NASA Technical Memorandum 105858
ICOMP–92–18

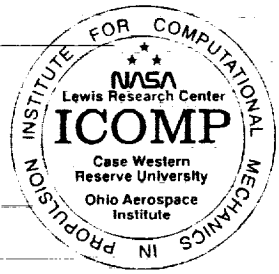
# Application of Vector-Valued Rational Approximations to the Matrix Eigenvalue Problem and Connections With Krylov Subspace Methods

Avram Sidi
*Technion-Israel Institute of Technology*
*Haifa, Israel*

*and Institute for Computational Mechanics in Propulsion*
*Lewis Research Center*
*Cleveland, Ohio*



September 1992




NASA

# APPLICATION OF VECTOR-VALUED RATIONAL APPROXIMATIONS TO THE MATRIX EIGENVALUE PROBLEM AND CONNECTIONS WITH KRYLOV SUBSPACE METHODS

Avram Sidi
Computer Science Department
Technion - Israel Institute of Technology
Haifa 32000, Israel
and
Institute for Computational Mechanics in Propulsion
NASA Lewis Research Center
Cleveland, Ohio 44135, U.S.A.

The results of this paper were presented at the International Meeting on Approximation, Interpolation, and Summability, Tel-Aviv, June 1990, and at the International Congress on Extrapolation and Rational Approximation, Tenerife, January 1992.

**Abstract**

Let $F(z)$ be a vector-valued function $F : C \rightarrow C^N$, which is analytic at $z = 0$ and meromorphic in a neighborhood of $z = 0$, and let its Maclaurin series be given. In a recent work [Si6] by the author, vector-valued rational approximation procedures for $F(z)$ that are based on its Maclaurin series, were developed, and some of their convergence properties were analyzed in detail. In particular, a Koenig type theorem concerning their poles and a de Montessus type theorem concerning their uniform convergence in the complex plane were given. With the help of these theorems it was shown how optimal approximations to the poles of $F(z)$ and the principal parts of the corresponding Laurent series expansions can be obtained. In the present work we use these rational approximation procedures in conjunction with power iterations to develop bona fide generalizations of the power method for an arbitrary $N \times N$ matrix that may be diagonalizable or not. These generalizations can be used to obtain simultaneously several of the largest distinct eigenvalues and corresponding eigenvectors and other vectors in the invariant subspaces. We provide interesting constructions for both nondefective and defective eigenvalues and the corresponding invariant subspaces, and present a detailed convergence theory for them. This is made possible by the observation that vectors obtained by power iterations with a matrix are actually coefficients of the Maclaurin series of a vector-valued rational function, whose poles are the reciprocals of some or all of the nonzero eigenvalues of the matrix being considered, while the principal parts of the Laurent expansions of this rational function are vectors in the corresponding invariant subspaces. In addition, it is shown that the generalized power methods of this work are equivalent to some Krylov subspace methods, among them the methods of Arnoldi and Lanczos. Thus, the theory of the present work provides a set of completely new results and constructions for these Krylov subspace methods. This theory suggests at the same time a new mode of usage for these Krylov subspace methods that has been observed to possess computational advantages over their common mode of usage.

# 1 Introduction

Let $F(z)$ be a vector-valued function, $F : C \to C^N$, which is analytic at $z = 0$ and meromorphic in a neighbourhood of $z = 0$, and let its Maclaurin series be given as

$$F(z) = \sum_{m=0}^{\infty} u_m z^m, \tag{1.1}$$

where $u_m$ are fixed vectors in $C^N$.

In a recent work by the author [Si6] three types of vector-valued rational approximation procedures that are entirely based on the expansion in (1.1) were proposed. For each of these procedures the rational approximations have two indices, $n$ and $k$, attached to them, and thus form a two-dimensional table akin to the Padé table or the Walsh array. Let us denote the $(n, k)$ entry of this table by $F_{n,k}(z)$. Then $F_{n,k}(z)$, if it exists, is defined to be of the form

$$F_{n,k}(z) = \frac{\sum_{j=0}^{k} c_j^{(n,k)} z^{k-j} F_{n+\nu+j}(z)}{\sum_{j=0}^{k} c_j^{(n,k)} z^{k-j}} \equiv \frac{P_{n,k}(z)}{Q_{n,k}(z)} \quad \text{with } c_k^{(n,k)} = Q_{n,k}(0) = 1, \tag{1.2}$$

where $\nu$ is an arbitrary but otherwise fixed integer $\geq -1$, and

$$F_m(z) = \sum_{i=0}^{m} u_i z^i, \ m = 0, 1, 2, ...; \quad F_m(z) \equiv 0 \quad \text{for } m < 0, \tag{1.3}$$

and the $c_j^{(n,k)}$ are scalars that depend on the approximation procedure being used.

If we denote the three approximation procedures by SMPE, SMMPE, and STEA, then the $c_j^{(n,k)} \equiv c_j$, for each of the three procedures, are defined such that they satisfy a linear system of equations of the form

$$\sum_{j=0}^{k-1} u_{ij} \, c_j = -u_{ik}, \ 0 \leq i \leq k-1; \ c_k = 1, \tag{1.4}$$

where $u_{ij}$ are scalars defined as

$$u_{ij} = \begin{cases} (u_{n+i}, u_{n+j}) & \text{for SMPE,} \\ (q_{i+1}, u_{n+j}) & \text{for SMMPE,} \\ (q, u_{n+i+j}) & \text{for STEA.} \end{cases} \tag{1.5}$$

Here $(\cdot \ , \ \cdot)$ is an inner product - not necessarily the standard Euclidean inner product - whose homogeneity property is such that $(\alpha x, \beta y) = \overline{\alpha}\beta(x, y)$ for $x, y \in C^N$ and $\alpha, \beta \in C$. The vectors

$q_1, q_2, ...,$ form a linearly independent set, and the vector $q$ is nonzero. Obviously, $F_{n,k}(z)$ exists if the linear system in (1.4) has a solution for $c_0, c_1, ..., c_{k-1}$.

It is easy to verify that for SMPE the equations in (1.4) involving $c_0, c_1, ..., c_{k-1}$ are the normal equations for the least squares problem

$$\min_{c_0,c_1,...,c_{k-1}} \left\| \sum_{j=0}^{k-1} c_j \, u_{n+j} \; + \; u_{n+k} \right\|, \tag{1.6}$$

where the norm $\|\cdot\|$ is that induced by the inner product $(\cdot\,,\,\cdot)$, namely, $\|x\| = \sqrt{(x,x)}$.

As is clear from (1.2) and (1.3), the numerator of $F_{n,k}(z)$ is a vector-valued polynomial of degree at most $n + \nu + k$, whereas its denominator is a scalar polynomial of degree at most $k$.

As can be seen from (1.4) and (1.5), the denominator polynomial $Q_{n,k}(z)$ is constructed from $u_n, u_{n+1}, ..., u_{n+k}$ for SMPE and SMMPE, and from $u_n, u_{n+1}, ..., u_{n+2k-1}$ for STEA. Once the denominators have been determined, the numerators involve $u_0, u_1, ..., u_{n+\nu+k}$ for all three approximation procedures.

The approximation procedures above are very closely related to some vector extrapolation methods. In fact, as is stated in Theorem 2.3 in Section 2 of [Si6], $F_{n,k}(z)$ for SMPE, SMMPE, and STEA are obtained by applying some generalized versions of the minimal polynomial extrapolation (MPE), the modified minimal polynomial extrapolation (MMPE), and the topological epsilon algorithm (TEA), respectively, to the vector sequence $F_m(z), m = 0, 1, 2, ...$ . For early references pertaining to these methods and their description, see the survey paper of Smith, Ford, and Sidi [SmFSi], and for recent developments pertaining to their convergence, stability, implementation, and other additional properties, see the papers by Sidi [Si1, Si2, Si5], Sidi and Bridger [SiB], Sidi, Ford, and Smith [SiFSm], and Ford and Sidi [FSi]. The above mentioned generalizations of vector extrapolation methods are given in [SiB, eqs.(1.16) and (1.17).].

A detailed convergence analysis for the approximations $F_{n,k}(z)$ as $n \to \infty$ was given in [Si6], whose main results can be verbally summarized as follows: (1) Under certain conditions the denominators $Q_{n,k}(z)$ converge, and their zeros, $k$ in number, tend to the $k$ poles of $F(z)$ that are

closest to the origin. This is a Koenig type result and is proved in Theorems 4.1 and 4.5 of [Si6], where the precise rates of convergence are also given for both simple and multiple poles of $F(z)$, and optimal approximations to multiple poles are constructed in a simple way. (2) Under the same conditions $F_{n,k}(z)$ converges to $F(z)$ uniformly in any compact subset of the circle containing the above mentioned $k$ poles of $F(z)$ with these poles excluded. This is a de Montessus type theorem and is proved in Theorem 4.2 of [Si6]. (3) The principal parts of the Laurent expansions of $F(z)$ about its poles, simple or multiple, can be constructed from $F_{n,k}(z)$ only. This construction, along with its convergence theory, is provided in Theorem 4.3 of [Si6].

It turns out that the denominator polynomials $Q_{n,k}(z)$ are very closely related to some recent extensions of the power method for the matrix eigenvalue problem, see [SiB, Section 6] and [Si3]. Specifically, if the vectors $u_m$ of (1.1) are obtained from $u_m = Au_{m-1}, m = 1, 2, ...$, with $u_0$ arbitrary, and $A$ being a complex $N \times N$ and, in general, nondiagonalizable matrix, then the reciprocals of the zeros of the polynomial $Q_{n,k}(z)$ are approximations to the $k$ largest distinct and, in general, defective eigenvalues of $A$, counted according to their multiplicities, under certain conditions. In Section 3 of the present work we provide precise error bounds for these approximations for $n \to \infty$ that are based on the results of Theorems 4.1 and 4.5 of [Si6]. While the approximations to nondefective eigenvalues have optimal accuracy in some sense, those that correspond to defective eigenvalues do not. In this paper we also show how approximations of optimal accuracy to defective eigenvalues can be constructed solely from $Q_{n,k}(z)$, providing their convergence theory for $n \to \infty$ at the same time. We then extend the treatment of [SiB] and [Si3] to cover the corresponding invariant subspaces in general, and the corresponding eigenvectors in particular. For example, we actually show how the eigenvectors corresponding to the largest distinct eigenvalues, whether these are defective or not, can be approximated solely in terms of the vectors $u_j$, and provide precise rates of convergence for them. The key to these results is the observation that the vector-valued power series $\sum_{m=0}^{\infty} u_m z^m$ actually represents a vector-valued *rational* function $F(z)$ whose poles are the reciprocals of some or all of the nonzero eigenvalues of $A$, depending on the spectral decomposition of $u_0$, and that corresponding eigenvectors (and certain combinations of eigenvectors and principal vectors) are related to corresponding principal parts of the Laurent expansions of the function $F(z)$. The main results of Section 3 pertaining to eigenvalues are given in Theorem 3.1, while those pertaining to eigenvectors and invariant subspaces are given in Theorem 3.2 and the subsequent paragraphs. A detailed description of the properties of the power iterations $u_m = Au_{m-1}, m = 1, 2, ...$, is provided

in Section 2.

