguments used in the last theorem,

$$\lim_{k\to\infty} B'_k = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ & & & 0 \\ & \bigcirc & & \vdots \\ & & & 0 \end{bmatrix} .$$

So for a particular k, let $B'_k = (b_{ij}^{(k)})$, and let ξ denote the unknown to be found. We know $(I-\xi N)^{-1} = I+\xi N+\xi^2N^2+\cdots+\xi^{n-1}N^{n-1}$, the $(1,n)$ element of $(I-\xi N)^{-1}B_k'$ is

$$1+\xi b_{1,n-1}^{(k)}+\xi^2 b_{1,n-2}^{(k)}+\cdots+\xi^{n-1}b_{1,1} ,$$

the $(1,n-1)$ element of $(I-\xi N)^{-1}B_k'$ is

$$b_{1,n-1}^{(k)}+\xi b_{1,n-2}+\cdots+\xi^{n-2}b_{1,1} .$$

The crucial ratio becomes

$$\text{(3.5.4)} \quad \frac{b_{1,n}^{(k+1)}}{b_{1,n-1}^{(k+1)}} = \frac{1+\xi b_{1,n-1}^{(k)}+\xi^2 b_{1,n-2}^{(k)}+\cdots+\xi^{n-1}b_{1,1}^{(k)}}{b_{1,n-1}^{(k)}+\xi b_{1,n-2}^{(k)}+\cdots+\xi^{n-2}b_{1,1}^{(k)}} = \phi .$$

The above equation can be solved because the complex numbers are algebraically complete.

We know that when k is large, $b_{1,j}^{(k)}$ is small for all $j \neq n$. In (3.5.4), when $b_{1,j}^{(k)}$ tend to zero, ξ tend to infinity. Therefore, there exists k so large, such that $b_{1,j}^{(k)}$ are so small that a solution ξ whose absolute value inverse $1/|\xi| \leq 1/(i+1)$.

Hence, $m(i)'$ is an integer so large (and thus k so large), that the absolute value inverse of a solution ξ of (3.5.4) is less than $1/(i+1)$. Choose $\psi_{i+1} = 1/\xi$.

Step 4: Now repeat steps 1 through 3 to find $\psi_{i+2},\psi_{i+3},\cdots$.

Thus we have a sequence of numbers that tend to zero, and, by arguments of the last theorem and choice of ψ_i, we know that using these numbers as shifts would force the vector iteration not to converge.

Remark. Through the intimate relation between SQR and inverse iteration with shifts, it is surprising to learn, by the last example, that SQR may not converge when operated on a general matrix even though the shifts tend towards an eigenvalue.

CHAPTER FOUR

The General Case

We are now in a position to summarize some of our results in the previous two chapters into a theorem.

Let C be a general complex $n \times n$ matrix. We adopt the following notation:

Notation: $\lambda_1, \ldots, \lambda_m$ are eigenvalues of C such that if the null space of $C - \lambda_i I$ is of dimension j, then λ_i appears j times among $\lambda_1, \ldots, \lambda_m$. Let P be a non-singular matrix such that $P^{-1}CP = J$, the Jordan canonical form of C. Then the column vectors of P are the generalized eigenvectors of C. Let $x_i^{(1)}$ be the column vector of P such that $Cx_i^{(1)} = \lambda_i x_i^{(1)}$, and $x_i^{(j)}$ be the column vector of P such that $Cx_i^{(j)} = \lambda_i x_i^{(j)} + x_i^{(j-1)}$.

Definition. Let $V = \sum_i \sum_j \alpha_{ij} x_i^{(j)}$ be a vector expressed in the basis of generalized eigenvectors mentioned above. We say V is *deficient* in λ_i if for all k such that $\lambda_k = \lambda_i$, $\alpha_{kj} = 0$ for all j.

Theorem. *Let* C *be a complex* $n \times n$ *matrix, let* $V^{(0)} = \sum_i \sum_j \alpha_{ij}^{(0)} x_i^{(j)}$ *be the initial vector and* V *a limit vector of the* RQI.