In Section 4 we present a short review of general projection methods and Krylov subspace methods for the matrix eigenvalue problem. Of particular interest to us are the methods of Arnoldi [A] and Lanczos [L], which are described in some detail in this section.

In Section 5 we show that the extensions of the power method developed and analyzed in Section 3 are very closely related to Krylov subspace methods. In particular, we show that the reciprocals of the poles and the corresponding residues of the rational approximations $F_{n,k}(z)$ obtained from the SMPE, SMMPE, and STEA procedures are the Ritz values and the Ritz vectors, respectively, of certain Krylov subspace methods for the matrix $A$ starting with the power iteration $u_n$. Specifically, the methods of Arnoldi and Lanczos are related to the $F_{n,k}(z)$ obtained from the SMPE and STEA procedures, respectively, precisely in this sense. The main results of Section 5 are summarized in Theorem 5.4 and Corollary 5.5.

Now the Ritz values and Ritz vectors obtained from Krylov subspace methods are normally used as approximations to nondefective eigenpairs. They are not very effective for defective eigenpairs. Since we know that the generalized power methods based on the SMPE, SMMPE, and STEA procedures are related to Krylov subspace methods, the constructions for approximating defective eigenvalues and their corresponding invariant subspaces that originate from generalized power methods and that are given in Section 3 are entirely new as far as Krylov subspace methods are concerned. Similarly, all of the convergence results of Section 3, whether they pertain to defective or nondefective eigenvalues and their corresponding invariant subspaces, are new and totally different from the known analyses provided by Kaniel [K], Paige[Pai], and Saad[Sa1,Sa2]. Some of these analyses can also be found in Parlett [Par2] and Golub and Van Loan[GV]. The last two references also give a very thorough treatment of the computational aspects of Krylov subspace methods.

In Section 6 we show how the Ritz values and Ritz vectors obtained in a stable way from the common implementations of the Aroldi and Lanczos methods can be used in constructing the approximations to the defective eigenvalues and their corresponding invariant subspaces in general and eigenvectors in particular.

In view of the connection between the Krylov subspace methods and vector-valued rational approximations of [Si6] and the corresponding generalized power methods of the present work, we can summarize the main contributions of this paper as follows:

(i) It is shown that Krylov subspace methods for the matrix eigenvalue problem have a basis in analytic function theory and in rational approximations in the complex plane.

(ii) A mode of usage of Krylov subspace methods akin to the power method, in which one first iterates on an arbitrary initial vector many times and only then applies Krylov subspace methods, is proposed. This mode produces approximations only to the largest eigenvalues and their corresponding invariant subspaces.

(iii) The output from Krylov subspace methods, namely, the Ritz values and Ritz vectors, are used in constructing optimal approximations to defective eigenvalues and the corresponding eigenvectors and invariant subspaces. (The Ritz values and Ritz vectors by themselves are not good approximations to defective eigenvalues and corresponding subspaces.)

(iv) A *complete* convergence theory for the generalized power methods is provided.

(v) Numerical experience shows that in many cases the mode of usage proposed in this work and mentioned in (ii) above may produce the accuracy that is achieved by applying the Arnoldi method in the commonly known way using less storage and less computational work when the matrix being treated is large and sparse.

Before closing this section we would like to note that the eigenvalue problem for defective matrices has received some attention in the literature. The problem of approximating the largest eigenvalue of a matrix when this eigenvalue is defective has been considered by Ostrowski[O], who proposes an extension of the Rayleigh quotient and inverse iteration and gives a thorough analysis for this extension. Parlett and Poole [ParPo] consider the properties of a wide range of projection methods within the framework of defective matrices. The convergence of the QR method for defective Hessenberg matrices has been analyzed in detail by Parlett [Par1]. The problem of determining the Jordan canonical form of nondefective matrices has been treated in Golub and Wilkinson [GW]. The use of power iterations in approximating defective eigenvalues is also treated to some extent in Wilkinson [W, Chap. 7] and Householder [H, Chap. 7].

Finally, we mention that the results of [Si6] as well as the application of vector-valued rational approximations to the matrix eigenvalue problem were motivated by the developments in a recent work by the author [Si4] on the classical Padé approximants.

## 2 Properties of Power Iterations

Let $A$ be an $N \times N$ matrix, which, in general, is complex and nondiagonalizable. Let $u_0$ be a given arbitrary vector in $C^N$, and generate the vectors $u_1, u_2, ...$, according to

$$u_{j+1} = Au_j, \quad j \geq 0. \tag{2.1}$$

Denote by $s$ be index of $A$, i.e., the size of the largest Jordan block of $A$ with zero eigenvalue. Then $u_m$ is of the form

$$u_m = \sum_{j=1}^{M} \left[ \sum_{l=0}^{p_j} \tilde{a}_{jl} \binom{m}{l} \right] \lambda_j^m, \quad \text{for } m \geq s, \tag{2.2}$$

where $\lambda_j$ are some or all of the *distinct nonzero* eigenvalues of $A$, which we choose to order such that

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_M| > 0, \tag{2.3}$$

$p_j + 1 \equiv \omega_j$ are positive integers less than or equal to the dimension of the invariant subspace of $A$ belonging to the eigenvalue $\lambda_j$, and $\tilde{a}_{jl}, 0 \leq l \leq p_j$, are linearly independent vectors in this invariant subspace. It turns out that the vector $\tilde{a}_{jp_j}$ is an *eigenvector* of $A$ corresponding to $\lambda_j$, while the vectors $\tilde{a}_{ji}, \quad i = 0, 1, ..., p_j - 1$, are combinations of eigenvectors *and* principal vectors of $A$ corresponding to the eigenvalue $\lambda_j$. What is more, the subspaces

$$Y_i = \text{span } \{\tilde{a}_{jl}, i \leq l \leq p_j\}, \quad i = 0, 1, ..., p_j,$$

are invariant subspaces of $A$ corresponding to the eigenvalue $\lambda_j$, and satisfy $Y_0 \supset Y_1 \supset ... \supset Y_{p_j}$.

Whether all distinct nonzero eigenvalues are present among $\lambda_1, \lambda_2, ..., \lambda_M$, the exact values of the $\omega_j$ , and the precise composition of the vectors $\tilde{a}_{jl}$, all depend on the spectral decomposition of the initial vector $u_0$. For a detailed derivation of the above see [SiB, Section 2].

Before we go on, we will only mention how to determine the maximum value that $\omega_j$ can assume. Suppose that the Jordan canonical form of $A$ has several Jordan blocks whose eigenvalues are all

equal to $\lambda_j$. Then the largest value that $\omega_j$ can assume is the size of the largest of these blocks. In general, for a randomly chosen vector $u_0$, $\omega_j$ will take on its maximum value. In cases $\omega_j$ is theoretically less than this maximum value, rounding errors on a computer will ultimately force $\omega_j$ to take on its maximum value.

It is obvious from the above that

$$k_0 \equiv \sum_{j=1}^{M}(p_j + 1) = \sum_{j=1}^{M} \omega_j \leq N \tag{2.4}$$

and

$$\tilde{a}_{ji},\ 0 \leq i \leq p_j,\ 1 \leq j \leq M, \quad \text{are linearly independent.} \tag{2.5}$$

Also the minimal polynomial of the matrix $A$ with respect to the vector $u_s$ has degree $k_0 = \sum_{j=1}^{M} \omega_j$, i.e.,

$$k_0 = \min\left\{k:\ \left(\sum_{i=0}^{k} \beta_i A^i\right) u_s = 0,\ \beta_k = 1\right\}.$$

If defined as a monic polynomial, this polynomial is unique and divides the minimal polynomial of $A$, which, in turn, divides the characteristic polynomial of $A$. Furthermore, the minimal polynomial of $A$ with respect to $u_s$ is also the minimal polynomial of $A$ with respect to $u_m$ for all $m \geq s$. Consequently, any set of vectors $\{u_m, u_{m+1}, \ldots, u_{m+k}\}$ is linearly independent for $m \geq s$ provided $k < k_0$.

Applying now Lemma 3.1 of [Si6] in conjunction with (2.2), we conclude that the vector-valued power series $\sum_{m=0}^{\infty} u_m z^m$ represents the vector-valued *rational* function

$$F(z) = \sum_{j=1}^{M} \sum_{i=0}^{p_j} \frac{a_{ji}}{(1 - \lambda_j z)^{i+1}} + G(z), \tag{2.6}$$

in which the vectors $a_{ji}$ are uniquely determined in terms of the $\tilde{a}_{jl}$ from

$$\tilde{a}_{jl} = \sum_{i=l}^{p_j} a_{ji} \binom{i}{i-l},\quad 0 \leq l \leq p_j,\quad 1 \leq j \leq M, \tag{2.7}$$

and hence form a linearly independent set, and $G(z)$ is a vector-valued polynomial of degree at most $s - 1$. In fact, $G(z)$ is in the invariant subspace of $A$ corresponding to the zero eigenvalue. Also, $a_{jp_j} = \tilde{a}_{jp_j}$, i.e., $a_{jp_j}$ is an eigenvector of $A$ corresponding to the eigenvalue $\lambda_j$, while for each $i, 0 \leq i \leq p_j - 1$, $a_{ji}$ is some other vector in the invariant subspace $Y_i$ corresponding to the eigenvalue $\lambda_j$, and involves principal vectors as well as eigenvectors.

When the matrix $A$ is diagonalizable, $p_j = 0$ for all $j$ in (2.2) and hence in (2.6). If, in addition, $A$ is normal, then its eigenvectors form an orthogonal set with respect to the standard Euclidean inner product, namely, $(x, y) = x^*y$, where $x^*$ stands for the hermitian conjugate of $x$. Consequently, the vectors $\tilde{a}_{j0} = a_{j0}$ in (2.2) and (2.6) are orthogonal with respect to this inner product when $A$ is normal.

Now that we have shown that the power series $\sum_{m=0}^{\infty} u_m z^m$ represents a rational function $F(z)$ that is analytic at $z = 0$ and has poles $z_j = \lambda_j^{-1}$ of respective multiplicities $\omega_j = p_j + 1$, $j = 1, 2, ..., M$, we can apply any one of the approximation procedures SMPE, SMMPE, or STEA to the power series $\sum_{m=0}^{\infty} u_m z^m$ to obtain the vector-valued rational approximations $F_{n,k}(z)$ to $F(z)$. We can then apply the theorems of Sections 4 and 5 of [Si6] to construct approximations to the eigenvalues $\lambda_j$ and the vectors $a_{ji}$ in (2.6) and (2.7).

It is important to note that the linear independence of the vectors $a_{jl}$ is an important condition for the convergence of the SMPE and SMMPE procedures, but is not needed for the STEA procedure. In addition, we assume throughout that

$$\begin{vmatrix} (q_1, a_{10}) & \cdots & (q_1, a_{1p_1}) & \cdots & (q_1, a_{t0}) & \cdots & (q_1, a_{tp_t}) \\ \vdots & & \vdots & & \vdots & & \vdots \\ (q_k, a_{10}) & \cdots & (q_k, a_{1p_1}) & \cdots & (q_k, a_{t0}) & \cdots & (q_k, a_{tp_t}) \end{vmatrix} \neq 0 \quad \text{for SMMPE,} \tag{2.8}$$

and that

$$\prod_{j=1}^{t} (q, a_{jp_j}) \neq 0 \quad \text{for STEA.} \tag{2.9}$$

No additional assumption is needed for SMPE.

In order for (2.8) to hold it is necessary (but not sufficient) that the two sets of vectors $\{a_{ji} : 0 \leq i \leq p_j, 1 \leq j \leq t\}$ and $\{q_1, ..., q_k\}$, each be linearly independent, as has already been assumed.

# 3 Theoretical Development of Generalized Power Methods

In light of the developments of the previous section and Theorems 4.1, 4.3, and 4.5 of [Si6] and the developments of Section 5 in the same paper, we approach the matrix eigenvalue problem as

follows:

Given the vector $u_0$ that is picked arbitrarily, we generate the vectors $u_1, u_2, ...$, according to (2.1). We then fix the integers $n$ and $k$, and determine the coefficients $c_j^{(n,k)}$, $j = 0, 1, ..., k$, of the denominator polynomial of $F_{n,k}(z)$ for one of the procedures SMPE, SMMPE, and STEA, by using $u_n, u_{n+1}, ..., u_{n+k}$ for SMPE and SMMPE, and $u_n, u_{n+1}, ..., u_{n+2k-1}$, for STEA. By Theorem 4.1 of [Si6] the zeros of the polynomial $\hat{Q}_{n,k}(\lambda) \equiv \lambda^{-k} Q_{n,k}(\lambda^{-1}) = \sum_{j=0}^{k} c_j^{(n,k)} \lambda^j$ are approximations to the $k$ largest $\lambda_j$ in (2.2), counted according to their multiplicities $\omega_j$, provided the conditions stated in this theorem are satisfied. In case the matrix $A$ is normal, the zeros of the polynomial $\hat{Q}_{n,k}(\lambda)$, obtained from SMPE and STEA with the standard Euclidean inner product, are even better approximations to the eigenvalues $\lambda_j$ of $A$ as follows from Theorem 4.5 of [Si6].

## 3.1 Treatment of Eigenvalue Approximations

Theorem 3.1 below, which is of constructive nature, summarizes all the relevant results concerning the approximations to the $\lambda_j$. The corresponding approximations to eigenvectors and other vectors in the invariant subspaces are subsequently obtained with the help of the developments in Section 5 of [Si6], and the relevant results for this problem are summarized in Theorem 3.2 below.

We note that we have adopted in this section all of the notation of the previous sections.

**Theorem 3.1:** *Let the matrix $A$ and the vector sequence $u_m, m = 0, 1, 2, ...$, be as described in the preceding section. Let the positive integers $t$ and $k$ be such that*

$$|\lambda_t| > |\lambda_{t+1}| \quad \text{and} \quad k = \sum_{j=1}^{t} (p_j + 1) = \sum_{j=1}^{t} \omega_j. \tag{3.1}$$

*Determine the coefficients $c_j^{(n,k)}, j = 0, 1, ..., k$, for one of the procedures $SMPE, SMMPE$, and $STEA$, by utilizing $u_n, u_{n+1}, ...$, as described in (1.4) and (1.5). Then, under the additional conditions given in (2.8) and (2.9),*

$$\hat{Q}_{n,k}(\lambda) \equiv \sum_{j=0}^{k} c_j^{(n,k)} \lambda^j = \prod_{j=1}^{t} (\lambda - \lambda_j)^{\omega_j} + O(\epsilon(n)) \quad \text{as } n \to \infty, \tag{3.2}$$

*where*

$$\epsilon(n) = n^{\alpha} \left| \frac{\lambda_{t+1}}{\lambda_t} \right|^n, \tag{3.3}$$

*$\alpha$ being some nonnegative integer. In fact, if the $\lambda_j$ whose moduli are $|\lambda_t|$ are simple, then $\alpha = \overline{p}$, where $\overline{p} = \max\{p_j : |\lambda_j| = |\lambda_{t+1}|\}$. Consequently, the polynomial $\hat{Q}_{n,k}(\lambda)$, for $n \to \infty$, has $\omega_j$ zeros $\lambda_{jl}(n)$, $1 \le l \le \omega_j$, that tend to $\lambda_j$, $j = 1, 2, ..., t$. For each $j$ and $l$ we have*

$$\lambda_{jl}(n) - \lambda_j = O\left(\delta_j(n)^{1/\omega_j}\right) \quad as\ n \to \infty, \tag{3.4}$$

*where*

$$\delta_j(n) = n^{\overline{p}} \left| \frac{\lambda_{t+1}}{\lambda_j} \right|^n. \tag{3.5}$$

*Let us denote*

$$\hat{\lambda}_j(n) = \frac{1}{\omega_j} \sum_{l=1}^{\omega_j} \lambda_{jl}(n) \quad or \quad \hat{\lambda}_j(n) = \left[ \frac{1}{\omega_j} \sum_{l=1}^{\omega_j} \lambda_{jl}(n)^{-1} \right]^{-1}. \tag{3.6}$$

*Then*

$$\hat{\lambda}_j(n) - \lambda_j = O(\delta_j(n)) \quad as\ n \to \infty. \tag{3.7}$$

*Also, the $p_j$th derivative of $\hat{Q}_{n,k}(\lambda)$ has exactly one zero $\tilde{\lambda}_j(n)$ that tends to $\lambda_j$ and satisfies*

$$\tilde{\lambda}_j(n) - \lambda_j = O(\delta_j(n)) \quad as\ n \to \infty. \tag{3.8}$$

*Let the matrix $A$ be normal, i.e., $AA^* = A^*A$. Then $p_j = 0$ hence $\omega_j = 1$ for all $j$. If the $c_j^{(n,k)}$ are determined through the procedures SMPE and STEA with the standard Euclidean inner product, and $k$ is such that*

$$|\lambda_k| > |\lambda_{k+1}|, \tag{3.9}$$

*and provided $q = u_n$ for STEA, then (6.8) and (6.10) are substantially improved to read, respectively,*

$$\hat{Q}_{n,k}(\lambda) = \prod_{j=1}^{k} (\lambda - \lambda_j) + O\left( \left| \frac{\lambda_{k+1}}{\lambda_k} \right|^{2n} \right) \quad as\ n \to \infty, \tag{3.10}$$

*and, for $j = 1, ..., k$,*

$$\lambda_j(n) - \lambda_j = O\left( \left| \frac{\lambda_{k+1}}{\lambda_j} \right|^{2n} \right) \quad as\ n \to \infty, \tag{3.11}$$

*where $\lambda_j(n)$ is the unique zero of $\hat{Q}_{n,k}(\lambda)$ that tends to $\lambda_j$.*

We would like to note again that the result in (3.2) and (3.3) was originally given in [SiB, Section 6, Theorem 6.1], and those in (3.10) and (3.11) were originally given for SMPE in [Si3]. The rest of Theorem 6.1 is new. As mentioned in these papers, the methods contained in Theorem 3.1 are true extensions of the classical power method.

One important aspect of Theorem 3.1 is the construction of *optimal* approximations to *defective* eigenvalues through (3.6) and (3.7). From (3.4) it is clear that when $p_j = 0$ hence $\omega_j = 1$, which occurs automatically if $\lambda_j$ is a nondefective eigenvalue, the rate of convergence of the approximation corresponding to $\lambda_j$ is optimal. In case $\lambda_j$ is a defective eigenvalue and $p_j > 0$, the rate of convergence of each of its $\omega_j$ corresponding approximations is $1/\omega_j$ of the optimal rate. For this case (3.6) and (3.7) show how the poor approximations $\lambda_{ji}(n)$ can be combined in a simple way to give an optimal approximation, namely $\hat{\lambda}_j(n)$. Similarly, (3.8) shows that $\tilde{\lambda}_j(n)$, the zero of the $p_j$th derivative of $\hat{Q}_{n,k}(\lambda)$ that tends to $\lambda_j$, has the same optimal convergence rate as $\hat{\lambda}_j(n)$. The results in (3.10) and (3.11) show that the approximations obtained from SMPE and STEA for a normal matrix converge twice as fast as those obtained for a nonnormal diagonalizable matrix having the same spectrum.

## 3.2 Treatment of Invariant Subspace Approximations

For the treatment of the eigenvectors and invariant subspaces we need some preliminary work.

Let us rewrite (2.6) in the form

$$F(z) = \sum_{j=1}^{M} \sum_{i=0}^{p_j} \frac{d_{ji}}{(z - z_j)^{i+1}} + G(z), \tag{3.12}$$

where

$$z_j = \lambda_j^{-1} \quad \text{and} \quad d_{ji} = (-z_j)^{i+1} a_{ji} \quad \text{for all } j, i. \tag{3.13}$$

Thus the $d_{ji}$ are the coefficients of the principal part of the Laurent expansion of $F(z)$ about the pole $z_j$, $j = 1, ..., M$.

Consider the rational function

$$\hat{F}(z) = \frac{F(z) - F_{n+\nu}(z)}{z^{n+\nu+1}}, \tag{3.14}$$

which is analytic at $z = 0$ and has the Maclaurin series expansion

$$\hat{F}(z) = \sum_{i=0}^{\infty} u_{n+\nu+i+1} z^i. \tag{3.15}$$

By (3.12) we can write

$$\hat{F}(z) = \sum_{i=0}^{p_j} \frac{\hat{d}_{ji}}{(z - z_j)^{i+1}} + \hat{G}_j(z), \tag{3.16}$$

where

$$\hat{d}_{ji} = z_j^{-n-\nu-1} \sum_{l=i}^{p_j} \begin{pmatrix} -n-\nu-1 \\ l-i \end{pmatrix} z_j^{-l+i} d_{jl}, \tag{3.17}$$

and $\hat{G}_j(z)$ is analytic at $z_j$, i.e., as above, the $\hat{d}_{ji}$ are coefficients of the principal part of the Laurent expansion of $\hat{F}(z)$ about the pole $z_j, j = 1, ..., M$. Unlike before, both $\hat{F}(z)$ and the $\hat{d}_{ji}$ depend on $n$, in addition. The vector $\hat{d}_{jp_j}$, being a scalar multiple of the constant vector $d_{jp_j}$, is an eigenvector of $A$ corresponding to the eigenvalue $\lambda_j$. For $i \neq p_j$, the vector $\hat{d}_{ji}$, being a linear combination of the constant vectors $d_{jl}, i \leq l \leq p_j$, is in the invariant subspace $Y_i$, and, as is obvious from (3.17), the coefficients of the $d_{jl}$ in this linear combination are polynomials in $n$, up to the common multiplicative factor $z_j^{-n-\nu-1}$.