Suppose that $\rho_k = \rho(V^{(k)}) \to \rho$ *as* $k \to \infty$. *Then either* I. $\rho = \lambda_i$ *for some* i *and* $V^{(0)}$ *is not deficient in* λ_i: *in which case* V *is a vector in the generalized eigenspace of* λ_i, i.e., V *is a linear combination of generalized eigenvectors associated with* λ_i. *If further the weak* *condition* (3.5.2) *of Theorem 3.5 is true for each Jordan block whose associated eigenvalue* λ_j *equal* λ_i, *then* V *is actually an eigenvector such that* $CV = \lambda_i V$.

or II(a). $\rho = \lambda_i$ *for some* i *and* $V^{(0)}$ *is deficient in* λ_i,

or II(b). ρ *is not an eigenvalue of* C.

In either case of II $V = \sum_{j\in\Gamma} \psi_j x_j^{(1)}$ *where* $\Gamma \subseteq \{1,2,\ldots,m\}$ *and* $|\lambda_j-\rho| = m$, *a constant for all* $j \in \Gamma$. *Case II is unstable*.

Proof. (I) $\rho = \lambda_i$ and $V^{(0)}$ not deficient in λ_i: Let ℓ be the index such that $\alpha_{i\ell}^{(0)} \neq 0$ but $\alpha_{ij}^{(0)} = 0$ for $j > \ell$. Then one step of RQI gives (see (3.1.2) where $n = \ell$ there):
$\alpha_{i\ell}^{(k+1)} = \tau_k(\alpha_{i\ell}^{(k)}/(\lambda_i-\rho_k))$, and for j such that $\alpha_j \neq \alpha_i$

$$\begin{aligned}\alpha_{j1}^{(k+1)} &= \tau_k\alpha_{j1}^{(k)}((\lambda_j-\rho_k)^{-1} + \sum_{p=2}^{q}(\lambda_j-\rho_k)^{-p}\frac{\alpha_{jp}^{(k)}}{\alpha_{j1}^{(k)}}) \\ &= \tau_k\alpha_{j1}^{(k)}m_{j1}^{(k)} ,\end{aligned}$$

where q = Jordan block size of λ_j, and
$m_{j1}^{(k)} = (\lambda_j-\rho_k)^{-1} + \sum_{p=2}^{q}(\lambda_j-\rho_k)^{-p}\alpha_{jp}^{(k)}/\alpha_{j1}^{(k)}$. Now $\rho_k \to \lambda_i \neq \lambda_j$ implies there exists N_1 such that if $k \geq N_1$, $\max\{(\lambda_j-\rho_k)^p|\ 1\leq p\leq q\} \leq M_1$ for some constant M_1. Then, by the theorem in Chapter 3, section 3, we know $|\alpha_{jp}^{(k)}/\alpha_{j1}^{(k)}| \to 0$ as $k \to \infty$. There exists N_2 such that if $k \geq N_2$, $\max\{|\alpha_{jp}^{(k)}/\alpha_{j1}^{(k)}||\ 2\leq p\leq q\} \leq 1$. Let $N = \max\{N_1,N_2\}$. Then for $k \geq N$, $|m_{j1}^{(k)}| \leq qM_1$. Hence for $k \geq N$

$$\left|\frac{\alpha_{i\ell}^{(k+N)}}{\alpha_{j1}^{(k+N)}}\right| = \left(\prod_{s=1}^{k}\left|\frac{m_{j1}^{(N+s)}}{\lambda_i-\rho_{N+s}}\right|\right)\left|\frac{\alpha_{i\ell}^{(N)}}{\alpha_{j1}^{(N)}}\right| \to \infty$$

as $\rho_k \to \lambda_i$ as $k \to \infty$. Thus $\alpha_{j1}^{(k)} \to 0$ as $k \to \infty$, and as mentioned above $|\alpha_{jp}^{(k)}/\alpha_{j1}^{(k)}| \to 0$, hence $\alpha_{jp}^{(k)} \to 0$ as $k \to \infty$. Consequently, a limit vector V must be deficient in λ_j for all $\lambda_j \neq \lambda_i$, i.e., V is a vector in the generalized eigenspace of λ_i. If further the hypothesis of the theorem in Chapter 3, section 5 is fulfilled, then V is an eigenvector such that $CV = \lambda_i V$. (Note the hypothesis is automatically satisfied if all the eigenvectors corresponding to λ_i are of the lowest grade or if the coefficients of the higher grade eigenvectors are all zero in the starting vector).