Following now the developments in Section 5 of [Si6], we obtain the following constructive result for the $\hat{d}_{ji}$.

**Theorem 3.2:** *With the notation and conditions of Theorem 3.1, let us define, for* $1 \leq j \leq t$,

$$\zeta_j(n) = 1/\hat{\lambda}_j(n) \text{ or } \zeta_j(n) = 1/\bar{\lambda}_j(n), \tag{3.18}$$

*and, for* $0 \leq i \leq p_j$ *and* $1 \leq l \leq \omega_j$,

$$\hat{d}_{ji,l}(n) = (z - \zeta_j(n))^i \left. \frac{\sum_{r=1}^{k} c_r^{(n,k)} z^{k-r} \sum_{m=1}^{r} u_{n+\nu+m} z^{m-1}}{\sum_{r=0}^{k} c_r^{(n,k)} (k-r) z^{k-r-1}} \right|_{z=1/\lambda_{jl}(n)} \tag{3.19}$$

*and*

$$\hat{d}_{ji}(n) = \sum_{l=1}^{\omega_j} \hat{d}_{ji,l}(n). \tag{3.20}$$

*Then, for* $0 \leq i \leq p_j$, $\hat{d}_{ji}(n)$ *is an approximation to* $\hat{d}_{ji}$ *in (3.17) in the sense*

$$\limsup_{n\to\infty} \left| \hat{d}_{ji}(n) - \hat{d}_{ji} \right|^{1/n} \leq \left| \lambda_{t+1} \right|. \tag{3.21}$$

We would like to note that Theorem 3.2 actually contains the basic ingredients of a potentially bona fide numerical method for approximating the eigenvectors and other vectors in invariant subspaces corresponding to largest eigenvalues of $A$. The resulting method, which is described below, (i) makes use of only $u_n, u_{n+1}, ...,$ disregarding $u_0, u_1, ..., u_{n-1}$ entirely, and (ii) enables us to construct optimal approximations to the vectors $a_{ji}, 0 \leq i \leq p_j$, for $p_j = 0$ as well as $p_j > 0$. We now turn to these constructions.

### 3.2.1 Approximation of the Eigenvector $a_{jp_j}$

Let us first specialize the result of Theorem 3.2 to the case $i = p_j$. We have

$$\hat{d}_{jp_j} = \lambda_j^{n+\nu+1} d_{jp_j}, \tag{3.22}$$

so that (3.21) can also be written as

$$\limsup_{n\to\infty} \left|\lambda_j^{-n-\nu-1}\hat{d}_{jp_j}(n) - d_{jp_j}\right|^{1/n} \le \left|\frac{\lambda_{t+1}}{\lambda_j}\right|. \tag{3.23}$$

This clearly shows that the vector $\hat{d}_{jp_j}(n)$, as $n \to \infty$, aligns itself with the *constant* vector $d_{jp_j}$, which is proportional to the eigenvector $a_{jp_j}$, practically at the rate of $|\lambda_{t+1}/\lambda_j|^n$. It is thus sufficient to compute the vectors $\hat{d}_{ji,l}(n)$, $1 \le l \le \omega_j$, by (3.19), and then to form $\hat{d}_{ji}(n)$ by (3.20) as our approximation to the (appropriately normalized) eigenvector $a_{jp_j}$, and this is valid whether $p_j = 0$ or $p_j > 0$.

### 3.2.2 Approximation of the Vectors $a_{ji}$, $0 \le i \le p_j - 1$

Although the vector $a_{jp_j}$ (up to a multiplicative constant) can be determined from $\hat{d}_{jp_j}(n)$ in a rather painless manner, the determination of the remaining $a_{ji}$ from the $\hat{d}_{jl}(n)$ becomes somewhat involved. The reason for this is that the vectors $\hat{d}_{ji}$, apart from the scalar multiplicative factor $z_j^{-n-\nu-1}$, are linear combinations of the $d_{jl}$ hence of the $a_{jl}, i \le l \le p_j$, with coefficients that vary as functions of $n$, as can be seen from (3.17) and (3.13), and as has been mentioned before. This means that the $\hat{d}_{ji}$ do *not* have a fixed direction with varying $n$.

Let us now rewrite (3.17) in the form

$$T(n)\begin{bmatrix} d_{j0} \\ d_{j1} \\ \vdots \\ d_{jp_j} \end{bmatrix} = z_j^{n+\nu+1}\begin{bmatrix} \hat{d}_{j0} \\ \hat{d}_{j1} \\ \vdots \\ \hat{d}_{jp_j} \end{bmatrix}, \tag{3.24}$$

where $T(n)$ is the upper triangular matrix

$$T(n) = \begin{bmatrix} \tau_{00} & \tau_{01} & \cdots & \tau_{0p_j} \\ & \tau_{11} & \cdots & \tau_{1p_j} \\ & & & \vdots \\ & & & \tau_{p_jp_j} \end{bmatrix}, \quad \tau_{il} = \begin{pmatrix} -n-\nu-1 \\ l-i \end{pmatrix} z_j^{-l+i} \text{ all } i \text{ and } l. \tag{3.25}$$

Obviously, $T(n)$ is invertible since its diagonal elements are unity. Thus,

$$\begin{bmatrix} d_{j0} \\ d_{j1} \\ \vdots \\ d_{jp_j} \end{bmatrix} = T(n)^{-1} \begin{bmatrix} \hat{d}_{j0} \\ \hat{d}_{j1} \\ \vdots \\ \hat{d}_{jp_j} \end{bmatrix} z_j^{n+\nu+1}, \tag{3.26}$$

where $T(n)^{-1}$ is also upper triangular, its diagonal elements being unity.

Now since all elements of $T(n)$ are polynomials in $n$, and since its determinant is unity, the elements of $T(n)^{-1}$ turn out to be polynomials in $n$, i.e., the matrix $T(n)^{-1}$ can grow at most polynomially as $n \to \infty$. If we denote the nonzero elements of $T(n)^{-1}$ by $\rho_{il}, i \le l \le p_j, 0 \le i \le p_j$, then we can write (3.26) in the form

$$d_{ji} = z_j^{n+\nu+1} \sum_{l=i}^{p_j} \rho_{il}\ \hat{d}_{jl}, \quad 0 \le i \le p_j. \tag{3.27}$$

Let us replace $\hat{d}_{jl}$ in (3.27) by $\left[\left(\hat{d}_{jl} - \hat{d}_{jl}(n)\right) + \hat{d}_{jl}(n)\right]$, and invoke (3.21). After some manipulation we obtain

$$\limsup_{n\to\infty} \left| d_{ji} - z_j^{n+\nu+1} \sum_{l=i}^{p_j} \rho_{il}\ \hat{d}_{jl}(n) \right|^{1/n} \le \left| \frac{\lambda_{t+1}}{\lambda_j} \right|. \tag{3.28}$$

This implies that the vector $\sum_{l=i}^{p_j} \rho_{il}\ \hat{d}_{jl}(n)$ aligns itself with the fixed vector $d_{ji}$ as $n \to \infty$ practically at the rate of $|\lambda_{t+1}/\lambda_j|^n$. We leave the details of the proof of (3.28) to the reader.

We note that (3.28) shows how to construct a good approximation to $d_{ji}$ from the $\hat{d}_{jl}(n)$ and $\lambda_j$, provided $\lambda_j$ is known. Since $\lambda_j$ is not known, however, the vector $\sum_{l=i}^{p_j} \rho_{il}\hat{d}_{jl}(n)$ cannot be constructed. We, therefore, propose to replace $\lambda_j$ in the matrix $T(n)^{-1}$ by the known approximations $\zeta_j(n)$. Also, in this case, it can be shown that (3.28) remains valid. Again, we leave the details of the proof to the reader.

Before closing this section, we must mention that the developments of this section are meant to be theoretical, in general. Although they can be used for computational purposes for small values of $k$, their use for large $k$ is likely to introduce numerical instabilities in many cases. These instabilities are mainly a result of our direct use of the power iterations $u_{n+i} = A^i u_n$, $i = 0, 1, \ldots$. They exhibit themselves first of all through the poor computed approximations to the $\lambda_j$, which ultimately affect the computed eigenvector approximations. This problem can be remedied by observing that the

approximations $F_{n,k}(z)$ that we developed and applied to the matrix eigenvalue problem are very tightly connected with Krylov subspace methods for some of which there exist computationally stable implementations. In particular, the SMPE and STEA procedures are related to the method of Arnoldi and the method of Lanczos, respectively, as we show in detail in the next two sections.

# 4 General Projection Methods and the Methods of Arnoldi and Lanczos for the Matrix Eigenproblem

## 4.1 General Projection Methods

Let $\{v_1, ..., v_k\}$ and $\{w_1, ..., w_k\}$ be two linearly independent sets of vectors in $C^N$, and define the $N \times k$ matrices $V$ and $W$ by

$$V = [v_1|v_2|\cdots|v_k] \quad \text{and} \quad W = [w_1|w_2|\cdots|w_k]. \tag{4.1}$$

In addition, let us agree to denote the subspaces span $\{v_1, ..., v_k\}$ and span $\{w_1, ..., w_k\}$ by $V$ and $W$, respectively.

In projection methods one looks for an approximate eigenvalue-eigenvector pair $(\lambda, x)$ with $x \in V$ that satisfies the condition

$$(y, Ax - \lambda x) = 0 \quad \text{for all } y \in W, \tag{4.2}$$

which can also be written in the equivalent form

$$W^*(A - \lambda I)V\xi = 0 \quad \text{for some } \xi \in C^k. \tag{4.3}$$

Here we have used the fact that $x \in V$ implies that $x = V\xi$ for some $\xi \in C^k$. Of course, (4.3) holds if and only if $\lambda$ is an eigenvalue of the matrix pencil $(W^*AV, W^*V)$, i.e., it satisfies the characteristic equation

$$\det(W^*AV - \lambda W^*V) = 0. \tag{4.4}$$

In general, (4.4) has $k$ solutions for $\lambda$, which are known as *Ritz values* in the literature. Given that $\lambda'$ is a Ritz value, the corresponding eigenvector $\xi'$ is a solution of the homogeneous system in (4.3). The eigenvector approximation corresponding to $\lambda'$ is now $x' = V\xi'$, and is known as a *Ritz vector.*

The different projection methods are characterized by the subspaces $V$ and $W$ that they employ. (Note that $V$ and $W$ are also called, respectively, the right and left subspaces.)

## 4.2 The Method of Arnoldi

In this method $V$ and $W$ are Krylov subspaces given by

$$V = V_{k-1} = \text{ span } \{u_0, Au_0, ..., A^{k-1}u_0\} \text{ and } W = W_{k-1} = V_{k-1}, \tag{4.5}$$

for some arbitrary vector $u_0$.

Arnoldi has given a very successful implementation of this method. In this implementation the vectors $A^i u_0, i = 0, 1, ...,$ are orthogonalized by a very special Gram-Schmidt process as follows:

Step 0. Let $v_1 = u_0/||u_0||$

Step 1. For $j = 1, ..., k-1$, do

Determine the scalar $h_{j+1,j} > 0$ and the vector $v_{j+1}$, such that $||v_{j+1}|| = 1$ and

$h_{j+1,j}v_{j+1} = Av_j - \sum_{i=1}^{j} h_{ij}v_i, \quad h_{ij} = (v_i, Av_j), \ 1 \le i \le j.$

(4.6)

Thus the $N \times k$ matrix $V = [v_1|v_2|\cdots|v_k]$ is unitary in the sense that $V^*V$ is the $k \times k$ identity matrix. As a result, $W^*V = V^*V = I$, and the generalized eigenvalue problem of (4.3) now becomes

$$H\xi = \lambda\xi, \tag{4.7}$$

where $H$ is the $k \times k$ upper Hessenberg matrix

$$H = \begin{bmatrix} h_{11} & h_{12} & \cdots & & h_{1k} \\ h_{21} & h_{22} & \cdots & & h_{2k} \\ & h_{32} & \cdots & & h_{3k} \\ & & \ddots & & \vdots \\ & & & h_{k,k-1} & h_{kk} \end{bmatrix}, \tag{4.8}$$

i.e., the Ritz values are the eigenvalues of $H$.

## 4.3 The Method of Lanczos

In this method $V$ and $W$ are the Krylov subspaces

$$V = V_{k-1} = \text{ span } \{u_0, Au_0, ..., A^{k-1}u_0\} \text{ and } W = W_{k-1} = \text{ span } \{q, A^*q, ..., (A^*)^{k-1}q\}, \tag{4.9}$$

for some arbitrary vectors $u_0$ and $q$.

The algorithm given by Lanczos generates one set of vectors $\{v_1, ..., v_k\}$ from the $A^i u_0$, $i = 0, 1, ..., k-1$, and another set of vectors $\{w_1, ..., w_k\}$ from the $(A^*)^i q$, $i = 0, 1, ..., k-1$, that satisfy the biorthogonality condition

$$w_i^* v_j = \delta_{ij}, \tag{4.10}$$

as long as the process does not break down. This is achieved by the following Algorithm:

Step 0. Set $v_1 = \sigma u_0$ and $w_1 = \tau q$ such that $(w_1, v_1) = 1$.

Step 1. For $j = 1, ..., k-1$, do

(a) Compute $\hat{v}_{j+1}$ and $\hat{w}_{j+1}$ by

$\hat{v}_{j+1} = Av_j - \alpha_j v_j - \beta_j v_{j-1}$

$\hat{w}_{j+1} = A^* w_j - \overline{\alpha}_j w_j - \overline{\delta}_j w_{j-1}$

(when $j = 1$ take $\beta_1 v_0 = \overline{\delta}_1 w_0 = 0$)

with $\alpha_j = (w_j, Av_j)$

(b) Choose $\delta_{j+1}$ and $\beta_{j+1}$ such that

$\delta_{j+1}\beta_{j+1} = (\hat{w}_{j+1}, \hat{v}_{j+1})$

and set

$v_{j+1} = \hat{v}_{j+1}/\delta_{j+1}$ and $w_{j+1} = \hat{w}_{j+1}/\overline{\beta}_{j+1}$.

(4.11)

By (4.10) the matrices $V$ and $W$ satisfy $W^*V = I$. As a result, the generalized eigenvalue problem of (4.3) becomes

$$H\xi = \lambda\xi, \tag{4.12}$$

where $H$ is the $k \times k$ tridiagonal matrix

$$H = \begin{bmatrix} \alpha_1 & \beta_2 & & & & \\ \delta_2 & \alpha_2 & \beta_3 & & & \\ & \delta_3 & \alpha_4 & \beta_4 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & & \beta_k \\ & & & & \delta_k & \alpha_k \end{bmatrix}, \tag{4.13}$$

and the Ritz values are the eigenvalues of $H$.

### 4.4 The Case of Hermitian $A$

The subspaces $V$ in (4.5) and (4.9) are identical. When $A$ is hermitian, i.e., $A^* = A$, and $q = u_0$, the subspaces $W$ in (4.5) and (4.9) become identical too. Thus the methods of Arnoldi and Lanczos become equivalent for the case under consideration. Furthermore, it can be shown that the elements $h_{ij}$ of the matrix $H$ in the method of Arnoldi satisfy $\overline{h_{i,i+1}} = h_{i+1,i}$ so that $h_{i,i+1} = h_{i+1,i} > 0$ for $i = 1, 2, ..., k-1$, while $h_{ij} = 0$ for $j \geq i+2$. The diagonal elements $h_{ii}$ are all real. That is to say, in the absence of roundoff, the matrix $H$ is real symmetric tridiagonal. If we pick $q = u_0$ and choose $\delta_j = \beta_j = \sqrt{(\hat{v}_j, \hat{v}_j)}$ in the method of Lanczos, then the matrix $H$ in (4.13) turns out to be real symmetric and is exactly the same as the one produced by the method of Arnoldi.

The properties of the Ritz values and Ritz vectors of the Lanczos method, as applied to hermitian matrices, have been analyzed by Kaniel [K], Paige[Pai], and Saad [Sa1]. The paper [Sa2] gives results for nonhermitian matrices.

## 5 Equivalence of Rational Approximation Procedures and Krylov Subspace Methods

We now go back to the rational approximation procedures SMPE, SMMPE, and STEA. In particular, we concentrate on the poles and residues of the rational functions $F_{n,k}(z)$.

### 5.1 Poles of $F_{n,k}(z)$ vs. Ritz Values

From the determinant representations of $F_{n,k}(z)$ that are given in Theorem 2.2 of [Si6], it follows that the denominator $Q_{n,k}(z)$ of $F_{n,k}(z)$ is a constant multiple of the determinant

$$D(\lambda) = \begin{vmatrix} 1 & \lambda & \cdots & \lambda^k \\ u_{00} & u_{01} & \cdots & u_{0k} \\ u_{10} & u_{11} & \cdots & u_{1k} \\ \vdots & \vdots & & \vdots \\ u_{k-1,0} & u_{k-1,1} & \cdots & u_{k-1,k} \end{vmatrix}, \tag{5.1}$$

where $\lambda = z^{-1}$ and $u_{ij}$ are as defined in (1.5). This implies that the zeros of the polynomial $D(\lambda)$ are the reciprocals of the zeros of $Q_{n,k}(z)$, or, equivalently, the reciprocals of the poles of $F_{n,k}(z)$.

In addition, they are the roots of a generalized eigenvalue problem as we show next.

**Theorem 5.1:** *Whatever the $u_{ij}$, the zeros of the polynomial $D(\lambda)$ in (5.1) are the eigenvalues of the matrix pencil $(X,T)$, where*

$$X = \begin{bmatrix} u_{01} & u_{02} & \cdots & u_{0k} \\ u_{11} & u_{12} & \cdots & u_{1k} \\ \vdots & \vdots & & \vdots \\ u_{k-1,1} & u_{k-1,2} & \cdots & u_{k-1,k} \end{bmatrix} \quad \text{and} \quad T = \begin{bmatrix} u_{00} & u_{01} & \cdots & u_{0,k-1} \\ u_{10} & u_{11} & \cdots & u_{1,k-1} \\ \vdots & \vdots & & \vdots \\ u_{k-1,0} & u_{k-1,1} & \cdots & u_{k-1,k-1} \end{bmatrix}, \tag{5.2}$$

*i.e., they satisfy the equation*

$$\det(X - \lambda T) = 0. \tag{5.3}$$

**Proof:** Multiply the $(j-1)$st column of $D(\lambda)$ by $\lambda$ and subtract from the $j$th column for $j = k+1, k, ..., 2$, in this order. This results in

$$D(\lambda) = \left| \begin{array}{c|c} 1 & 0 \cdots 0 \\ \hline u_{00} & \\ u_{10} & X - \lambda T \\ \vdots & \\ u_{k-1,0} & \end{array} \right| = \det(X - \lambda T), \tag{5.4}$$

thus proving the claim. □

When $u_{ij}$ are as in (1.5), Theorem 5.1 takes on the following interesting form.

**Theorem 5.2:** *Define the $N \times k$ matrices $V$ and $W$ by*

$$V = [u_n | u_{n+1} | \cdots | u_{n+k-1}] \tag{5.5}$$

*and*

$$\begin{aligned} W &= V && \text{for SMPE,} \\ W &= [q_1 | q_2 | \cdots | q_k] && \text{for SMMPE,} \\ W &= [q | A^* q | \cdots | (A^*)^{k-1} q] && \text{for STEA.} \end{aligned} \tag{5.6}$$

*Then, with $u_{ij}$ as defined by (1.5), the zeros of $D(\lambda)$ are the eigenvalues of the matrix pencil $(W^*AV, W^*V)$, i.e., they satisfy*

$$\det(W^*AV - \lambda W^*V) = 0. \tag{5.7}$$

*Consequently, the reciprocals of the poles of the rational approximations $F_{n,k}(z)$ obtained from the SMPE or SMMPE or STEA procedures are the Ritz values of the Krylov subspace methods whose right and left subspaces are column spaces of $V$ and $W$, respectively.*

**Proof:** Since Theorem 5.1 applies, all we need to show is that $X = W^*AV$ and $T = W^*V$ there. That $T = W^*V$ follows from (1.5), (5.2), (5.5), and (5.6). From (1.5), (5.2), and (5.6), we similarly have $X = W^*[u_{n+1}|\cdots|u_{n+k}]$. Using now the fact that $u_{j+1} = Au_j$, $j \geq 0$, we also have $[u_{n+1}|\cdots|u_{n+k}] = AV$. Consequently, $X = W^*AV$. Again, from $u_{j+1} = Au_j, j \geq 0$, we realize, in addition, that the right subspace for all three methods is none other than the Krylov subspace span $\{u_n, Au_n, ..., A^{k-1}u_n\}$. This completes the proof. □

## 5.2 Residues of $F_{n,k}(z)$ vs. Ritz Vectors

Turning Theorem 5.2 around, what we have is that the Ritz values obtained by applying the Krylov subspace methods whose left and right subspaces are column spaces of $V$ and $W$, respectively, are, in fact, the reciprocals of the poles of the corresponding rational approximations $F_{n,k}(z)$ to the meromorphic function $F(z) = \sum_{i=0}^{\infty} u_i z^i$. An immediate question that arises is, of course, whether there is any connection between the Ritz vectors and the $F_{n,k}(z)$. The answer, which is in the affirmative, is provided in Theorem 5.3 below.