(ii) By the theorem in Chapter 3, section 3, $\alpha_{jp}^{(k)} \to 0$ as $k \to \infty$ for all $p > 1$ and all j. Thus a vector can only be a linear combination of eigenvectors of the lowest grade. The conclusion then follows from the theorem in Chapter 2, section 7. The fact that this case is unstable is the result of the discussion of Chapter 2, section 8.

□

Remark. By this theorem, we have a complete characterization of limit vectors provided ρ_k converges. Thus the global picture of RQI on a general matrix is reduced to the convergence of ρ_k

Bibliography

[1] H.J. Buurma, A geometric proof of convergence for the QR method, Thesis, Rijksuniversiteit Te Groningen, 1970.

[2] K.L. Kim, A numerical study of a variant of the Rayleigh Quotient Iteration, Thesis, The University of Texas at Austin, 1970.

[3] Marcus and Minc, *A Survey of Matrix Theory and Matrix Inequalities*. Allyn and Bacon, Inc., Boston, 1964.

[4] A.M. Ostrowski, On the convergence of the Rayleigh Quotient iteration for the computation of the characteristic roots and vectors, Parts I-VI, *Arch. Rational Mech. Anal.*, 1 (1958), pp. 233-241; 2 (1959/59), pp. 423-428; 3 (1959), pp. 325-340, pp. 341-347, pp. 472-481; 4 (1959), pp. 153-165.

[5] B.N. Parlett and W. Kahan, On convergence of a practical QR algorithm, *Proc. IFIP Congress 1968*, pp. A25-A30.

[6] B.N. Parlett and W.G. Poole, Jr., A geometric theory for the QR, LU and power iterations, *SIAM J. Numerical Anal.* 10 (1973), pp. 389-411.

[7] B.N. Parlett, The Rayleigh Quotient Iteration and some generalizations for nonnormal matrices, *Mathematics of Computation* 28 (1974), pp. 679-693.

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

[9] J.H. Wilkinson, Inverse iteration in theory and in practice, *Symposia Mathematica* X (1972), pp. 361-379.

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

Computer Science Division
Electronic Research Laboratory
University of California at Berkeley 94720

2a. REPORT SECURITY CLASSIFICATION

2b. GROUP

3. REPORT TITLE

THE RAYLEIGH QUOTIENT ITERATION FOR NON-NORMAL MATRICES

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Scientific Final

5. AUTHOR(S) *(First name, middle initial, last name)*

Nai-fu Chen

6. REPORT DATE December 1975	7a. TOTAL NO. OF PAGES 80	7b. NO. OF REFS 9

8a. CONTRACT OR GRANT NO. ONR-N00014-69-A-0200-1017	9a. ORIGINATOR'S REPORT NUMBER(S) M-548
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Mathematics Branch Office of Naval Research Washington, D.C. 20360

13. ABSTRACT

The Rayleigh Quotient Iteration (RQI) is a method for computing eigenvectors and eigenvalues of a square matrix.

The behaviour, both local and global, of RQI with symmetric and normal matrices is almost completely understood. The vector sequence converges for almost all starting vectors.

In this paper, we investigate the global properties of RQI on non-normal matrices. Results on nearly normal matrices with real eigenvalues are obtained, and at the other extreme, results on completely degenerate matrices are also obtained. In particular, the question of global convergence of the vector iteration on a general matrix is reduced to the convergence of the scalar sequence of the Rayleigh quotients. In practice, the vector iteration always converges.

The main difficulty in the degenerate case is that the iteration function is discontinuous near the eigenvector. An example is used to display the sectorial behaviour of the iteration. Further we construct a sequence of numbers that converges to the eigenvalue. Yet if the numbers are used as shifts with inverse iteration, the vector sequence fails to converge.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
eigenvector						
eigenvalue						
iterative methods						
Rayleigh Quotient						
global convergence						
nonnormal matrix						