**Theorem 5.3:** *Let $\hat{\lambda}$ be a Ritz value of the Krylov subspace methods whose right and left subspaces are column spaces of, respectively, $V$ and $W$ in Theorem 5.2. Denote the corresponding Ritz vector by $\hat{x}$. Let $\nu = -1$ in the corresponding rational approximation $F_{n,k}(z)$, cf. (1.2). Provided $\hat{\lambda}$ is simple, $\hat{x}$ is a constant multiple of the residue of $F_{n,k}(z)$ at the pole $\hat{z} = 1/\hat{\lambda}$.*

**Proof:** Let us first determine the residue of $F_{n,k}(z)$ at the pole $\hat{z} = 1/\hat{\lambda}$. With $\nu = -1$

$$\text{Res } F_{n,k}(z)|_{z=\hat{z}} = \frac{P_{n,k}(\hat{z})}{Q'_{n,k}(\hat{z})} = \frac{\sum_{r=0}^{k} c_r \hat{z}^{k-r} F_{n+r-1}(\hat{z})}{Q'_{n,k}(\hat{z})}, \tag{5.8}$$

since $Q'_{n,k}(\hat{z}) \neq 0$ that follows from the assumption that $\hat{\lambda}$ is simple, which implies that $\hat{z}$ is a simple pole. By $F_{n+s}(z) = F_{n-1}(z) + \sum_{m=n}^{n+s} u_m z^m$ and $\sum_{r=0}^{k} c_r \hat{z}^{k-r} = 0$, we can rewrite (5.8) in the form, cf. Section 5 of [Si6],

$$\text{Res } F_{n,k}(z)|_{z=\hat{z}} = \frac{1}{Q'_{n,k}(\hat{z})} \sum_{r=1}^{k} c_r \hat{z}^{k-r} \sum_{m=n}^{n+r-1} u_m \hat{z}^m = \frac{\hat{z}^{n+k-1}}{Q'_{n,k}(\hat{z})} \sum_{m=0}^{k-1} \eta_m u_{n+m}, \tag{5.9}$$

where

$$\eta_m = \sum_{r=m+1}^{k} c_r \hat{\lambda}^{r-m-1}, \quad m = 0, 1, ..., k-1. \tag{5.10}$$

Let us now denote $\eta = (\eta_0, \eta_1, ..., \eta_{k-1})^T$. Then (5.9) implies that Res $F_{n,k}(z)|_{z=\hat{z}}$ is a scalar multiple of $V\eta$. Recall that the Ritz vector corresponding to $\hat{\lambda}$ is $V\hat{\xi}$, where $\hat{\xi} \in C^k$ and satisfies $W^*(A - \hat{\lambda}I)V\hat{\xi} = 0$, which, on account of Theorem 5.2, is the same as $(X - \hat{\lambda}T)\hat{\xi} = 0$. Thus in order to show that Res$F_{n,k}(z)|_{z=\hat{z}}$ is a constant multiple of the Ritz vector corresponding to the Ritz value $\hat{\lambda}$, it is sufficient to show that

$$(X - \hat{\lambda}T)\eta = 0. \tag{5.11}$$

From (5.2), the $(i+1)$st component of the $k$-dimensional vector $\tau = (X - \hat{\lambda}T)\eta$, $i = 0, 1, ..., k-1$, is

$$\tau_i = \sum_{m=0}^{k-1} (u_{i,m+1} - \hat{\lambda}u_{im})\eta_m, \tag{5.12}$$

which, by (5.10), becomes

$$\tau_i = \sum_{m=0}^{k-1} (u_{i,m+1} - \hat{\lambda}u_{im}) \sum_{r=m+1}^{k} c_r \hat{\lambda}^{r-m-1}. \tag{5.13}$$

Expanding and rearranging this summation, we obtain

$$\tau_i = -u_{i0} \left( \sum_{r=1}^{k} c_r \hat{\lambda}^r \right) + \sum_{m=1}^{k} u_{im} c_m. \tag{5.14}$$

Recalling that $\sum_{r=0}^{k} c_r \hat{\lambda}^r = 0$, we can rewrite (5.14) as

$$\tau_i = \sum_{m=0}^{k} u_{im} c_m. \tag{5.15}$$

Finally, from the assumption that $c_k = 1$ and from the fact that $c_0, c_1, ..., c_{k-1}$ satisfy the linear equations in (1.4), we conclude that

$$\tau_i = 0, \quad i = 0, 1, ..., k-1. \tag{5.16}$$

This completes the proof. □

## 5.3 Summary of $F_{n,k}(z)$ vs. Krylov Subspace Methods

We now combine the results of Theorems 5.2 and 5.3 to state the following equivalence theorem, which forms the main result of this section, and one of the main results of this work.

**Theorem 5.4:** *Let $F_{n,k}(z)$ be the rational approximation obtained by applying the SMPE or SMMPE or STEA procedure to the vector-valued power series $\sum_{m=0}^{\infty} u_m z^m$, where $u_m = A^m u_0$, $m =$*

$0, 1, ...,$ *are power iterations. Denote the reciprocals of the poles of* $F_{n,k}(z)$ *by* $\lambda'_1, ..., \lambda'_k$. *Setting* $\nu = -1$ *in the numerator of* $F_{n,k}(z)$, *denote the corresponding residues of* $F_{n,k}(z)$ *by* $x'_1, ..., x'_k$. *Next, denote by* $\lambda''_1, ..., \lambda''_k$ *and* $x''_1, ..., x''_k$, *respectively, the Ritz values and corresponding Ritz vectors produced by the Krylov subspace methods whose right subspace is* span $\{u_n, Au_n, ..., A^{k-1}u_{n-1}\}$ *and left subspaces are the column spaces of the matrices* $W$ *in (5.6). Then*

$$\lambda'_j = \lambda''_j, \quad j = 1, ..., k, \tag{5.17}$$

*and*

$$x'_j \propto x''_j, \text{ provided } \lambda'_j \text{ is simple.} \tag{5.18}$$

More can be said about the SMPE and STEA procedures versus the methods of Arnoldi and Lanczos, and this is done in Corollary 5.5 below.

**Corollary 5.5:** *With* $F_{n,k}(z), \lambda'_j, x'_j$, $j = 1, ..., k$, *as in Theorem 5.4, let* $\lambda''_j, x''_j, j = 1, ..., k$, *be the Ritz values and Ritz vectors produced by applying the k-step Arnoldi or Lanczos methods to the matrix* $A$, *starting with the vector* $u_n = A^n u_0$. *(That is to say, replace the initial vector* $u_0$ *in Step 0 of (4.6) or (4.11) by the nth power iteration* $u_n$.*) In addition, let* $q$ *be the same vector for the STEA procedure and the Lanczos method. Then the SMPE and STEA procedures are equivalent to the methods of Arnoldi and Lanczos, respectively, precisely in the sense of (5.17) and (5.18).*

Now that we have shown the equivalence of the methods of Arnoldi and Lanczos with the generalized power methods based on the SMPE and STEA approximation procedures, we realize that those results that we proved in Section 3 for the latter and that pertain to the nondefective as well as defective eigenvalues of $A$ are, in fact, new results for the former. That is to say, if we apply the methods of Arnoldi or Lanczos to the matrix $A$ starting with the $n$th power iteration $u_n = A^n u_0$ for large $n$, then the Ritz values are approximations to the $k$ largest distinct eigenvalues of $A$ counted according to the multiplicities that appear in (2.2). Similarly, the Ritz vectors can be used for constructing the approximations to the corresponding invariant subspaces. These points will be considered in greater detail in the next section.

## 5.4 Optimality Properties of the Arnoldi Method

In Section 1 we mentioned that the coefficients of $c_i^{(n,k)}$ of the denominator polynomial $Q_{n,k}(z)$ of $F_{n,k}(z)$ for the SMPE procedure are the solution to the optimization problem given in (1.6). If

we now pick the vectors $u_m$ as the power iterations $u_m = A^m u_0$, $m = 0, 1, ...$, then (1.6) reads

$$\min_{c_0, c_1, ..., c_{k-1}} \left\| \left( \sum_{j=0}^{k-1} c_j A^j + A^k \right) u_n \right\|. \tag{5.19}$$

Exploiting the fact that the method of Arnoldi is equivalent to the generalized power method based on the SMPE approximation procedure, we can state the following optimality properties for the Arnoldi method as applied to a *general* matrix $A$.

**Theorem 5.5:** *Let $\lambda_j', x_j'$, $j = 1, 2, ..., k$, be the Ritz values and appropriately normalized Ritz vectors, respectively, produced by applying the $k$-step Arnoldi method to the matrix $A$ starting with the power iteration $u_n = A^n u_0$. Let $\mathcal{P}_k$ denote the set of monic polynomials of degree exactly $k$, while $\pi_k$ denotes the set of polynomials of degree at most $k$. Then for $k < k_0$, cf. (2.4),*

$$\left\| \left[ \prod_{i=1}^{k} (A - \lambda_i' I) \right] u_n \right\| = \min_{f \in \mathcal{P}_k} \|f(A) u_n\| \equiv \epsilon_{n,k}, \tag{5.20}$$

$$x_j' = \left[ \prod_{\substack{i=1 \\ i \neq j}}^{k} (A - \lambda_i' I) \right] u_n, \tag{5.21}$$

$$(A - \lambda_j' I) x_j' = \left( \sum_{i=0}^{k-1} c_i^{(n,k)} A^i + A^k \right) u_n = \sum_{i=0}^{k-1} c_i^{(n,k)} u_{n+i} + u_{n+k}, \tag{5.22}$$

$$\begin{aligned} \|(A - \lambda_j' I) x_j'\| &= \min_{\lambda \in \boldsymbol{C}, g \in \mathcal{P}_{k-1}} \|(A - \lambda I) g(A) u_n\|, \\ &= \min_{\lambda \in \boldsymbol{C}} \|(A - \lambda I) x_j'\|, \\ &= \min_{g \in \mathcal{P}_{k-1}} \|(A - \lambda_j' I) g(A) u_n\|, \\ &= \epsilon_{n,k} \text{ independently of } j, \end{aligned} \tag{5.23}$$

*and*

$$((A - \lambda_j' I) x_j', \ g(A) u_n) = 0, \ \text{all } g \in \pi_{k-1}. \tag{5.24}$$

*For $k = k_0$, we have $A x_j' = \lambda_j' x_j'$.*

**Proof:** We start by noting that (5.24) is nothing but a restatement of the requirement that $A x_j' - \lambda_j' x_j'$ be orthogonal to the left subspace of the Arnoldi method, which is also its right subspace $V = \{g(A) u_n : g \in \pi_{k-1}\}$.

Since the Ritz values $\lambda'_j$, $j = 1, ..., k$, are the zeros of the monic polynomial $\hat{Q}_{n,k}(\lambda) = \sum_{i=0}^{k-1} c_i^{(n,k)} \lambda^i + \lambda^k$, we can write

$$\hat{Q}_{n,k}(\lambda) = \prod_{i=1}^{k} (\lambda - \lambda'_i). \tag{5.25}$$

Thus

$$\hat{Q}_{n,k}(A) = \sum_{i=0}^{k-1} c_i^{(n,k)} A^i + A^k = \prod_{i=1}^{k} (A - \lambda'_i I). \tag{5.26}$$

Combining (5.26) with (5.19), we obtain (5.20).

Provided $x'_j$ is as given by (5.21), the proofs of (5.22) and (5.23) are immediate.

To prove the validity of (5.21) it is sufficient to show that $x'_j \in V$ and that $(A - \lambda'_j I)x'_j$ is orthogonal to all the vectors in $V$. That $x'_j \in V$ is obvious from (5.21) itself. The fact that $c_i^{(n,k)}$, $i = 0, 1, ..., k-1$, are the solution of the optimization problem in (5.19) implies that the vector $\hat{Q}_{n,k}(A)u_n$ is orthogonal to every vector in $V$. But $\hat{Q}_{n,k}(A)u_n = (A - \lambda'_j I)x'_j$, as can be seen from (5.26). This completes the proof. □

Note that the proofs of (5.20) and (5.21) for hermitian matrices can also be found in [Par2, Chap. 12, pp. 239-240].

A few historical notes on the methods of Arnoldi and Lanczos are now in order.

Following the work of Arnoldi the equivalent form in (5.19) was suggested in a paper by Erdelyi [E], in the book by Wilkinson [W, pp. 583-584], and in the papers by Manteuffel [M] and Sidi and Bridger [SiBr]. The equivalence of the different approaches does not seem to have been noticed, however. For instance, [W] discusses both approaches without any attempt to explore the connection between them. With the exception of [SiBr], these works all consider the case $n = 0$. The case $n > 0$ and the limit as $n \to \infty$ are considered in [SiBr] and [Si3].

In his discussion of the power iterations in [H, Chap. 7], Householder gives determinantal representations of certain polynomials whose zeros are approximations to the largest eigenvalues of the matrix being considered. One of these representations, namely the one given in Eq. (16) in [H, p. 186], coincides with the determinant $D(\lambda)$ in (5.1) of the present work pertaining to the

STEA approximation procedure with $n \geq 0$. It is shown there that the zeros of $D(\lambda)$ tend to the $k$ largest eigenvalues of the matrix $A$ as $n \to \infty$, but a theorem as detailed as our Theorem 3.1 is not given. It is also mentioned in the same place that, apart from a constant multiplicative factor, the polynomials $D(\lambda)$ with $n = 0$ are precisely the so-called Lanczos polynomials given in Eq. (10) of [H, p. 23] that are simply $\det(\lambda I - H)$ with $H$ as given in (4.13). As we pointed out in this section, up to a constant multiplicative factor, $D(\lambda)$ with $n > 0$ is itself the Lanczos polynomial $\det(\lambda I - H)$ when the Lanczos method is being applied with $u_0$ replaced by $u_n = A^n u_0$. It is not clear to the author whether this connection between $D(\lambda)$ with $n > 0$ and the Lanczos method has been observed before or not.

# 6 Stable Numerical Implementations

In this section we concentrate on the implementation of the generalized power methods based on the SMPE and the STEA approximation procedures as these are related to the methods of Arnoldi and Lanczos respectively, and as good implementations for the latter are known. For example, the implementations in (4.6) and (4.11) are usually quite stable.

## 6.1 General Computational Considerations

The theoretical results of Section 3 all involve the limiting procedure $n \to \infty$. When $|\lambda_1|$ is larger (smaller) than 1, we may have difficulties in implementing the procedures above due to possible overflow (underflow) in the computation of the vectors $u_m$ for large $m$. This situation can be remedied easily as will be shown below.

We first observe that the denominator polynomial $Q_{n,k}(z)$ of the vector-valued rational approximation $F_{n,k}(z)$ remains unchanged when the vectors $u_n, u_{n+1}, u_{n+2}, \ldots$, are all multiplied by the same scalar, say $\alpha$, and so do its zeros. Consequently, the vectors $\hat{d}_{ji}(n)$ defined in Theorem 3.2 remain the same up to the multiplicative factor $\alpha$. That is to say, as far as the matrix eigenvalue problem is concerned, multiplication of the vectors $u_n, u_{n+1}, \ldots$, by the scalar $\alpha$ leaves the eigenvalue approximations unchanged and multiplies the eigenvector approximations by $\alpha$.

For the purpose of numerical implementation we propose to pick $\alpha = 1/||u_n||$, and we achieve this by the following simple algorithm that is also used in the classical power method:

Step 0. Pick $u_0$ arbitrarily such that $||u_0|| = 1$.

Step 1. For $m = 1, 2, ..., n$ do

$$w_m = Au_{m-1}$$
$$u_m = w_m/||w_m||. \tag{6.1}$$

Once the vector $u_n$ has been determined in this way, we apply the $k$-step Arnoldi or Lanczos methods to the matrix $A$ with this $u_n$ as the starting vector, and obtain the $k$ Ritz values and the corresponding Ritz vectors.

## 6.2 Treatment of Nondefective Eigenvalues

If $\lambda_j$, one of the largest $t$ distinct nonzero eigenvalues of $A$ that contribute to the power iterations $u_m$ exactly as in (2.2), is nondefective, i.e., it has $\omega_j = 1$, then, under the conditions of Theorem 3.1, there is precisely one Ritz value $\lambda_j(n)$ that tends to $\lambda_j$ with $\lambda_j(n) - \lambda_j = O(n^p|\lambda_{t+1}/\lambda_j|^n)$ as $n \to \infty$ if $A$ is nonnormal and $\lambda_j(n) - \lambda_j = O(|\lambda_{k+1}/\lambda_j|^{2n})$ as $n \to \infty$ if $A$ is normal. If $x_j$ is the eigenvector corresponding to $\lambda_j$, then the Ritz vector $x_j(n)$ corresponding to $\lambda_j(n)$ tends to $x_j$ with $\limsup_{n\to\infty} ||x_j(n) - x_j||^{\frac{1}{n}} \leq |\lambda_{t+1}/\lambda_j|$ in all cases, by Theorem 3.2. Thus the Ritz value and the corresponding Ritz vector are the required approximations to the eigenpair $(\lambda_j, x_j)$.

## 6.3 Treatment of Defective Eigenvalues

When the eigenvalue $\lambda_j$ is defective and has $\omega_j > 1$ in (2.2), then, under the conditions of Theorem 3.1, there are precisely $\omega_j$ Ritz values $\lambda_{jl}(n), 1 \leq l \leq \omega_j$, that tend to $\lambda_j$, each with the rate of convergence $O([n^p|\lambda_{t+1}/\lambda_j|]^{n/\omega_j})$ as $n \to \infty$. That is to say, the Ritz values for a defective eigenvalue are not as effective as the ones for nondefective eigenvalues. However, $\hat{\lambda}_j(n)$ and $\tilde{\lambda}_j(n)$ that are defined in Theorem 3.1 do enjoy the property that they tend to $\lambda_j$ with the optimal rate of convergence $O(n^p|\lambda_{t+1}/\lambda_j|^n)$ as $n \to \infty$, as in the case of a nondefective eigenvalue.

As for the invariant subspaces $Y_i$, $i = 0, 1, ..., p_j$, $p_j = \omega_j - 1$, the most basic result to use is Theorem 3.2. Acording to this theorem and the subsequent developments, the building blocks for the invariant subspaces are the vectors $\hat{d}_{ji,l}(n)$ that are defined by (3.19). Now the vector $\hat{d}_{ji,l}(n)$ is a constant multiple of Res $F_{n,k}(z)|_{z=z_{jl}(n)}$, where $z_{jl}(n) = 1/\lambda_{jl}(n)$, which, when $\nu = -1$, is a

constant multiple of the Ritz vector corresponding to $\lambda_{jl}(n)$ by Theorem 5.4. That is, once the Ritz vectors have been computed, they can be used to construct the vectors $\hat{d}_{ji,l}(n)$, which, in turn, are used in constructing the approximate invariant subspaces $Y_i$ with optimal accuracy.

Let us now show how the vector $\hat{d}_{ji,l}(n)$ is expressed in terms of the corresponding Ritz vector. For simplicity of notation we shall write $\hat{z} = z_{jl}(n) = 1/\lambda_{jl}(n)$. The Ritz vector corresponding to $\lambda_{jl}(n)$ is $\hat{x} = \sum_{i=1}^{k} \xi_i v_i$, where $v_1 = u_n$ and $(u_n, u_n) = 1$ by (6.1). We recall that for the method of Arnoldi the vectors $v_1, v_2, ..., v_k$ are actually the ones that would be obtained by orthogonalizing the power iterations $u_n, Au_n, ..., A^{k-1}u_n$ by the Gram-Schmidt process. For the method of Lanczos the vectors $v_1, v_2, ..., v_k$ are obtained by biorthogonalizing $u_n, Au_n, ..., A^{k-1}u_n$ against the vectors $q, A^*q, ..., (A^*)^{k-1}q$. In both cases we have

$$AV = VH + R, \tag{6.2}$$

where $H$ is the upper Hessenberg matrix of (4.8) for the Arnoldi method or the tridiagonal matrix of (4.13) for the Lanczos method, and thus it is upper Hessenberg in both cases. The matrix $R$ has all of its first $k-1$ columns equal to zero, and its $k$th column is $h_{k+1,k}v_{k+1}$.

From the way the vectors $v_1, v_2, ..., v_k$ are constructed it is easy to see that

$$V = \left[u_n | Au_n | \cdots | A^{k-1}u_n\right] B, \tag{6.3}$$

where $B$ is the upper triangular matrix

$$B = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1k} \\ & \beta_{22} & \cdots & \beta_{2k} \\ & & \ddots & \vdots \\ & & & \beta_{kk} \end{bmatrix}, \tag{6.4}$$

whose entries $\beta_{ij}$ are required. Substituting (6.3) in (6.2), we have

$$[Au_n | A^2u_n | \cdots | A^ku_n]B = [u_n | Au_n | \cdots | A^{k-1}u_n]BH + R. \tag{6.5}$$

By equating the $j$th columns of both sides of (6.5) for $j < k$, we obtain

$$\sum_{i=1}^{j} (A^i u_n)\beta_{ij} = \sum_{i=0}^{j} (A^i u_n)(BH)_{i+1,j} \tag{6.6}$$

as the matrices $B$ and $BH$ are upper triangular and upper Hessenberg, respectively. From the linear independence of the vectors $A^i u_n, i = 0, 1, ..., k-1$, (6.6) reduces to

$$\beta_{ij} = (BH)_{i+1,j}, \quad 0 \le i \le j; \ \beta_{0j} \equiv 0 \text{ all } j \ge 1. \tag{6.7}$$

Now $\beta_{11} = 1$ since $v_1 = u_n$. These equations can be solved in the order $i = 0, 1, ..., j$, $j = 1, 2, ..., k-1$, which amounts to computing the 1st, 2nd,...,$k$th columns of the matrix $B$, in this order. This can be accomplished as $h_{j+1,j} > 0$ for all $j$. Thus by letting $i = 0$ in (6.7), we obtain $\sum_{r=1}^{j+1} \beta_{1r} h_{rj} = 0$, which we solve for $\beta_{1,j+1}$. Next, letting $i = 1$, we obtain $\beta_{1j} = \sum_{r=1}^{j+1} \beta_{2r} h_{rj}$, which we solve for $\beta_{2,j+1}$. By letting $i = 2, 3, ..., j$, we obtain $\beta_{i+1,j+1}$, $i = 2, 3, ..., j$, in this order.

Suppose that the Ritz vector $\hat{x}$ has been computed in the form $\hat{x} = \sum_{i=1}^{k} \xi_i v_i$ and that the $\xi_i$ have been saved. Then, recalling also that $u_{n+i} = A^i u_n$, $i = 0, 1, ..., k-1$,

$$\hat{x} = \sum_{i=0}^{k-1} \sigma_i u_{n+i}, \tag{6.8}$$

and the coefficient of $u_n$ is given by

$$\sigma_0 = \sum_{j=1}^{k} \beta_{1j} \xi_j. \tag{6.9}$$

Similarly, from (3.19), the coefficient of $u_n$ in $\hat{d}_{ji,l}(n)$ (setting $\nu = -1$ there) is given by

$$\sigma_0' = (\hat{z} - \zeta_j(n))^i \frac{\sum_{r=1}^{k} c_r^{(n,k)} \hat{z}^{k-r}}{\sum_{r=0}^{k} c_r^{(n,k)} (k-r) \hat{z}^{k-r-1}} = -(\hat{z} - \zeta_j(n))^i \frac{c_0^{(n,k)} \hat{z}^k}{Q_{n,k}'(\hat{z})}. \tag{6.10}$$

Now if we denote the Ritz values by $\lambda_1', ..., \lambda_k'$ and set $z_i' = 1/\lambda_i'$, $i = 1, ..., k$, then we can show that

$$\sigma_0' = -(\hat{z} - \zeta_j(n))^i \frac{\hat{z}}{\prod\limits_{\substack{r=1 \\ z_r' \neq \hat{z}}}^{k} (1 - z_r'/\hat{z})}, \tag{6.11}$$

so that

$$\hat{d}_{ji,l}(n) = \frac{\sigma_0'}{\sigma_0} \hat{x} = -\frac{(\hat{z} - \zeta_j(n))^i}{\prod\limits_{\substack{r=1 \\ z_r' \neq \hat{z}}}^{k} (1 - z_r'/\hat{z})} \frac{\hat{z}}{\sum_{j=1}^{k} \beta_{1j} \xi_j} \hat{x}, \tag{6.12}$$

which is the desired result.

With this we can now go on to compute the approximations to the eigenvector $a_{jp_j}$ and the vectors $a_{ji}$, $0 \leq i < p_j$, precisely as described in Sections 3.2.1 and 3.2.2, respectively. For example, the vector $\hat{d}_{jp_j}(n) = \sum_{l=1}^{\omega_j} \hat{d}_{jp_j,l}(n)$ is the approximation to the eigenvector $a_{jp_j}$ the error in which is, roughly speaking, $O(|\lambda_{t+1}/\lambda_j|^n)$ as $n \to \infty$.

## References


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

[E] I. Erdelyi, An iterative least-square algorithm suitable for computing partial eigensystems, *SIAM J. Numer. Anal.*, Ser. B, **2** (1965), pp. 421-436.

[FSi] W.F. Ford and A. Sidi, Recursive algorithms for vector extrapolation methods, *Appl. Numer. Math.*, **4** (1988), pp. 477-489.

[GV] G.H. Golub and C.F. Van Loan, *Matrix Computations*, Second Edition, Johns Hopkins University Press, Baltimore, 1989.

[GW] G.H. Golub and J.H. Wilkinson, Ill-conditioned eigensystems and the computation of the Jordan canonical form, *SIAM Rev.*, **18** (1976), pp. 578-619.

[H] A.S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964.

[K] S. Kaniel, Estimates for some computational techniques in linear algebra, *Math. Comp.*, **20** (1966), pp. 369-378.

[L] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, *J. Res. Nat. Bur. Standards*, **45** (1950), pp. 255-282.

[M] T.A. Manteuffel, Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration, *Numer. Math.*, **31** (1978), pp. 183-203.

[O] A.M. Ostrowski, On the convergence of the Rayleigh quotient iteration for the computation of characteristic roots and vectors, IV and VI, *Arch. Rat. Mech. Anal.*, **3** (1959), pp. 341-347 and **4** (1959/60), pp. 153-165.

[Pai] C.C. Paige, The computation of eigenvalues and eigenvectors of very large sparse matrices, Ph.D. dissertation, University of London, 1971.

[Par1] B.N. Parlett, Global convergence of the basic $QR$ algorithm on Hessenberg matrices, *Math. Comp.*, **22** (1968), pp. 803-817.

[Par2] B.N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[ParPo] B.N. Parlett and W.G. Poole, A geometric theory for the $QR, LU$ and power iterations, *SIAM J. Numer. Anal.*, **10** (1973), pp. 389-412.

[Sa1] Y. Saad, Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices, *Lin. Alg. Appl.*, **34** (1980), pp. 269-295.

[Sa2] Y. Saad, On the rates of convergence of the Lanczos and the block Lanczos methods, *SIAM J. Numer. Anal.*, **17** (1980), pp. 687-706.

[Si1] A. Sidi, Convergence and stability properties of minimal polynomial and reduced rank extrapolation algorithms, *SIAM J. Numer. Anal.*, **23** (1986), pp. 197-209. Originally appeared as NASA TM - 83443 (1983).

[Si2] A. Sidi, Extrapolation vs. projection methods for linear systems of equations, *J. Comp. Appl. Math.*, **22** (1988), pp. 71-88.

[Si3] A. Sidi, On extensions of the power method for normal operators, *Lin. Alg. Appl.*, **120** (1989), pp. 207-224.

[Si4] A. Sidi, Quantitative and constructive aspects of the generalized Koenig's and de Montessus's theorems for Padé approximants, *J. Comp. Appl. Math.*, **29** (1990), pp. 257- 291.

[Si5] A. Sidi, Efficient implementation of minimal polynomial and reduced rank extrapolation methods, *J. Comp. Appl. Math.*, **36** (1991), pp. 305-337.

[Si6] A. Sidi, Rational approximations from power series of vector-valued meromorphic functions. Submitted.

[SiB] A. Sidi and J. Bridger, Convergence and stability analyses for some vector extrapolation methods in the presence of defective iteration matrices, *J. Comp. Appl. Math.*, **22** (1988), pp. 35-61.

[SiFSm] A. Sidi, W.F. Ford, and D.A. Smith, Acceleration of convergence of vector sequences, *SIAM J. Numer. Anal.*, **23** (1986), pp. 178-196. Originally appeared as NASA TP-2193 (1983).

[SmFSi] D.A. Smith, W.F. Ford, and A. Sidi, Extrapolation methods for vector sequences, *SIAM Rev.*, **29** (1987), pp. 199-233. Erratum: Correction to "Extrapolation methods for vector sequences", *SIAM Rev.*, **30** (1988), pp. 623-624.

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

# REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188



| 1. AGENCY USE ONLY (*Leave blank*) | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | September 1992 | Technical Memorandum |

**4. TITLE AND SUBTITLE**

Application of Vector-Valued Rational Approximations to the Matrix Eigenvalue Problem and Connections With Krylov Subspace Methods

**5. FUNDING NUMBERS**

WU-505-62-21

**6. AUTHOR(S)**

Avram Sidi

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135-3191

**8. PERFORMING ORGANIZATION REPORT NUMBER**

E-7317

**9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)**

National Aeronautics and Space Administration
Washington, D.C. 20546-0001

**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

NASA TM-105858
ICOMP-92-18

**11. SUPPLEMENTARY NOTES**

Avram Sidi, Institute for Computational Mechanics in Propulsion, Lewis Research Center (work funded under Space Act Agreement C99066G) and Technion-Israel Institute of Technology, Computer Science Department, Haifa 3200, Israel.

**12a. DISTRIBUTION/AVAILABILITY STATEMENT**

Unclassified - Unlimited
Subject Category 64

**12b. DISTRIBUTION CODE**

**13. ABSTRACT** ***(Maximum 200 words)***

Let $F(z)$ be a vectored-valued function $F: C \to C^N$, which is analytic at $z = 0$ and meromorphic in a neighborhood of $z = 0$, and let its Maclaurin series be given. In a recent work [Si6] by the author, vector-valued rational approximation procedures for $F(z)$ that are based on its Maclaurin series were developed, and some of their convergence properties were analyzed in detail. In particular, a Koenig type theorem concerning their poles and a de Montessus type theorem concerning their uniform convergence in the complex plane were given. With the help of these theorems it was shown how optimal approximations to the poles of $F(z)$ and the principle parts of the corresponding Laurent series expansions can be obtained. In the present work we use these rational approximation procedures in conjunction with power iterations to develop bona fide generalizations of the power method for an arbitrary $N \times N$ matrix that may be diagonalizable or not. These generalizations can be used to obtain simultaneously several of the largest distinct eigenvalues and corresponding eigenvectors and other vectors in the invariant subspaces. We provide interesting constructions for both nondefective and defective eigenvalues and the corresponding invariant subspaces, and present a detailed convergence theory for them. This is made possible by the observation that vectors obtained by power iterations with a matrix are actually coefficients of the Maclaurin series of a vector-valued rational function, whose poles are reciprocals of some or all of the nonzero eigenvalues of the matrix being considered, while the principal parts of the Laurent expansions of this rational functions are vectors in the corresponding invariant subspaces. In addition, it is shown that the generalized power methods of this work are equivalent to some Krylov subspace methods, among them the methods of Arnoldi and Lanczos. Thus, the theory of the present work provides a set of completely new results and constructions for these Krylov subspace methods. This theory suggests at the same time a new mode of usage for these Krylov subspace methods that has been observed to possess computational advantages over their common mode of usage.

| 14. SUBJECT TERMS | 15. NUMBER OF PAGES |
|---|---|
| Vector-valued rational approximations; Matrix eigenproblem; Krylov subspace methods; Generalized power methods | 34 |
| | 16. PRICE CODE A03 |

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | |