Numerical Algorithms
https://doi.org/10.1007/s11075-022-01381-0

ORIGINAL PAPER

Computing eigenvalues of semi-infinite quasi-Toeplitz matrices

D. A. Bini[1] · B. Iannazzo[2] · B. Meini[1] · J. Meng[3] · L. Robol[1]

Received: 29 March 2022 / Accepted: 15 July 2022

Abstract
A quasi-Toeplitz (QT) matrix is a semi-infinite matrix of the form $A = T(a) + E$ where $T(a)$ is the Toeplitz matrix with entries $(T(a))_{i,j} = a_{j-i}$, for $a_{j-i} \in \mathbb{C}$, $i, j \geq 1$, while E is a matrix representing a compact operator in ℓ^2. The matrix A is finitely representable if $a_k = 0$ for $k < -m$ and for $k > n$, given $m, n > 0$, and if E has a finite number of nonzero entries. The problem of numerically computing eigenpairs of a finitely representable QT matrix is investigated, i.e., pairs $(\lambda, \boldsymbol{v})$ such that $A\boldsymbol{v} = \lambda\boldsymbol{v}$, with $\lambda \in \mathbb{C}$, $\boldsymbol{v} = (v_j)_{j\in\mathbb{Z}^+}$, $\boldsymbol{v} \neq 0$, and $\sum_{j=1}^{\infty} |v_j|^2 < \infty$. It is shown that the problem is reduced to a finite nonlinear eigenvalue problem of the kind $WU(\lambda)\boldsymbol{\beta} = 0$, where W is a constant matrix and U depends on λ and can be given in terms of either a Vandermonde matrix or a companion matrix. Algorithms relying on Newton's method applied to the equation $\det WU(\lambda) = 0$ are analyzed. Numerical experiments show the effectiveness of this approach. The algorithms have been included in the CQT-Toolbox [Numer. Algorithms 81 (2019), no. 2, 741–769].

Keywords Toeplitz matrices · Eigenvalues · Infinite matrices · Nonlinear eigenvalue problem · MATLAB · Operators · Spectrum

Mathematics Subject Classification (2010) 15A18 · 15B05 · 47A75 · 65F15 · 65H17

1 Introduction

A quasi-Toeplitz (QT) matrix A is a semi-infinite matrix that can be written as $A = T(a) + E$ where $T(a) = (t_{i,j})_{i,j\in\mathbb{Z}^+}$ is Toeplitz, i.e., $t_{i,j} = a_{j-i}$ for a given sequence $\{a_k\}_{k\in\mathbb{Z}}$, and E is compact, that is, E is the limit of a sequence of semi-infinite matrices E_i of finite rank. Here, convergence means that $\lim_{i\to\infty} \|E - E_i\|_s = 0$, where

Dedicated to Claude Brezinski on the occasion of his 80th birthday.

✉ J. Meng
mengjie@ouc.edu.cn

Extended author information available on the last page of the article

$\|\cdot\|_s$ is the operator norm induced by the vector norm $\|\boldsymbol{v}\|_s = (\sum_{i=1}^{\infty}|v_i|^s)^{\frac{1}{s}}$, for $\boldsymbol{v} = (v_i)_{i\in\mathbb{Z}^+}$, and the value of $s \geq 1$ depends on the specific context where the mathematical model originates.

Matrices of this kind are encountered in diverse applications related to semi-infinite domains. For instance, the analysis of queuing models, where buffers have infinite capacity, leads to QT matrices where the compact correction reproduces the boundary conditions of the model while the Toeplitz part describes the inner action of the stochastic process. A typical paradigm in this framework is given by random walks in the quarter plane. Some references in this regard can be found in the books [5, 31, 33], and in the more recent papers [29, 35, 36]. Another classical and meaningful example concerns the class of matrices that discretize boundary value problems by means of finite differences. In this case, the Toeplitz part of the QT matrix describes the inner action of the differential operator, while the compact correction expresses the boundary conditions imposed on the differential system. In this framework, it is worth citing the two books [24, 25], which are a relevant reference on a very close subject concerning Generalized Locally Toeplitz matrices (GLT) and their applications, where a rich literature is cited.

Computational aspects in the solution of matrix equations with QT matrices in bidimensional random walk have been recently investigated in [6, 7, 11], while generalizations including probabilistic models with restarts are analyzed in [8]. Other applications of QT matrices have been considered in [3, 4, 10], concerning matrix functions and means, and in [30, 37] concerning Sylvester equations. Important sources of theoretical properties of QT matrices are given in the books [14, 15], and [18]. In [9] a suitable Matlab toolbox, the CQT-toolbox, has been introduced for performing arithmetic operations with QT matrices including the four arithmetic operations and the more relevant matrix factorizations.

The analysis of the spectrum of Toeplitz matrices subjected to finite rank perturbations (localized impurities) has been performed in several papers in the literature. We refer in particular to [12] where semi-infinite and bi-infinite Toeplitz matrices are considered, and to [13] where the case of band matrices of finite but large size, with the perturbation in the $m \times m$ upper-left block, is analyzed. Spectral analysis of general operators, including the case of QT matrices, with specific attention to the control of the approximation error and with interesting applications to different fields, is carried out in [20]. The eigenvalue problem for Jacobi operators, including tridiagonal Toeplitz matrices plus a compact correction, is considered in [39], while in [21] the eigenvalue problem for infinite matrices having a finite number of nonzero entries in each column is treated by means of an infinite-dimensional QR iteration. A pseudospectral collocation method for approximating eigenvalues of evolution operators for linear renewal equations has been considered in [19].

1.1 Main results

In this paper, we deal with the computation of the eigenvalues of QT matrices, a topic that was not covered in the CQT-Toolbox of [9]. Namely, we are interested in the design and analysis of algorithms for computing the eigenvalues λ and the corresponding eigenvectors $\boldsymbol{v}$ of a given QT matrix A, that is, $\boldsymbol{v}$ is such that $A\boldsymbol{v} = \lambda\boldsymbol{v}$ and $\boldsymbol{v} \in \ell^s$ where $1 \le s < \infty$. Here ℓ^s is the set of vectors $\boldsymbol{x} = (x_i)_{i\ge 1}$ such that $\|\boldsymbol{x}\|_s < \infty$. For the sake of simplicity, in the following we will set $s = 2$ and use $\|\cdot\|$ to denote the 2-norm. The attention is restricted to the case where A is finitely representable, i.e., $A = T(a) + E$, where $T(a)$ is a band Toeplitz matrix determined by a finite number of parameters $a_{-m}, \ldots, a_n$ for $m, n > 0$, with $a_{-m}, a_n \neq 0$, E is a matrix having infinitely many rows and columns but a finite number of nonzero entries. A matrix of this kind represents a bounded linear operator from ℓ^2 in ℓ^2. We associate with the matrix A the Laurent polynomial $a(z) = \sum_{i=-m}^{n} a_i z^i$.

Recall that the spectrum of a bounded operator A is the set of $\lambda \in \mathbb{C}$ such that $A - \lambda I$ is not invertible, and the essential spectrum is the set of $\lambda \in \mathbb{C}$ such that $A - \lambda I$ is not Fredholm. We wish to point out that, not all the points of the spectrum or of the essential spectrum are necessarily eigenvalues of A. Moreover, while for a Toeplitz matrix A the set of eigenvalues does not contain isolated points and can be explicitly determined by the image $a(\mathbb{T})$ of the unit circle $\mathbb{T}$ through the Laurent polynomial $a(z)$ and by the winding number of $a(z) - \lambda$ (see [14]), for a general QT matrix having a nontrivial compact correction the set of eigenvalues may contain a continuous part and a discrete part, the latter is formed by a set of isolated eigenvalues. As an example, see Fig. 1.

We prove that any isolated eigenvalue λ of a QT matrix A is the solution of a finite nonlinear eigenvalue problem of the form

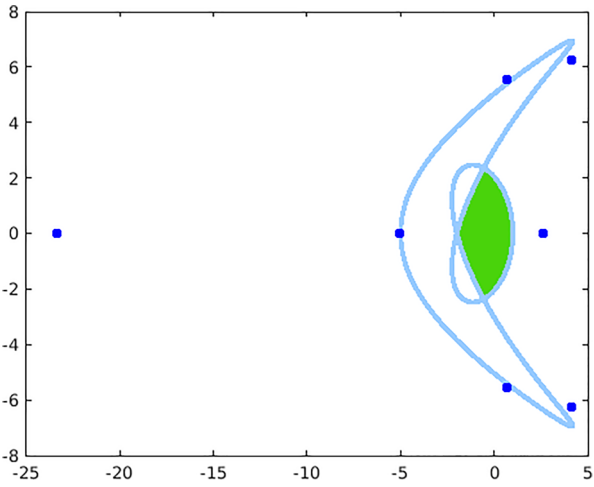

Fig. 1 Isolated eigenvalues (blue dots) and dense set of eigenvalues (green area) of the QT matrix associated with $a(z) = z^{-3} - 3z^{-2} + 2z^{-1} - 2z + z^2 + 2z^3$ and with the correction $E = (e_{i,j})$, $e_{1,j} = -2j$, $e_{2,j} = -2(10 - j)$, $e_{3,j} = -2$, for $j = 1, \ldots, 9$, $e_{i,j} = 0$ elsewhere

$$WU(\lambda)\boldsymbol{\gamma} = 0, \quad \boldsymbol{\gamma} \in \mathbb{C}^p \setminus \{0\},$$

where W is a $q \times k$ constant matrix and $U(\lambda)$ is a $k \times p$ matrix-valued function whose size p and entries depend on λ in an implicit way. Here $k, q > 0$ are integers depending on the given matrix A, while p is the number of zeros ξ_j, $j = 1, \ldots, p$ of modulus less than 1 of the Laurent polynomial $a(z) - \lambda$. It is well known that the value of p is given by $p = m + w$ where w is the winding number of $a(z) - \lambda$. Thus, it takes constant values on each connected component Ω of the set $\mathbb{C} \setminus a(\mathbb{T})$ (see Fig. 2 for an example). Note that while p depends on λ, it is locally constant on $\mathbb{C} \setminus a(\mathbb{T})$, and thus we will not write explicitly the dependence on λ.

We consider two different forms of $U = (u_{i,j})$: the *Vandermonde version* and the *Frobenius version*. In the former version, U can be chosen as the Vandermonde matrix with entries $u_{i,j} = \xi_j^{i-1}$, $i = 1, \ldots, k$, $j = 1, \ldots, p$, provided that $\xi_i \neq \xi_j$ for $i \neq j$. In the latter, U is the truncation to size $k \times p$ of the matrix $[I; G; G^2; \ldots]$ (we adopted the Matlab notation where ";" separates block rows of the matrix), where $G = F^p$ is the p-th power of the $p \times p$ companion (Frobenius) matrix F associated with the monic polynomial $s(z) = \prod_{j=1}^{p}(z - \xi_j) = z^p + \sum_{j=0}^{p-1} s_j z^j$.

This formulation of the problem allows us to detect those components Ω that constitute continuous sets of eigenvalues (for $q < p$), and to design numerical algorithms for computing the isolated eigenvalues of A (for $q \geq p$) by solving the corresponding nonlinear eigenvalue problem. Nonlinear eigenvalue problems have recently received much attention in the literature. Here we refer to the survey paper [27], to the subsequent paper [26], to the more recent works [23] and [28], and to the references therein.

Our algorithms follow the classical approach of applying Newton's iteration, as done in [23] and [27], to the scalar equation $f(\lambda) = 0$, where $f(\lambda) = \det(WU(\lambda))$ by relying on the Jacobi identity $f(\lambda)/f'(\lambda) = 1/\mathrm{trace}((WU(\lambda))^{-1} WU'(\lambda))$ whenever $p = q$, or $q > p$ where W is modified to only take p rows into account; this last case yields eigenvectors that will need to be checked a posteriori against the remaining $q - p$ equations. Here, the main problem is to exploit the specific features of the

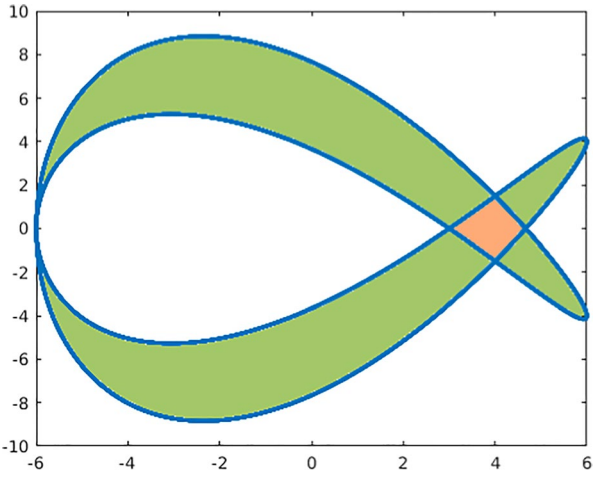

Fig. 2 Connected components of $\mathbb{C} \setminus a(\mathbb{T})$ for the Laurent polynomial $a(z) = 3z^{-3} - 2z^{-2} + z^{-1} - z - 4z^2 - 3z^3$. In green and orange the components with winding numbers 1 and 2, respectively. In white the components with winding number 0

function $f(\lambda)$ through the design of efficient algorithms to compute $U(\lambda)$ and $U'(\lambda)$ in both the Vandermonde and in the Frobenius formulation. This analysis leads to the algorithmic study of some interesting computational problems such as computing the winding number of $a(z) - \lambda$, or computing the coefficients of the polynomial factor $s(z)$ having zeros of modulus less than 1 together with their derivatives with respect to λ, or computing $G = F^p$ and the derivative of G^j for $j = 1, 2, \ldots$, with respect to λ. We will accomplish the above tasks by relying on the combination of different computational tools such as Graeffe's iteration [34], the Wiener-Hopf factorization of $a(z) - \lambda$ computed by means of the cyclic reduction algorithm [5], and the Barnett factorization of F^p [1].

The algorithms based on the Vandermonde and on the Frobenius versions require either the computation of the zeros of the Laurent polynomial $a(z) - \lambda$ and the selection of those zeros $\xi_1, \ldots, \xi_p$ having modulus less than 1, or the computation of the coefficients of the factor $\prod_{j=1}^{p}(z - \xi_j)$. In principle, the latter approach is less prone to numerical instabilities and avoids the theoretical difficulties encountered when there are multiple or clustered zeros. This fact is confirmed by numerical tests and analysis.

Our procedure uses Newton's iteration as an effective tool for refining a given approximation to an eigenvalue. In order to numerically compute all the eigenvalues we have combined Newton's iteration with a heuristic strategy based on choosing as starting approximations the eigenvalues of the $N \times N$ matrix A_N given by the leading principal submatrix of A of sufficiently large size. In fact, we may show that for any $\epsilon > 0$, the ϵ-pseudospectrum of A_N gets closer to any isolated eigenvalue of A as N gets large.

One could argue that a large value of N would provide an approximation of the isolated eigenvalues of A, directly. Nevertheless, our approach requires only a rough approximation of the isolated eigenvalues and thus a smaller value of N, followed by Newton's iteration, to compute the eigenvalues with the same accuracy. Numerical experiments show the effectiveness of this approach: examples are shown where in order to obtain full-precision approximations of the eigenvalues of A from the eigenvalues of A_N would require large values of N (of the order of millions), while starting Newton's iteration with the eigenvalues of A_N for moderate values of N (of the order of few hundreds) provides very accurate approximations in few steps. It is interesting to observe that a similar remark, concerning a different context, i.e., spectral factorization performed by means of Newton's iteration, has been pointed out in [17, page 4789].

1.2 Paper organization

The paper is organized as follows. In Section 2 we recall some preliminary properties that are useful in the subsequent analysis. In particular, Section 2.1 deals with the eigenvalues of $T(a)$ while Section 2.2 deals with the eigenvalues of $T(a) + E$. Section 3 concerns the reduction of the original eigenvalue problem for QT operators to the form of a nonlinear eigenvalue problem for finite matrices in the Frobenius and in the Vandermonde versions. Section 4 concerns further algorithmic

issues. In particular, an efficient method for computing the winding number of a Laurent polynomial is designed based on the Graeffe iteration; the problem of computing a factor of the polynomial $a(z) - \lambda$ together with its derivative with respect to λ is analyzed relying on the Barnett factorization and on the solution of a linear system associated with a resultant matrix; morever, in the same section we prove the regularity of the function $\det(WU(\lambda))$ to which Newton's iteration is applied. In Section 5 we investigate on the relationships between the isolated eigenvalues of A and the eigenvalues of A_N when N gets large. The results of some numerical experiments are reported in Section 6. Finally, Section 7 draws the conclusions and describes some open problems.

The algorithms, implemented in Matlab, have been added to the CQT-Toolbox of [9]. The main functions are `eig_single` and `eig_all`. The former computes a single eigenvalue of a QT matrix starting from a given approximation, and, optionally, an arbitrary number of components of the corresponding eigenvector, the latter provides the computation of all the eigenvalues. Other related functions integrate the package. More information, together with the description of other auxiliary functions and optional parameters can be found at https://numpi.github.io/cqt-toolbox while the software can be downloaded at https://github.com/numpi/cqt-toolbox.

2 Preliminaries

Let $a(z) = \sum_{i=-m}^{n} a_i z^i$ be a Laurent polynomial where $a_{-m}, a_n \neq 0$, and define $T(a) = (t_{i,j})_{i,j=1,2,\ldots}$, $t_{i,j} = a_{j-i}$, the Toeplitz matrix associated with $a(z)$. Given a semi-infinite matrix $E = (e_{i,j})_{i,j=1,2,\ldots}$, such that $e_{i,j} = 0$ for $i > k_1$, or $j > k_2$, the matrix $A = T(a) + E$ represents a bounded linear operator from the set $\ell^2 = \{(v_i)_{i\in\mathbb{Z}^+} : v_i \in \mathbb{C}, \sum_{i=1}^{\infty} |v_i|^2 < \infty\}$ to itself. Denote by $\mathcal{B}(\ell^2)$ the set of bounded linear operators from ℓ^2 to itself and by $\mathbb{T}$ the unit circle in the complex plane.

Recall that A is invertible if there exists $B \in \mathcal{B}(\ell^2)$ such that $AB = BA = I$, where I is the identity on $\mathcal{B}(\ell^2)$. Moreover, A is Fredholm if there exists $B \in \mathcal{B}(\ell^2)$ such that $AB - I$ and $BA - I$ are compact, i.e., A is invertible modulo compact operators. Recall also that for $A \in \mathcal{B}(\ell^2)$ the spectrum of A is defined as

$$\mathrm{sp}(A) = \{\lambda \in \mathbb{C} : A - \lambda I \text{ is not invertible}\}$$

while the essential spectrum is defined as

$$\mathrm{sp}_{\mathrm{ess}}(A) = \{\lambda \in \mathbb{C} : A - \lambda I \text{ is not Fredholm}\},$$

so that $\mathrm{sp}_{\mathrm{ess}}(A) \subset \mathrm{sp}(A)$.

It is well known that for a Laurent polynomial $a(z)$, $T(a)$ is invertible in ℓ^s if and only if $a(z) \neq 0$ for $z \in \mathbb{T}$ and $\mathrm{wind}(a) = 0$ (see [15, Corollary 1.11]), where $\mathrm{wind}(a)$ is the winding number of the curve $a(\mathbb{T})$.

In the case where $a(z)$ is a Laurent polynomial, we may write

$$\text{wind}(a) = \frac{1}{2\pi}\int_0^{2\pi} e^{\mathrm{i}t}\frac{a'(e^{\mathrm{i}t})}{a(e^{\mathrm{i}t})}dt, \tag{1}$$

where $a'(z) = \sum_{j=-m}^{n} j a_j z^{j-1}$ is the first derivative of $a(z)$. Notice that $\text{wind}(a-\lambda)$ is constant for λ in each connected component Ω of the set $\mathbb{C}\setminus a(\mathbb{T})$. Consequently, we have (see [15, Corollary 1.12])

$$\text{sp}(T(a)) = a(\mathbb{T}) \cup \{\lambda \in \mathbb{C}\setminus a(\mathbb{T}) \ : \ \text{wind}(a-\lambda)\neq 0\}, \tag{2}$$

moreover,

$$\text{sp}_{\text{ess}}(T(a)) = a(\mathbb{T}). \tag{3}$$

We say that $(\lambda, \boldsymbol{v})$ is an eigenpair (eigenvalue, eigenvector) if $A\boldsymbol{v} = \lambda\boldsymbol{v}$ and $\boldsymbol{v} \in \ell^2 \setminus \{0\}$.

2.1 Eigenvalues of $T(a)$

The following results from [15] characterize the eigenpairs of the Toeplitz operator $T(a)$. In our statements and throughout the paper, we used a slightly different notation with respect to [15]. Namely, we denote the entries of $T(a)$ as $(T(a))_{i,j} = a_{j-i}$, while in the classical literature they are denoted as $(T(a))_{i,j} = a_{i-j}$. The reason is that this notation is more suited to fit the context of Markov chains and queueing models where these matrices play an important role.

Lemma 1 [15, *Proposition* 1.20] *Let $1 \le s \le \infty$. For a Laurent polynomial $a(z)$, a point $\lambda \notin a(\mathbb{T})$ is an eigenvalue of $T(a)$ as an operator on ℓ^s if and only if $r := \text{wind}\,(a-\lambda) > 0$. Moreover, the kernel of $T(a)-\lambda I$ has dimension r and if $\boldsymbol{v} \in \ker\,(T(a)-\lambda I)$ then $\boldsymbol{v}$ is exponentially decaying.*

If $\lambda \in a(\mathbb{T})$ a similar result can be given. Let $\tau_1, \ldots, \tau_q$ be the distinct zeros of $a(z)-\lambda$ of modulus 1 and multiplicity $\alpha_1, \ldots, \alpha_q$, respectively. Define

$$c(z) = (a(z)-\lambda)/\prod_{j=1}^{q}\left(1-\frac{z}{\tau_j}\right)^{\alpha_j}, \tag{4}$$

so that $c(z)$ is a Laurent polynomial having no zero on $\mathbb{T}$. Then we have the following.

Lemma 2 [15, *Proposition* 1.22] *Let $1 \le s \le \infty$. For a Laurent polynomial $a(z)$, a point $\lambda \in a(\mathbb{T})$ is an eigenvalue of $T(a)$ as an operator on ℓ^s if and only if $r := \text{wind}(c) > 0$. Moreover, the kernel of $T(a)-\lambda I$ has dimension r and if $\boldsymbol{v} \in \ker\,(T(a)-\lambda I)$ then $\boldsymbol{v}$ is exponentially decaying.*

Observe that according to the above lemmas, the eigenvalues of $T(a)$ belong to the set $\text{sp}\,(T(a))$, which, in turn, can be explicitly described by means of (2).

Let Ω be a connected component of the set $\mathbb{C} \setminus a(\mathbb{T})$. The function wind $(a - \lambda)$ is constant on Ω, and this means that if the winding number is $r > 0$ then all the values $\lambda \in \Omega$ are eigenvalues of $T(a)$ of (geometric) multiplicity r, while if $r \leq 0$ then no $\lambda \in \Omega$ is eigenvalue of $T(a)$. We recall Proposition 1.25 from [15].

Lemma 3 *If $\lambda \in a(\mathbb{T})$ is in the boundary of Ω, and $\xi =$ wind $(a - \mu)$ for $\mu \in \Omega$, then $\xi \geq$ wind(c), where $c(z)$ is defined in (4).*

From the above results it follows that (compare with Corollary 1.26 in [15]) if λ lies on the boundary of Ω such that wind$(a - \mu) \leq 0$ for $\mu \in \Omega$ then λ cannot be an eigenvalue of $T(a)$. That is, the eigenvalues of $T(a)$ belong necessarily to those components Ω for which wind$(a - \lambda) > 0$ and to their boundaries. Therefore $T(a)$ cannot have isolated eigenvalues.

2.2 Eigenvalues of $T(a) + E$

From the definition of spectrum and of essential spectrum it follows that

$$\mathrm{sp}_{\mathrm{ess}}(T(a)) = \mathrm{sp}_{\mathrm{ess}}(T(a) + E) \subset \mathrm{sp}(T(a) + E)$$

for any compact operator E. In fact, to prove the equality, if $A = T(a) + E$ is a QT matrix, then $A - \lambda I$ is not Fredholm iff $B(A - \lambda I) - I$ and $(A - \lambda I)B - I$ are not compact for any bounded operator B. That is, iff $B(T(a) - \lambda I) - I + BE$ and $(T(a) - \lambda I)B - I + EB$ are not compact. This is equivalent to say that $B(T(a) - \lambda I) - I$ and $(T(a) - \lambda I)B - I$ are not compact, i.e., $T(a) - \lambda I$ is not Fredholm.

Another interesting property is given by the following.

Proposition 1 *If $A = T(a) + E$ is a QT matrix, then* $\mathrm{sp}(T(a)) \subset \mathrm{sp}(A)$.

The above result is an immediate consequence of the following

Lemma 4 *If the QT matrix $A = T(a) + E$ is invertible on ℓ^s $(1 \leq s \leq \infty)$, then $T(a)$ is also invertible on ℓ^s.*

Proof Since $A = T(a) + E$ is invertible, then $0 \notin \mathrm{sp}(A)$. This implies $0 \notin \mathrm{sp}_{\mathrm{ess}}(A) = \mathrm{sp}_{\mathrm{ess}}(T(a)) = a(\mathbb{T})$ so that $a(z) \neq 0$ for all $z \in \mathbb{T}$. To show $T(a)$ is invertible, it is sufficient to show that the winding number of $a(z)$ is 0, that is, wind$(a) = 0$. To this end, suppose wind$(a) = m$ and $m \neq 0$, then A is a Fredholm operator and it follows from [38, Theorem 2.8] and [15, Theorem 1.9] that the index of A is Ind $A =$ Ind$T(a) = m \neq 0$. On the other hand, since A is invertible, it follows that dim Ker $A =$ dim Coker $A = 0$, where dim Ker A is the dimension of the kernel of A and dim Coker A is the dimension of the cokernel of A. It follows from [15, page 9] that the index of A is Ind $A =$ dim Ker $A-$ dim Coker $A = 0$, from which we get a contradiction. Hence, we must have wind$(a) = 0$.

Observe that, in general, $\lambda \in \text{sp}\,(T(a)+E)$ does not imply $\lambda \in \text{sp}\,(T(a))$, as the following example shows. Denote by trid (α,β,γ) the tridiagonal Toeplitz matrix associated with the Laurent polynomial $\alpha z^{-1}+\beta+\gamma z$. Let $T(a)=\text{trid}\,(-2,5,-2)$, so that $T(a)=UU^T$, $U=\text{trid}(0,2,-1)$. Set $A=T(a)-4\boldsymbol{e}_1\boldsymbol{e}_1^T$, where $\boldsymbol{e}_1=[1,0,\ldots]^T$, so that $A=U\,\text{diag}\,(0,1,1,\ldots)U^T$. Then $0\in\text{sp}\,(A)$ since A is not invertible (but it is Fredholm), while $0\notin\text{sp}\,(T(a))$ since $T(a)$ is invertible being U and U^T invertible operators. That is, adding a compact correction E to $T(a)$ there may be eigenvalues of $A=T(a)+E$ not belonging to $\text{sp}(T(a))$.

The following two results are useful for our analysis.

Proposition 2 *Let $\lambda\notin a(\mathbb{T})$ and $w=\text{wind}\,(a-\lambda)$. Then the Laurent polynomial $a(z)-\lambda$ has $p=m+w$ zeros of modulus less than* 1.

Proof Since $a(e^{\mathfrak{i}t})'=\mathfrak{i}e^{\mathfrak{i}t}a'(e^{\mathfrak{i}t})$, from (1) we get $\frac{1}{2\pi\mathfrak{i}}\int_0^{2\pi}\frac{a(e^{\mathfrak{i}t})'}{a(e^{\mathfrak{i}t})-\lambda}dt=w$, which implies that the number p of zeros and the number $\hat{m}$ of poles of $a(z)-\lambda$ in the open unit disk are such that $p-\hat{m}=w$. Since $\hat{m}=m$, it follows $p=m+w$.

A similar result holds for $\lambda\in a(\mathbb{T})$.

Proposition 3 *Let $\lambda\in a(\mathbb{T})$ and suppose that $a(z)-\lambda$ has q zeros $\tau_1,\ldots,\tau_q$ of modulus* 1 *with multiplicities $\alpha_1,\ldots,\alpha_q$, let $c(z)$ be the Laurent polynomial in* (4). *Then, $a(z)-\lambda$ has $p=m+w-(\alpha_1+\ldots+\alpha_q)$ zeros of modulus less than* 1, *where $w=\text{wind}(c)$.*

3 Computational analysis

In this section, we aim at the design and analysis of numerical algorithms for computing the eigenvalues of the finitely representable QT matrix $A=T(a)+E$ belonging to a given connected component Ω of $\mathbb{C}\setminus a(\mathbb{T})$, together with the corresponding eigenvectors. For the sake of simplicity, the case $\lambda\in a(\mathbb{T})$ is not treated in this paper.

If $E=0$ then the spectrum and the essential spectrum of $T(a)$ are explicitly known (see (2), and (3)). Moreover, an eigenvalue λ, together with its multiplicity, can be explicitly characterized in terms of the winding number wind $(a-\lambda)$, if $\lambda\notin a(\mathbb{T})$ (see Lemma 1). Therefore the case of interest is $E\neq 0$.

Recall the following notations: $a(z)=\sum_{i=-m}^{n}a_iz^i$, while k_1 is the row size of the non zero part of the correction E. We set $q=\max(m,k_1)$, and denote p the number of zeros of modulus less than 1 of the Laurent polynomial $a(z)-\lambda$. In view of Proposition 2 we have $p=m+\text{wind}(a-\lambda)$, moreover p is constant for $\lambda\in\Omega$. Finally, for a given matrix A, we denote by $A_{r\times s}$ the leading principal submatrix of A of size $r\times s$, i.e., the submatrix formed by the entries in the first r rows and in the first s columns. If $r=s$ we write A_r in place of $A_{r\times s}$.

3.1 Reduction to a nonlinear eigenvalue problem

Consider an eigenpair $(\lambda, \boldsymbol{v})$ of $A = T(a) + E$ so that $\boldsymbol{u} := (A - \lambda I)\boldsymbol{v} = 0$. Observe that the condition $u_k = 0$ for $k \geq q + 1$ can be written as the linear difference equation

$$\sum_{j=-m}^{n} a_j v_{k+j} - \lambda v_k = 0, \quad k \geq q + 1, \tag{5}$$

whose characteristic polynomial is $b(z) = (a(z) - \lambda)z^m$. The dimension of the space of solutions of (5) that belong to ℓ^2 depends on λ and coincides with the number p of roots of $a(z) - \lambda$ with modulus less than 1. Our two approaches differ in the way the basis of the latter space is chosen.

If $\boldsymbol{v}^{(j)}$, $j = 1, \ldots, p$ is a basis of the space of solutions, then we may write the eigenvector $\boldsymbol{v}$ as a linear combination of $\boldsymbol{v}^{(j)}$, i.e., $\boldsymbol{v} = \sum_{j=1}^{p} \alpha_j \boldsymbol{v}^{(j)}$. Therefore, we may say that $(\lambda, \boldsymbol{v})$ is an eigenpair for A if and only if $\boldsymbol{v} = \sum_{j=1}^{p} \alpha_j \boldsymbol{v}^{(j)}$ and the conditions $u_1 = \ldots = u_q = 0$ are satisfied.

The latter conditions form a nonlinear system in q equations and p unknowns which can be written as

$$\begin{aligned} &HV(\lambda)\boldsymbol{\alpha} = \lambda V_{q\times p}(\lambda)\boldsymbol{\alpha}, \quad H = A_{q\times\infty} \\ &V(\lambda) = [\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \ldots, \boldsymbol{v}^{(p)}], \quad \boldsymbol{\alpha} \in \mathbb{C}^p, \quad \text{wind}\,(a - \lambda) = p - m. \end{aligned} \tag{6}$$

In fact, λ and the p components of $\boldsymbol{\alpha}$, normalized such that $\|\boldsymbol{\alpha}\| = 1$, form a set of p unknowns. It is clear that the system (6) is in the form of a nonlinear eigenvalue problem (NEP).

This system has a nontrivial solution $\boldsymbol{\alpha}$ for a given λ if and only if λ is eigenvalue of A corresponding to the eigenvector $\boldsymbol{v} = V(\lambda)\boldsymbol{\alpha}$. Notice that, for $p > q$, this system has always a solution since the matrix $HV(\lambda) - \lambda V_{q\times p}(\lambda)$ has more columns than rows so that $\ker(HV(\lambda) - \lambda V_{q\times p}(\lambda)) \neq \{0\}$ and the multiplicity of λ is given by $p - \text{rank}(HV(\lambda) - \lambda V_{q\times p}(\lambda))$.

If $p = q$, (6) provides a balanced nonlinear eigenvalue problem that we are going to analyze.

If $p < q$ and if the pair $(\lambda, \boldsymbol{\alpha})$ solves (6), then it solves also the balanced nonlinear eigenvalue problem

$$H_{p\times\infty} V(\lambda)\boldsymbol{\alpha} = \lambda V_p(\lambda)\boldsymbol{\alpha}, \tag{7}$$

formed by the first p equations of (6). Thus, we may look for solutions $(\lambda, \boldsymbol{\alpha})$ of (7), and, if any, we may check if these are also solutions of (6).

We may express the NEP (6) in a more convenient form by using the Toeplitz structure of $T(a)$. This is the subject of the next section.

3.2 A different formulation

Let $Z = (z_{i,j})$ be the shift matrix defined by $z_{i,i+1} = 1$, $z_{i,j} = 0$ elsewhere. Then for any solution $\boldsymbol{v}$ of the linear difference equation (5), the shifted vector $Z^k\boldsymbol{v}$ is still

a solution for any $k \geq 0$. Moreover, if $\boldsymbol{v} \in \ell^2$ then also $Z^k\boldsymbol{v} \in \ell^2$, and if $\boldsymbol{v}^{(1)}$ and $\boldsymbol{v}^{(2)}$ are linearly independent, then also $Z^k\boldsymbol{v}^{(1)}$ and $Z^k\boldsymbol{v}^{(2)}$ are linearly independent. To show the latter implication, assume that there exists a linear combination $\boldsymbol{v} = \alpha_1\boldsymbol{v}^{(1)} + \alpha_2\boldsymbol{v}^{(2)} \neq 0$ such that $Z^k\boldsymbol{v} = 0$. Then, $v_i = 0$ for $i \geq k+1$. But since $a_{-m} \neq 0$, we find that $v_k = \ldots = v_1 = 0$, i.e., $\boldsymbol{v} = 0$, which is a contradiction.

Therefore, if the columns of $V(\lambda)$ are a basis of the space of the solutions in ℓ^2, then also the columns of $U(\lambda) = Z^m V(\lambda)$ form a basis of the same space. This implies that the columns of $U(\lambda)$ are linear combinations of the columns of $V(\lambda)$. That is, there exists a non singular $p \times p$ matrix $S(\lambda)$ such that $U(\lambda) = V(\lambda)S(\lambda)$ whence we have $Z^m V(\lambda) = V(\lambda)S(\lambda)$.

If we multiply the rows from $m+1$ to $2m$ of the Toeplitz matrix $T(a) - \lambda I$ by $V(\lambda)$ we get

$$\begin{bmatrix} a_{-m} & \ldots & a_{-1} & a_0 - \lambda & a_1 & \ldots & a_n & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & a_{-m} & \cdots & a_{-1} & a_0 - \lambda & a_1 & \cdots & a_n \end{bmatrix} V(\lambda) = 0.$$

Observing that $V(\lambda) = [V_{m\times p}(\lambda); Z^m V(\lambda)]$, we may rewrite the identity as

$$\begin{bmatrix} B & (T(a) - \lambda I)_{m,\infty} \end{bmatrix} \begin{bmatrix} V_{m\times p}(\lambda) \\ Z^m V(\lambda) \end{bmatrix} = 0, \quad B = \begin{bmatrix} a_{-m} & \ldots & a_{-1} \\ & \ddots & \vdots \\ & & a_{-m} \end{bmatrix}.$$

Since $Z^m V(\lambda) = V(\lambda)S(\lambda)$, we get

$$(T(a) - \lambda I)V(\lambda) = \begin{bmatrix} -B \\ 0_{\infty\times m} \end{bmatrix} . V_{m\times p}(\lambda)S(\lambda)^{-1}$$

On the other hand, relying once again on the property $Z^m V(\lambda) = V(\lambda)S(\lambda)$, we find that

$$\begin{aligned} (A - \lambda I)V(\lambda) &= -\begin{bmatrix} B \\ 0_{\infty\times m} \end{bmatrix} V_{m\times p}(\lambda)S(\lambda)^{-1} + EV(\lambda) \\ &= \left(-\begin{bmatrix} B & 0_{m\times\infty} \\ 0_{\infty\times m} & 0_{\infty\times\infty} \end{bmatrix} V(\lambda) + EZ^m V(\lambda) \right) S(\lambda)^{-1} \end{aligned}$$

so that

$$(A - \lambda I)V(\lambda) = MV(\lambda)S(\lambda)^{-1}, \quad M = \begin{bmatrix} -B & 0_{m\times\infty} \\ 0_{\infty\times m} & 0_{\infty\times\infty} \end{bmatrix} + EZ^m. \tag{8}$$

The possibly nonzero rows of the matrix M are the first $q = \max(m, k_1)$ rows, which form the $q \times \infty$ matrix $N = M_{q\times\infty}$, i.e., $M = [N; 0]$. It is interesting to observe that the matrix EZ^m is obtained by shifting the columns of E to the right of m places. This implies that the matrix N takes one of the following forms

$$N = \begin{bmatrix} -B & E_1 \\ 0_{(q-m)\times m} & E_2 \end{bmatrix}, \quad N = \begin{bmatrix} -B & E_1 \end{bmatrix},$$

depending on whether $q > m$ or $q = m$, respectively, where $E = [E_1;E_2]$, and E_1 has size $m \times \infty$ while E_2 has size $(q-m) \times \infty$. In other words, the submatrices E and B do not overlap. This fact allows us to rewrite (6) as a set of q equations in p unknowns in the more convenient form

$$NV(\lambda)\boldsymbol{\beta} = 0, \quad N = \begin{bmatrix} -B & E_1 \\ 0_{(q-m)\times m} & E_2 \end{bmatrix}, \quad \boldsymbol{\beta} = S(\lambda)^{-1}\boldsymbol{\alpha}. \tag{9}$$

Another observation is that multiplying equation (9) on the left by any invertible matrix provides an equivalent formulation of the NEP@. In particular, if $q > m$, consider the rank revealing QR factorization $E_2 = QR$ of the matrix E_2, assume that $\text{rank}(E_2) = r_2$ and denote $\widetilde{R}$ the $r_2 \times \infty$ matrix formed by the first r_2 rows of R so that $R = [\widetilde{R};0]$ and we may write $E_2 = Q[\widetilde{R};0_{(q-m-r_2)\times\infty}]$.

Multiplying (9) to the left by diag (I_m, Q^*), where Q^* is the transposed Hermitian of Q, yields

$$WV(\lambda)\boldsymbol{\beta} = 0, \quad W = \begin{bmatrix} -B & E_1 \\ 0_{r_2\times m} & \widetilde{R} \end{bmatrix}, \quad \boldsymbol{v} = V(\lambda)S(\lambda)\boldsymbol{\beta} = Z^m V(\lambda)\boldsymbol{\beta}. \tag{10}$$

Observe that W is a constant matrix of full rank, with $m + \text{rank}(E_2) = m + r_2$ rows, while $V(\lambda)$ is a matrix depending on λ. The eigenvalue problem for QT matrices is reduced to the NEP (10) which can take different forms according to the way a basis of the solution of the difference equation (5) is chosen.

We may conclude with the following result.

Theorem 1 *Let Ω be a connected component of the set $\mathbb{C} \setminus a(\mathbb{T})$, let $\lambda \in \Omega$ and $p = m + \text{wind}(a - \lambda)$. Let $V(\lambda)$ be a matrix whose p columns form a basis of the space of solutions of the difference equation (5) belonging to ℓ^2. If $p > q$ then all $\lambda \in \Omega$ are eigenvalues of $T(a) + E$. If $p \le q$ then $\lambda \in \Omega$ is eigenvalue of $A = T(a) + E$ corresponding to the eigenvector $\boldsymbol{v} \in \ell^2$ iff there exists $\boldsymbol{\beta} \in \mathbb{C}^p \setminus \{0\}$ which solves the nonlinear eigenvalue problem $WV(\lambda)\boldsymbol{\beta} = 0$ of (10). In this case, $\boldsymbol{v} = Z^m V(\lambda)\boldsymbol{\beta}$.*

In what follows, without loss of generality, we assume that the nonlinear eigenvalue problem is balanced. This case is encountered if $p = q$ or if $p < q$ where we consider the subset of the first p equations in (10).

3.3 Choosing a basis: Vandermonde and Frobenius versions

Let the zeros ξ_i of $a(z) - \lambda$ be simple and ordered as

$$|\xi_1| \le \cdots \le |\xi_p| < 1 \le |\xi_{p+1}| \le \cdots \le |\xi_{m+n}|,$$

and let $V(\lambda) = (\xi_j^{i-1})_{i\in\mathbb{Z}^+, j=1,\ldots,p}$ be the $\infty \times p$ Vandermonde matrix associated with $\xi_1, \ldots, \xi_p$. The columns $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(p)}$ of $V(\lambda)$ provide a basis of the set of solutions of the difference equation (5) that belong to ℓ^2, so that $\boldsymbol{v}$ is an eigenvector of A corresponding to λ if and only if there exists $\boldsymbol{\alpha} = (\alpha_i) \in \mathbb{C}^p \setminus \{0\}$ such that $\boldsymbol{v} = \sum_{j=1}^p \alpha_j \boldsymbol{v}^{(j)}$

and (6) is satisfied. The same argument can be applied in the case of confluent zeros considering a generalized Vandermonde matrix.

The formulation (10) where $V(\lambda)$ is the (generalized) Vandermonde matrix associated with the roots ξ_i of $a(z) - \lambda$ is referred to as the *Vandermonde version* of the problem. It is well known that the zeros of a polynomial are severely ill-conditioned if they are clustered. This may make the choice of the basis $\boldsymbol{v}^{(i)}$, given by the columns of the Vandermonde matrix, unsuited in some problems. A way to overcome this issue is to consider the *Frobenius version* of the NEP obtained in the following way.

For the sake of notational simplicity, in the following we write V in place of $V(\lambda)$. For simple roots, write the Vandermonde matrix V in the form $V = [V_p; V_p D^p; V_p D^{2p}; \ldots]$, with $D = \mathrm{diag}(\xi_1, \ldots, \xi_p)$, and define $U := VV_p^{-1}$. Recall that $V_p D^p V_p^{-1} = F^p$, where $F = Z_p - \boldsymbol{e}_p[s_0, s_1, \ldots, s_{p-1}]$ denotes the companion (Frobenius) matrix associated with the polynomial $s(z) = (z - \xi_1) \cdots (z - \xi_p) = \sum_{i=0}^{p-1} s_i z^i + z^p$, see for instance [5]. Here, $\boldsymbol{e}_p = [0, \ldots, 0, 1]^T \in \mathbb{R}^p$. For multiple roots, a similar construction can be made with the generalized Vandermonde matrix and where D is a block diagonal matrix whose diagonal blocks are Jordan blocks associated with the distinct roots of $a(z) - \lambda$ having modulus smaller than 1.

Denote $G := F^p$ so that the columns of $U = VV_p^{-1} = [I; G; G^2; \ldots]$ provide a different basis of the set of solutions of the linear difference equation (5). The NEP (10) can be equivalently rewritten as

$$WU\boldsymbol{\gamma} = 0, \quad \boldsymbol{v} = Z^m U\boldsymbol{\gamma}, \tag{11}$$

We refer to (11) as the *Frobenius version* of the problem. Observe that in the Frobenius form, it is not relevant if the roots of $a(z) - \lambda$ are multiple or numerically clustered, in fact the matrix $G = F^p$ exists and can be computed independently of the location of the roots of $s(z)$.

Notice that if $m + r_2 = p$, then the matrix W can be partitioned into $p \times p$ blocks as $W = [W_0, W_1, W_2, \ldots]$ and WU can be rewritten in terms of a matrix power series as $WU = \sum_{i=0}^{\infty} W_i G^i$. The following result provides information in this regard [5, Chapter 3].

Theorem 2 *Assume that* $a(z) = \sum_{i=-m}^{n} a_i z^i$, *where* $a_{-m}, a_n \neq 0$, *has roots* ξ_i, $i = 1, \ldots, m+n$ *such that* $|\xi_1| \leq \cdots \leq |\xi_p| < 1 \leq |\xi_{p+1}| \leq \ldots \leq |\xi_{m+n}|$ *and denote* $s(z) = \prod_{i=1}^{p}(z - \xi_i)$. *Define* $A_k = (a_{j-i+kp-m+p})_{i,j=1,p}$ *for* $k = -1, 0, 1, \ldots$ *where we assume* $a_\ell = 0$ *if* $\ell < -m$ *or* $\ell > n$. *Let* F *be the Frobenius matrix associated with the factor* $s(z)$. *Then* $G = F^p$ *is the unique solution of the matrix equation*

$$\sum_{k=-1}^{\infty} A_k X^{k+1} = 0, \tag{12}$$

having minimum spectral radius $\rho(G)$, *moreover*, $\rho(G) = |\xi_p|$.

Notice that the blocks A_k defined in the above theorem are obtained by partitioning the Toeplitz matrix $T\left(z^{m-p}a(z)\right)$ into $p \times p$ blocks which are themselves Toeplitz. Moreover, since $T\left(z^{m-p}a(z)\right)$ is a banded matrix, then $A_k = 0$ for k sufficiently large. In the literature, there are several effective algorithms for the numerical computation of G, based on fixed point iterations or on doubling techniques. We refer the reader to [2, 5, 16], and [17], for more details.

4 The numerical algorithms

In this section we describe our algorithms to refine a given approximation of an eigenvalue λ of $A = T(a) + E$, while in Section 5 we will discuss how to get the initial approximation. The algorithms require: a function $g(x) : \mathbb{C} \to \mathbb{C}$ such that the fixed point iteration $\lambda_{\nu+1} = g(\lambda_\nu)$ converges locally to the eigenvalue λ, solution of the NEP (10), and a choice of the basis $V(\lambda)$ of the solutions of (5) belonging to ℓ^2.

The general scheme is reported in the Template Algorithm 1. This algorithm, for an initial approximation $\lambda_0 \in \mathcal{U}_w := \{\lambda \in \mathbb{C} \setminus a(\mathbb{T}) : \text{wind}\,(a-\lambda) = w\}$ of the eigenvalue, provides either a more accurate approximation to the corresponding eigenpair, or a message with the following possible cases: 1) all the elements in the set $\mathcal{U}_w$ are eigenvalues; 2) the generated sequence exited from $\mathcal{U}_w$; 3) it holds $p < q$ and the approximated solution solves the first p equations but not the full NEP (10); 4) convergence did not occur after the maximum number of allowed iterations.

Algorithm 1 Template algorithm.

Input: the coefficients of $a(z) = \sum_{i=-m}^{n} a_i z^i$ and the correction matrix E; an error bound $\epsilon > 0$; an initial approximation $\lambda_0 \in \mathbb{C}$; an upper bound `maxit` to the number of iterations; a function $g(x) : \mathbb{C} \to \mathbb{C}$ defining a fixed point iteration to solve the NEP (10); a rule to generate $V(\lambda)$.

Output: An approximation μ to an eigenvalue λ, the vector $\boldsymbol{\beta}$ providing an approximation to the corresponding eigenvector according to (10), together with a message.

1: Construct the matrix W of (10) together with the scalar r_2; compute $w_0 = \text{wind}(a - \lambda_0)$, $p_0 = m + w_0$ and $q = m + r_2$. Set $\nu = 0$.
2: **while** $\nu <$ `maxit` **do**
3: Compute $w = \text{wind}(a - \lambda_\nu)$, $p = m + w$.
4: if $w \neq w_0$ then output '`out of` $\mathcal{U}_{w_0}$' and stop; otherwise set $w_0 = w$.
5: if $p > q$ then output '`continuous set of eigenvalues`' and stop.
6: if $p = q$ then perform one step of the fixed point iteration $\lambda_{\nu+1} = g(\lambda_\nu)$; set $\nu = \nu + 1$, compute $\boldsymbol{\beta}$ and $\boldsymbol{v}$ according to (10), compute the residual error `res` $= \|((A-\lambda_\nu I)\boldsymbol{v})_{q\times 1}\|/\|\boldsymbol{v}_{q\times 1}\|$; if `res` $\leq \epsilon$, output '`isolated eigenvalue (p=q)`' together with $\mu = \lambda_\nu$, $\boldsymbol{\beta}$ and exit, otherwise continue from step 3;
7: if $p < q$ then perform one step of the fixed point iteration $\lambda_{\nu+1} = g(\lambda_\nu)$ applied to (10) restricted to the first p components. Check if the residual error in the first p components is less than ϵ. If not, continue from step 3, otherwise check if $\text{rank}(WV(\lambda_\nu))$ is less than p. If so, output '`isolated eigenvalue (p<q)`' together with $\mu = \lambda_\nu$ and $\boldsymbol{\beta}$, and exit; otherwise output '`non converging sequence (p<q)`' and exit;
8: Stop if the maximum number of iterations `maxit` has been reached; in this case, output '`Maximum number of iterations exceeded`'.
9: **end while**

Now, we deal with algorithmic issues encountered in the design of the fixed point iterations to solve the nonlinear eigenvalue problem (10). This analysis is needed to design algorithms to implement the function $g(x)$ used in the Template Algorithm 1.

We essentially analyze Newton's iteration applied to the determinantal versions of the problem, that is, $\det(WV) = 0$, $\det(WU) = 0$, in the Vandermonde and in the Frobenius forms, respectively. Before doing that, we discuss on how to compute the winding number of $a(z) - \lambda$, since this is a fundamental step in the design of the overall algorithm.

4.1 Computing the winding number

The winding number w of the Laurent polynomial $a(z) - \lambda$ can be computed in different ways. The most elementary one is to express w as $w = p - m$, where p is the number of zeros of $a(z) - \lambda$ of modulus less than 1. Any root-finding algorithm applied to the polynomial $z^m(a(z) - \lambda)$ can be used for this purpose, for instance, the

command `roots` of Matlab provides approximations to all the roots of $z^m(a(z) - \lambda)$, and we may count how many roots have modulus less than 1. This approach has the drawback that polynomial roots are ill-conditioned when clustered, so that we may encounter instability if there are clusters of roots of modulus close to 1.

A second approach is based on equation (1) that expresses w as ratio of two integrals. The integrals can be approximated by the trapezoid rule at the Fourier points using two FFTs. In this case, the presence of roots of the polynomial close to the unit circle may lead to a large number of Fourier points with a consequent slow down of the CPU time.

A third approach, which is the one we have implemented, relies on Graeffe's iteration [34], which is based on the following observations. Given a polynomial $b(z)$ of degree $m + n$, the polynomial $c(z) = b(z)b(-z)$ is formed by monomials of even degree, i.e., there exists a polynomial $b_1(z)$ of degree $m + n$ such that $b_1(z^2) = c(z)$. Therefore, the roots of $b_1(z)$ are the square of the roots of $b(z)$. Consider the sequence defined by the Graeffe iteration $b_{k+1}(z^2) = b_k(z)b_k(-z)$ with initial value $b_0(z) = b(z)$. It turns out that the winding number of $b_k(z)$ is constant. Moreover, if $b(z)$ has m zeros of modulus less than 1 and n zeros of modulus greater than 1, then the limit for $k \to \infty$ of $b_k(z)/\theta_k$ is z^m. Here θ_k is the coefficient of maximum modulus of $b_k(z)$. This means that there exists an index k such that the coefficient of z^m in $b_k(z)$ has modulus greater than $\frac{1}{2}\|b_k(z)\|_1$, where $\|b_k(z)\|_1$ is the sum of the moduli of all the coefficients of $b_k(z)$. In view of Rouché theorem, the latter inequality is a sufficient condition to ensure that $b_k(z)$ has m roots of modulus less than 1.

Indeed, if there are zeros of modulus 1 then this procedure might not terminate. Therefore, if the number of Graeffe iterations exceeds a given upper bound, then the explicit computation of the polynomial roots is performed. These arguments support Algorithm 2 for counting the number of roots of a polynomial of modulus less than 1.

Algorithm 2 Count roots

Input: The coefficients of a polynomial $b(z) = \sum_{i=0}^{d} b_i z^i$ of degree d; an upper bound `maxit` to the number of iterations.
Output: Either the number of zeros of $b(z)$ of modulus less than 1, or the message `'failure'`.
1: Set $b_0(z) = b(z)$, $\nu = 0$.
2: **while** $\nu <$`maxit` **do**
3: Compute the coefficients of $c(z) = \sum_{i=0}^{d} c_i z^i$ such that $c(z^2) = b_\nu(z)b_\nu(-z)$, let h be the minimum index such that $|c_h| = \max_i |c_i|$, and set $b_{\nu+1}(z) = c(z)/c_h$.
4: $\nu = \nu + 1$
5: If $\|b_\nu(z)\|_1 < 2$ then output h and exit
6: **end while**
7: Compute the zeros of $c(z)$ and output the number of zeros of modulus less than 1 together with the warning: `'Reached the maximum number of iterations'`

4.2 Implementing Newton's iteration

In this section we analyze the computational issues concerning the implementation of Newton's iteration applied either to $f_V(\lambda) = \det \boldsymbol{\Phi}_V(\lambda)$, where $\boldsymbol{\Phi}_V(\lambda) = WV(\lambda)$ in the Vandermonde approach, or to $f_F(\lambda) = \det \boldsymbol{\Phi}_F(\lambda)$, where $\boldsymbol{\Phi}_F(\lambda) = WU(\lambda)$ in the Frobenius approach. We use the symbol $\boldsymbol{\Phi}(\lambda)$ to denote either $\boldsymbol{\Phi}_V(\lambda)$ or $\boldsymbol{\Phi}_F(\lambda)$, similarly we do for $f(\lambda)$. In all cases, $\boldsymbol{\Phi}(\lambda)$ is assumed to be a $p \times p$ matrix. This is true if $q = p$, and also in the case where $q > p$ when we consider only the first p rows of $WV(\lambda)$ or of $WU(\lambda)$.

Since $U(\lambda) = V(\lambda)V_p^{-1}$ then we have $\boldsymbol{\Phi}_V(\lambda) = \boldsymbol{\Phi}_F(\lambda)V_p$ so that $f_V(\lambda) = f_F(\lambda)\det V_p(\lambda)$. We recall that if the function $f(\lambda)$ has continuous second derivative, then Newton's method applied to the equation $f(\lambda) = 0$, given by $z_{\nu+1} = z_\nu - f(\lambda_\nu)/f'(\lambda_\nu)$, locally converges to a zero of $f(\lambda)$. The convergence is at least quadratic if the zero is simple, it is linear if the zero is multiple. If $\boldsymbol{\Phi}(\lambda)$ has entries with continuous second derivative, then also $f(\lambda) = \det \boldsymbol{\Phi}(\lambda)$ has continuous second derivative and for the Newton's correction $f(\lambda)/f'(\lambda)$ we have

$$f(\lambda)/f'(\lambda) = 1/\mathrm{trace}(\boldsymbol{\Phi}(\lambda)^{-1}\boldsymbol{\Phi}'(\lambda)). \tag{13}$$

A simple calculation shows that if $\boldsymbol{\Phi}(\lambda) = P(\lambda)Q(\lambda)$ then

$$f(\lambda)/f'(\lambda) = 1/\big(\mathrm{trace}(P(\lambda)^{-1}P'(\lambda)) + \mathrm{trace}(Q(\lambda)^{-1}Q'(\lambda))\big). \tag{14}$$

In particular, since $U(\lambda) = V(\lambda)V_p$, assuming $f_F(\lambda)$ and $f_V(\lambda)$ differentiable, we have

$$f_F(\lambda)/f_F'(\lambda) = f_V(\lambda)/f_V'(\lambda) + \mathrm{trace}(V_p(\lambda)^{-1}V_p'(\lambda)).$$

4.2.1 Vandermonde version

In order to apply Newton's iteration in the Vandermonde version, we have to assume that the roots $\xi_i(\lambda)$ of the polynomial $a(z) - \lambda$ have continuous second derivative. It is well known that if the coefficients of a polynomial $p_\lambda(z)$ of degree ν are analytic functions of λ, and if for a given λ_0 the polynomial has simple roots $\xi_1, \ldots, \xi_\nu$, then for λ in a neighborhood of λ_0, there exist $\xi_1(\lambda), \ldots, \xi_\nu(\lambda)$ analytic functions that are roots of $p_\lambda(z)$ and $\xi_i(\lambda_0) = \xi_i$, for $i = 1, \ldots, \nu$. Indeed, the polynomial $z^m(a(z) - \lambda)$ has coefficients that are analytic for $\lambda \in \mathbb{C}$, therefore $\xi_i(\lambda)$ are analytic functions as long as the zeros remain simple. In this subsection we assume this condition.

In order to compute the Newton correction by means of (13) we need to compute the entries of the Vandermonde matrix $V(\lambda)$. Therefore we assume we are given a polynomial rootfinder which approximates the roots of $z^m(a(z) - \lambda)$ so that we may select the p roots of modulus less than 1. For this task we rely on the Matlab command '`roots`'. Then we need to compute $V'(\lambda)$, i.e., the derivative of the entries of $V(\lambda)$. Concerning this task we have $(\xi_j^i)' = i\xi_j^{i-1}\xi_j'$. Moreover, since $a(\xi_j) - \lambda = 0$, taking the derivative of this equation yields $a'(\xi_j)\xi_j' - 1 = 0$, whence $\xi_j' = 1/a'(\xi_j)$. Therefore, we are able to implement the Newton iteration where the Newton correction takes the form (13) with $(V'(\lambda))_{i,j} = (i-1)\xi_j^{i-2}/a'(\xi_j)$.

4.2.2 Frobenius version

Consider the case $\boldsymbol{\Phi}(\lambda) = WU(\lambda)$, where $U(\lambda)$ is the matrix defined in Section 3.3. In order to evaluate the Newton correction, we have to compute the matrix G of minimal spectral radius which solves the matrix equation (12), then evaluate, the powers G^j and their derivatives $(G^j)'$, for $j \geq 0$.

Firstly, we discuss on how to compute G. This matrix can be obtained by the coefficients of the polynomial $s(z)$ collecting the zeros of $a(z) - \lambda$ of modulus less than 1, which yields the Frobenius matrix F and in turn $G = F^p$. In our implementation we

compute directly the matrix G as the solution of minimal spectral radius of equation (12) (compare Theorem 2). For this task, the algorithm of Cyclic Reduction, having a quadratic convergence, can be effectively applied [2]. It is worth pointing out that the first row of $-G$ contains the coefficients $s_0, \dots, s_{p-1}$ of the sought monic factor $s(z)$, so that these coefficients are known once the matrix G has been computed.

Secondly, we show how to compute the derivative of the coefficients $s_0, \dots, s_{p-1}$ of $s(z)$ with respect to λ. The polynomial $z^m(a(z) - \lambda)$ can be factorized as $z^m(a(z) - \lambda) = s(z)u(z)$, where $u(z)$ has zeros of modulus greater than or equal to 1, and $s(z)$ has zeros of modulus less than 1. Therefore, setting $\widehat{p} = m + n - p$, we have the equation

$$\begin{bmatrix} a_{-m} \\ \vdots \\ a_0 - \lambda \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} u_0 & & & & \\ u_1 & u_0 & & & \\ \vdots & \ddots & \ddots & & \\ u_{\widehat{p}} & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & u_0 \\ & & \ddots & \ddots & u_1 \\ & & & \ddots & \vdots \\ & & & & u_{\widehat{p}} \end{bmatrix} \begin{bmatrix} s_0 \\ s_1 \\ \vdots \\ s_p \end{bmatrix} = \begin{bmatrix} s_0 & & & & \\ s_1 & s_0 & & & \\ \vdots & \ddots & \ddots & & \\ s_p & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & s_0 \\ & & \ddots & \ddots & s_1 \\ & & & \ddots & \vdots \\ & & & & s_p \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_{\widehat{p}} \end{bmatrix}. \tag{15}$$

Denote by U and S the two matrices in the above equation and observe that they have size $(m + n + 1) \times (p + 1)$ and $(m + n + 1) \times (\widehat{p} + 1)$. Since $s_p = 1$ and $u_{\widehat{p}} = a_n$, then $s'_p = u'_{\widehat{p}} = 0$. Set $\boldsymbol{u} = [u_1, \dots, u_{\widehat{p}}]^T$, $\boldsymbol{s} = [s_0, \dots, s_p]^T$. Taking derivatives with respect to λ, and denoting $\boldsymbol{e}_{m+1}$ the vector with null components except the $(m + 1)$-st which is 1, yields the system $-\boldsymbol{e}_{m+1} = U\boldsymbol{s}' + S\boldsymbol{u}'$ which can be rewritten as

$$[\widehat{U}, \widehat{S}] \begin{bmatrix} \widehat{\boldsymbol{s}}' \\ \widehat{\boldsymbol{u}}' \end{bmatrix} = -\boldsymbol{e}_{m+1}, \tag{16}$$

where $\widehat{\boldsymbol{s}} = [s_0, \dots, s_{p-1}]^T$, $\widehat{\boldsymbol{u}} = [u_0, \dots, u_{\widehat{p}-1}]^T$, and $\widehat{U}$ and $\widehat{S}$ are the matrices obtained from U and S, respectively, by removing the last column and the last row. This is a system formed by $m + n$ equations and $m + n$ unknowns. Moreover, the matrix $[\widehat{U}, \widehat{S}]$ is invertible since it is a resultant matrix associated with polynomials having no zeros in common. Therefore we have

$$\begin{bmatrix} \widehat{\boldsymbol{s}}' \\ \widehat{\boldsymbol{u}}' \end{bmatrix} = -[\widehat{U}, \widehat{S}]^{-1} \boldsymbol{e}_{m+1}.$$

Thirdly, we explain how to compute G' using the derivatives of $s_0, \dots, s_{p-1}$. We rely on the Barnett factorization [1] that provides an LU factorization of the matrix $G = F^p$:

$$F^p = -\mathcal{L}^{-1}\mathcal{U}, \quad \mathcal{L} = \begin{bmatrix} s_p & & & \\ s_{p-1} & s_p & & \\ \vdots & \ddots & \ddots & \\ s_1 & \dots & s_{p-1} & s_p \end{bmatrix}, \quad \mathcal{U} = \begin{bmatrix} s_0 & s_1 & \dots & s_{p-1} \\ & s_0 & \ddots & \vdots \\ & & \ddots & s_1 \\ & & & s_0 \end{bmatrix}, \tag{17}$$

where $\mathcal{L}$ and $\mathcal{U}$ are lower triangular and upper triangular Toeplitz matrices, respectively. Applying the Barnett factorization (17) to our problem, we have

$$(F^p)' = -\mathcal{L}^{-1}\mathcal{U}' + \mathcal{L}^{-1}\mathcal{L}'\mathcal{L}^{-1}\mathcal{U}, \tag{18}$$

where $\mathcal{L}'$ and $\mathcal{U}'$ are the derivatives of $\mathcal{L}$ and $\mathcal{U}$, respectively, that are determined by the derivative s_j', $j = 0, 1, \ldots, p$. We may observe that the cost of computing $(F^p)'$ by means of (18) amounts to $O(p^3)$ arithmetic operations which, due to the triangular Toeplitz structure and to the fast algorithms for triangular Toeplitz matrix inversion and for Toeplitz-vector multiplication can be lowered to $O(p^2 \log p)$.

Finally, we discuss on how to compute G^j and $(G^j)'$ given G'. From the relation $G^j = G^{j-1}G$ we obtain $(G^j)' = (G^{j-1})'G + G^{j-1}G'$. This expression allows us to compute $(G^i)'$ and G^i for $i = 1, \ldots, k$ according to the following equations

$$\begin{aligned} G^i &= GG^{i-1} \\ (G^i)' &= (G^{i-1})'G + G^{i-1}G' \end{aligned} \qquad i = 2, \ldots, k.$$

Clearly, the cost of this computation is $2(k-1)$ matrix multiplications and $k-1$ matrix additions, for an overall cost of $2(k-1)p^3 + O(kp^2)$ arithmetic operations.

In our implementation, we have adopted the algorithm based on the Barnett factorization for its simplicity, but other effective techniques can be used. For instance, a different approach is based on the structure of F and on the fact that $F' = -\boldsymbol{e}_p \boldsymbol{s}'^T$. Indeed, $(F^k)'$ is such that $(F^k)' = (F^{k-1})'F + F^{k-1}F'$, $F' = -\boldsymbol{e}_p \boldsymbol{s}'^T$. This implies that

$$(F^k)' = (F^{k-1})'F + \boldsymbol{f}_{k-1}\boldsymbol{s}'^T, \quad \boldsymbol{f}_k = F\boldsymbol{f}_{k-1}, \quad k = 1, \ldots, p,$$

where $\boldsymbol{f}_0 = -\boldsymbol{e}_p$. A careful computational analysis shows that this computation can be performed in $O(p^2)$ arithmetic operations.

A slightly different approach can be carried out as follows. Recall that the last row of F is $-\boldsymbol{s}^T$ and that $F' = -\boldsymbol{e}_p \boldsymbol{s}'^T$. Given $\boldsymbol{s}$ and $\boldsymbol{s}'$, write

$$(F^p)' = -\sum_{\ell=0}^{p-1} F^{p-1-\ell}\boldsymbol{e}_p \boldsymbol{s}'^T F^\ell = -\sum_{\ell=0}^{p-1} \boldsymbol{v}_{p-1-\ell}\boldsymbol{s}'^T F^\ell,$$

where $\boldsymbol{v}_\ell = F^\ell \boldsymbol{e}_p$. Observe that the vector $\boldsymbol{v}_\ell$ is such that $\boldsymbol{v}_\ell = \begin{bmatrix} 0 \ldots 0 \sigma_p \ldots \sigma_{p-\ell} \end{bmatrix}^T$, with $\sigma_p = 1$ and $\sigma_{p-\ell} = -\sum_{h=1}^{\ell} s_{p-h}\sigma_{p-\ell+h}$, for $\ell = 1, \ldots, p-1$. For the rows $\boldsymbol{r}_1^T, \ldots, \boldsymbol{r}_p^T$ of $(F^p)'$ we have

$$\begin{aligned} \boldsymbol{r}_\ell^T &= \sigma_{p-\ell+1}\boldsymbol{s}'^T + \boldsymbol{r}_{\ell-1}^T F \\ &= \sigma_{p-\ell+1}\boldsymbol{s}'^T + (F^p)'_{\ell-1,p}\boldsymbol{s}^T + [0\ (F^p)'_{\ell-1,1}\ \cdots\ (F^p)'_{\ell-1,p-1}], \quad \ell = 2, \ldots, p \end{aligned}$$

and $\boldsymbol{r}_1^T = \sigma_p \boldsymbol{s}'^T = \boldsymbol{s}'^T$. The cost of the procedure is given by the computation of $\sigma_1, \ldots, \sigma_p$ that requires $p^2 - p$ operations, and the recursion for $\boldsymbol{r}_1^T, \ldots, \boldsymbol{r}_p^T$ that requires about $4p^2$ operations.

4.2.3 Convergence of Newton's iteration

We have seen that in the Vandermonde formulation, the function $f_V(\lambda)$ is holomorphic in $\mathbb{C} \setminus a(\mathbb{T})$ as long as the roots of the Laurent polynomial $a(z) - \lambda$ are simple. Here we prove that the function $f_F(\lambda)$ is holomorphic in $\mathbb{C} \setminus a(\mathbb{T})$ under no additional condition. We rely on the implicit function theorem for functions of complex variable given in the following form [22, Theorem 15].

Theorem 3 *Let $F : \mathcal{V} \subset \mathcal{C}^k \times \mathcal{C}^q \to \mathcal{C}^q$ be a holomorphic mapping such that the linear mapping $\frac{\partial F}{\partial w}(z_0, w_0) : \mathcal{C}^q \to \mathcal{C}^q$ is invertible, where $(z_0, w_0) \in \mathcal{V}$. Then there are neighborhoods $\mathcal{U}$ and $\mathcal{A}$, $(z_0, w_0) \in \mathcal{U}$, $z_0 \in \mathcal{A}$, and a holomorphic mapping $g : \mathcal{A} \to \mathbb{C}^q$, such that $F(z, w) = F(z_0, w_0)$ if and only if $w = g(z)$ for $(z, w) \in \mathcal{U}$.*

Observe that for $\lambda \in \Omega$ the winding number of $a(z) - \lambda$ is constant, where Ω is a connected component of $\mathbb{C} \setminus a(\mathbb{T})$. Therefore, the polynomial $z^m(a(z) - \lambda)$ has $p = m + w$ roots of modulus less than 1 and $\widehat{p} = m + n - p$ roots of modulus greater than 1. Thus, there exists the Wiener-Hopf factorization $z^m(a(z) - \lambda) = s(z)u(z)$, where $s(z)$ is the monic polynomial of degree p, with coefficients s_i, $i = 0, \ldots, p$, having roots of modulus less than 1, while $u(z)$, of degree $\widehat{p}$ and coefficients u_i, $i = 0, \ldots, \widehat{p}$, has roots of modulus greater than 1. Consider the function $F(\lambda; s_0, \ldots, s_{p-1}, u_0, \ldots, u_{\widehat{p}}) = Us - \widehat{a} = Su - \widehat{a}$, where $s = (s_0, \ldots, s_{p-1}, 1)^T$, $u = (u_0, \ldots, u_{\widehat{p}})^T$, $\widehat{a} = (a_{-m}, \ldots, a_{-1}, a_0 - \lambda, a_1, \ldots, a_n)^T$, and where the matrices U and S are defined in (15). The function F is defined in $\mathbb{C} \times \mathbb{C}^{m+n+1}$ and takes values in $\mathbb{C}^{m+n+1}$. A direct computation shows that the matrix of partial derivatives of F with respect to s_i and to u_j is given by $[\widetilde{U}, S]$, where $\widetilde{U}$ is the matrix obtained by removing the last column of U. This matrix is invertible since its last row is $[0, \ldots, 0, 1]$ and the leading principal submatrix of size $m + n$ coincides with $[\widehat{U}, \widehat{S}]$ in (16) that is invertible.

Therefore, we may apply Theorem 3 to the function F with $k = 1$, $q = m + n + 1$, where $F(z_0, w_0) = 0$, and conclude with the following result.

Theorem 4 *Let Ω be any connected component of $\mathbb{C} \setminus a(\mathbb{T})$. Then, for $\lambda \in \Omega$ the function $f_F(\lambda) = \det \boldsymbol{\Phi}_F(\lambda)$ is holomorphic.*

5 Choosing the initial approximation

The algorithms presented in the previous sections can be used for refining a given approximation to an isolated eigenvalue of a QT matrix A, once an initial approximation is available. In this section, we investigate the problem of determining initial approximations to each isolated eigenvalue of A. More specifically, we show that, if A is Hermitian then for any isolated eigenvalue λ of A, and for any $\epsilon > 0$ there exists an integer N and an eigenvalue μ of the $N \times N$ leading principal submatrix A_N (finite section) of A such that $|\lambda - \mu| \leq \epsilon$. That is, for each isolated eigenvalue λ of A we

may find a sufficiently close approximation to λ among the eigenvalues of the $N \times N$ matrix A_N for a sufficiently large value of N.

For non-Hermitian matrices we have a weaker result: we show that for any eigenvalue λ of A and for any positive ϵ, there exists $N_0 > 0$ such that for any $N \geq N_0$, λ belongs to the ϵ-pseudospectrum sp_ϵ of A_N defined as $\mathrm{sp}_\epsilon(A_N) = \{z \in \mathbb{C} \ : \ \|(A_N - zI)^{-1}\| \geq \epsilon^{-1}\}$.

This fact enables us to implement a heuristic approach that, given A, selects a sufficiently large value of N, computes all the eigenvalues of A_N and applies to each eigenvalue of A_N one of the fixed point methods described in the previous section, and finally selects the values for which the numerical convergence occurs.

Since we do not have an explicit formal relation between ϵ and N, and since we do not have a theoretical bound to the radius of the convergence neighborhood of Newton's iteration, this strategy remains a heuristics approach. Nevertheless, from our implementation and from the experiments that we performed, this strategy turns out to be practically effective.

5.1 The case of Hermitian matrices

If A is Hermitian then the Bauer-Fike theorem provides a helpful tool to show that the isolated eigenvalues of A can be approximated by the eigenvalues of A_N.

Let $A = T(a) + E$ be a QT matrix, $a(z) = \sum_{j=-m}^{n} a_j z^j$, E compact correction with support $h_1 \times h_2$, i.e., its entries outside the leading $h_1 \times h_2$ submatrix are zero. Let A_N be the $N \times N$ leading principal submatrix of A. Let $A\boldsymbol{v} = \lambda \boldsymbol{v}$ be such that λ is an isolated eigenvalue of A and $\boldsymbol{v} = (v_i)$ has exponential decay, i.e., $\lim_j |v_j|^{\frac{1}{j}} = \xi$ for $0 < \xi < 1$, and $\sum_j |v_j|^2 = 1$. Denote by $Y \in \mathbb{C}^{n \times n}$ the lower triangular Toeplitz matrix whose first column is $(a_n, \ldots, a_1)^T$. Due to the exponential decay of v_i, for any $\epsilon > 0$ there exists $N_0 > 0$ such that for any $N \geq N_0$ it holds that $\|Y\boldsymbol{w}_N\| \leq \|Y\| \, \|\boldsymbol{w}_N\|\epsilon$, where $\boldsymbol{w}_N = (v_{N+1}, \ldots, v_{N+n})^T$.

If $N > \max(m, n, h_1, h_2, N_0)$ set $\boldsymbol{v}_N = (v_1, \ldots, v_N)^T$, $\boldsymbol{u}_N = [0_{N-n}; Y\boldsymbol{w}_N]$ and rewrite the condition $A\boldsymbol{v} = \lambda \boldsymbol{v}$ as

$$A_N \boldsymbol{v}_N + \boldsymbol{u}_N = \lambda \boldsymbol{v}_N. \tag{19}$$

Defining $C_N = \frac{1}{\boldsymbol{v}_N^* \boldsymbol{v}_N} \boldsymbol{u}_N \boldsymbol{v}_N^*$, we may rewrite (19) as $(A_N + C_N)\boldsymbol{v}_N = \lambda \boldsymbol{v}_N$. That is, λ is eigenvalue of an $N \times N$ matrix which differs from A_N by the correction C_N. Observe also that the matrix C_N satisfies the inequality $\|C_N\| \leq \frac{1}{\|\boldsymbol{v}_N\|} \|Y\| \cdot \|\boldsymbol{w}_N\| \leq \frac{1}{\|\boldsymbol{v}_N\|} \epsilon$.

That is, we may look at an isolated eigenvalue λ of A as an eigenvalue of a finite matrix obtained by perturbing the finite matrix A_N. Therefore we may invoke the classical perturbation theorems for eigenvalues of finite matrices. For instance we can apply the Bauer-Fike theorem.

Theorem 5 (Bauer-Fike) *Let A be a diagonalizable matrix, i.e., there exists S such that $S^{-1}AS = D$, D diagonal, and let $\|\cdot\|$ be an absolute norm. Then, for any eigenvalue λ of $A + C$ there exists an eigenvalue μ of A such that $|\lambda - \mu| \leq \|C\| \cdot \|S\| \cdot \|S^{-1}\|$.*

Observe that the p-norms are absolute, i.e., $\|\boldsymbol{v}\| = \|(|v_i|)\|$ for any $\boldsymbol{v} = (v_i)$.

Therefore, if A is Hermitian, then A_N is Hermitian and consequently S can be chosen to be unitary so that for the 2-norm we have $\|S\| = \|S^{-1}\| = 1$ and by the Bauer-Fike theorem we may conclude that for any eigenvalue λ of $A_N + C_N$, that is for any isolated eigenvalue λ of A, there exists an eigenvalue λ_N of A_N such that $|\lambda - \lambda_N| \leq \|C_N\| \leq \epsilon/\|\boldsymbol{v}_N\|$. Therefore, $|\lambda_N - \lambda| \to 0$ exponentially with N.

5.2 The general case

The case of nonsymmetric matrices seems more tricky. In fact, the Bauer-Fike theorem can be still applied if A_N is diagonalizable but the bound turns into

$$|\lambda_N - \lambda| \leq \frac{\|\boldsymbol{w}_N\|}{\|\boldsymbol{v}_N\|}\|Y\| \cdot \|S_N\| \cdot \|S_N^{-1}\|$$

where $S_N^{-1}A_N S_N = D$ is a diagonal matrix. Therefore, in this case we need that A_N be diagonalizable and that $\lim_N \|\boldsymbol{w}_N\| \cdot \|S_N\| \cdot \|S_N^{-1}\| = 0$. This condition is satisfied if, say, the condition number $\|S_N\| \cdot \|S_N^{-1}\|$ is uniformly bounded from above by a constant.

Unfortunately, the condition number of S_N may grow very fast with N. Think for instance to the tridiagonal matrix $\mathrm{trid}(1/2, 0, 2) = \widehat{D}^{-1}\mathrm{trid}(1, 0, 1)\widehat{D}$ where $\widehat{D} = \mathrm{diag}(1, 2, 2^2, \ldots, 2^{N-1})$, having $S_N = Q_N\widehat{D}$ as eigenvector matrix with Q orthogonal. Clearly $\mathrm{cond}\,(S_N) = \mathrm{cond}\,(\widehat{D}) = 2^{N-1}$.

On the other hand, from (19) we find that if λ is not eigenvalue of A_N, then $(A_N - \lambda I)^{-1}\boldsymbol{u}_N = -\boldsymbol{v}_N$, that is, $\|(A_N - \lambda I)^{-1}\| \geq \|\boldsymbol{v}_N\|/\|\boldsymbol{u}_N\| \geq \gamma\epsilon^{-1}$ for some constant $\gamma > 0$. This implies that $\lambda \in \mathrm{sp}_{\gamma^{-1}\epsilon}(A_N)$ for any $N > N_0$.

Therefore, we may say that for any eigenvalue λ of the QT matrix A and for any $\epsilon > 0$ there exists an integer N_0 such that for any $N \geq N_0$ the matrix A_N has an ϵ -pseudo eigenvalue μ equal to λ. This fact motivates using the eigenvalues of A_N, for sufficiently large values of N, as starting approximations for Newton's iteration.

6 Implementation and numerical results

We have implemented the algorithms described in the previous sections in Matlab and added them to the CQT-Toolbox of [9]. The functions allow the computation in high precision arithmetic relying on the package `Advanpix`, see https://advanpix.com. The main functions are `eig_single` and `eig_all`. The function `eig_single` computes the approximation of a single eigenvalue by relying on Newton's iteration, in both the Vandermonde and the Frobenius version, starting from a given approximation λ_0. The function `eig_all` computes approximations to all the eigenvalues starting from the eigenvalues of the matrix A_N for $N = \gamma \max(h_1, h_2, m + n)$, where γ is a small constant that can be set by the user, by default $\gamma = 3$. The iterations are halted if the modulus of the difference between two subsequent approximations is less than $10^3 u$, where u is the machine precision,

and if this value is not smaller than the value obtained at the previous step. After the halting condition is satisfied, a further Newton step is applied to refine the approximation. The iterations are halted with the failure flag if $\text{wind}(\lambda_k) \neq \text{wind}(\lambda_{k-1})$ for some k or if $|\lambda_k|$ is larger than $\|A\|_\infty$ or if the maximum number of 20 iterations has been reached. For detailed information, including the description of other auxiliary functions and optional parameters, see https://numpi.github.io/cqt-toolbox, while one can download the software at https://github.com/numpi/cqt-toolbox.

6.1 The tests

We have performed several tests to validate our algorithms. Here, we describe the results of the most meaningful ones. In the following, we denote by `am` and `ap` two vectors such that `am`$= [a_0, a_{-1}, \ldots, a_{-m}]$ and `ap`$= [a_0, a_1, \ldots, a_n]$, where $a(z) = \sum_{i=-m}^{n} a_i z^i$ is the Laurent polynomial associated with the QT matrix $A = T(a) + E$. We refer to Algorithm V for the Vandermonde approach and Algorithm F for the Frobenius approach. The tests have been run on a laptop with Intel I5 CPU and with Matlab version R2021b.

In Test 1 we have set $m = 3$ and $n = 2$, where `am`$= [0, -1, 1, -1]$, `ap`$= [0, -1, -1]$. We have applied two kinds of corrections, namely, the 20×100 matrix E_2 having null entries except the last column which is equal to $[1, 2, 3, \ldots, 20]^T$, and the 3×100 matrix E_1 having null entries except in the last column which is equal to $8[1, 2, 3]^T$. We refer to these two corrections as Case 1 and Case 2, respectively.

In Test 2 we have set $m = 7$ and $n = 2$ where `am`$= [0, -1, 1, -1, 0, 0, 0, 1]$, `ap`$= [0, -1, -1]$. We have applied two kinds of corrections, namely, E_1 and E_2, where E_1 is the same as in Test 1, while E_2 has size 7×100 with null entries except the last column which is equal to $8[1, 2, 3, \ldots, 7]^T$. We refer to these two corrections as Case 1 and Case 2, respectively.

Test 3 has been designed in order to show that the Vandermonde approach may strongly suffer of numerical instability when the characteristic equation $a(z) - \lambda = 0$ has some clustered roots that, consequently, are ill-conditioned. For this test, we have constructed $a(z)$ in terms of a Mignotte-like polynomial [32]. More precisely, we set $a(z) = z^{-m} b(z)$, where $b(z)$ is of the form $b(z) = (10^{-1} + z)^3 + 10 z^{n+m}$. This polynomial has a very tight cluster of 3 zeros close to 10^{-1}. In our test we set $m = 10$ and $n = 2$ and $E = 10^{-5}[0_{12}, I_{12}]$.

In all the three tests the matrix A is not symmetric.

6.2 Details on the implementation

The algorithms have been applied in the double precision floating point arithmetic. The basins of attraction have been constructed as follows. A generic point in the picture, corresponding to the complex number λ_0 has been colored with a color depending on the limit of the sequence generated by fixed point iteration $\lambda_{k+1} = g(\lambda_k)$ for $k \geq 0$. Different colors, randomly generated, have been used for different limits. Different levels of gray have been used to denote that the iteration has been halted with

no convergence. The color light green has been used for the values λ_0 belonging to a continuous set of eigenvalues.

6.3 The results

In the figures where the eigenvalues are reported, red circles indicate the eigenvalues of the finite section A_N, blue dots represent isolated eigenvalues of A, while red circles containing a green dot represent eigenvalues of A_N that belong to a continuous set of eigenvalues. The light blue curve denotes the set $a(\mathbb{T})$. In Fig. 4 displaying the basins of attraction, the light green area indicates a continuous set of eigenvalues. For this set of figures, Algorithm V has been applied.

Figure 3 displays the eigenvalues of $A + E_1$, for the matrix of Test 1, and the basins of attraction of Newton's iteration, together with a zoom of a specific area.

Figure 4 displays the analogous images for the matrix $A + E_2$ of Test 1. Here, it is interesting to observe the existence of a connected component formed by a continuous set of eigenvalues denoted by a green triangle-shaped figure. Observe also that the corresponding red circles in this component contain a green dot.

The smallest value of N_0 for which the number of computed eigenvalues is constant for $N \geq N_0$ is $N_0 = 400$ for the Case 1, while it is $N_0 = 200$ for the Case 2. In

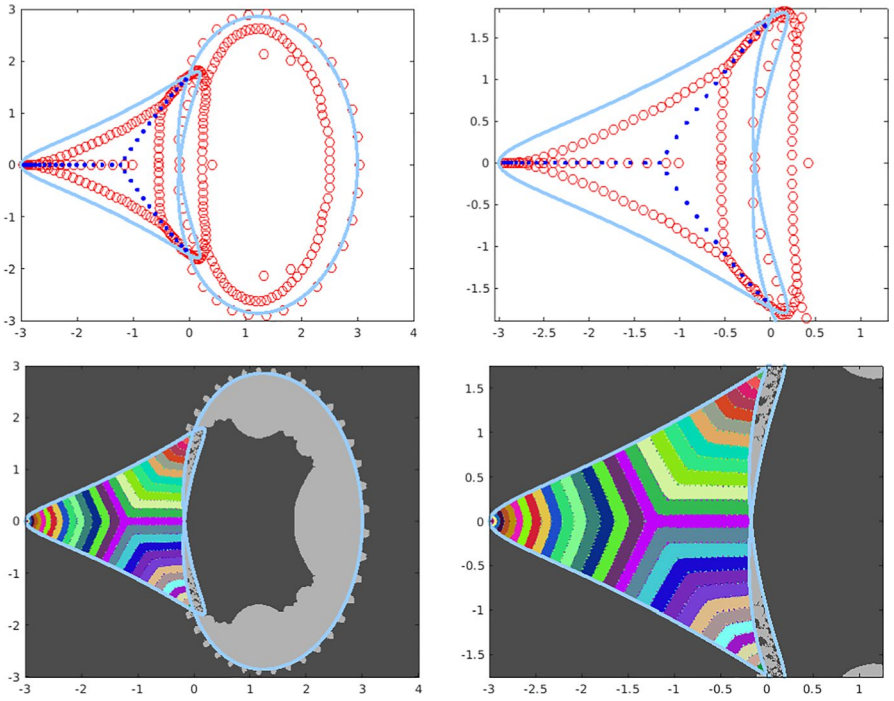

Fig. 3 Test 1, Case 1: Eigenvalues of the QT matrix A (blue dots) and of the finite section A_N (red circles), together with the basins of attraction for Newton's iteration computed by Algorithm V. On the right the zoom of a portion

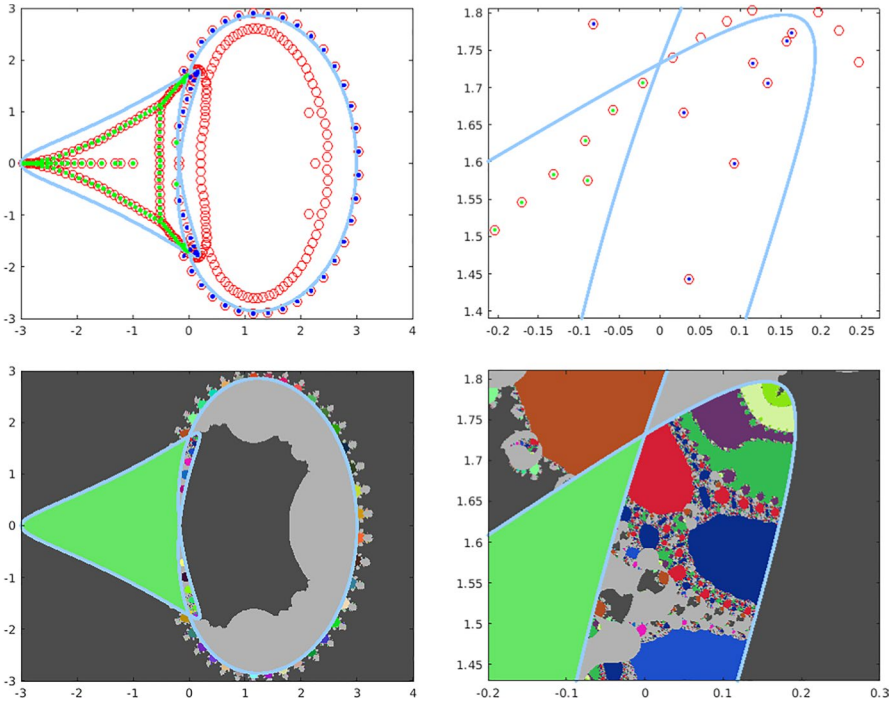

Fig. 4 Test 1, Case 2: Eigenvalues of the QT matrix A (blue dots) and of the finite section A_N (red circles), together with the basins of attraction for Newton's iteration computed by Algorithm V. On the right the zoom of a portion

both cases, the geometry of the basins of attraction, together with the distribution of the eigenvalues of A_N, explains why Newton's iteration converges to all the eigenvalues, when starting from the eigenvalues of A_N for a quite small value of N, even though the latter eigenvalues are far from the eigenvalues of A. This latter property is more evident in Case 1, where several blue dots are not contained inside red circles, see Fig. 3, zoomed part.

The number of iterations to arrive at convergence is quite small and is the same for both algorithms. Namely, concerning Case 1, it ranges from 3 to 18 with the avergae value of 7.5; concerning Case 2, it ranges from 3 to 10 with average 3.3.

Table 1 Test1, Case 1: Distances of some eigenvalues of A from the closest eigenvalue of A_N for different values of N. A "–" denotes a value below 1.e-15

$\lambda \setminus N$	200	400	800	1600	3200	6400
–4.0e-01±1.2e+00i	4.1e-04	1.3e-08	–			
–3.1e-01±1.3e+00i	3.0e-03	3.9e-06	5.3e-12	–		
–2.2e-01±1.5e+00i	5.4e-03	6.1e-05	5.9e-09	–		
–1.4e-01±1.6e+00i	2.6e-02	2.0e-03	2.7e-05	1.9e-10	–	
–5.9e-02±1.6e+00i	6.5e-02	3.5e-02	1.1e-02	3.5e-04	9.2e-07	3.8e-13

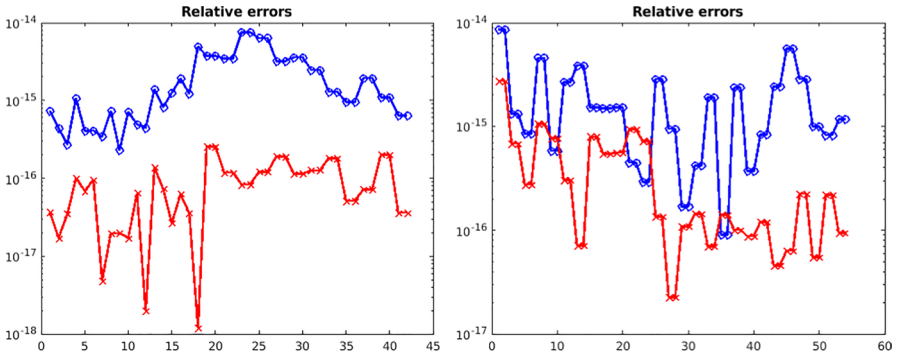

Fig. 5 Test 1. Relative errors in each eigenvalue computed with Algorithm V (blue circle) and with Algorithm F (red cross). Case 1 and case 2 on the left and on the right, respectively. Eigenvalues are sorted with respect to the real part

Another interesting issue to investigate, independently of the algorithm used, is to analyze how large must be N in order that the eigenvalues of A_N approximate all the eigenvalues of A within the machine precision $u = 2.22e{-}16$ so that no step of Newton's iteration would be necessary. It turns out that for the Test 1, Case 1, almost all the eigenvalues are well approximated already for $N = 800$, while there are few eigenvalues that require a pretty larger size. Table 1 shows a few significant cases. Typically, the eigenvalues closest to the light blue curve are the ones that need a large value of the truncation level N to be properly approximated by a corresponding eigenvalue of A_N. For instance, from Table 1 it turns out that $N = 3200$ is not enough to approximate the rightmost eigenvalue. Even $N = 6400$ does not provide a full accuracy approximation. A similar situation holds for the Case 2.

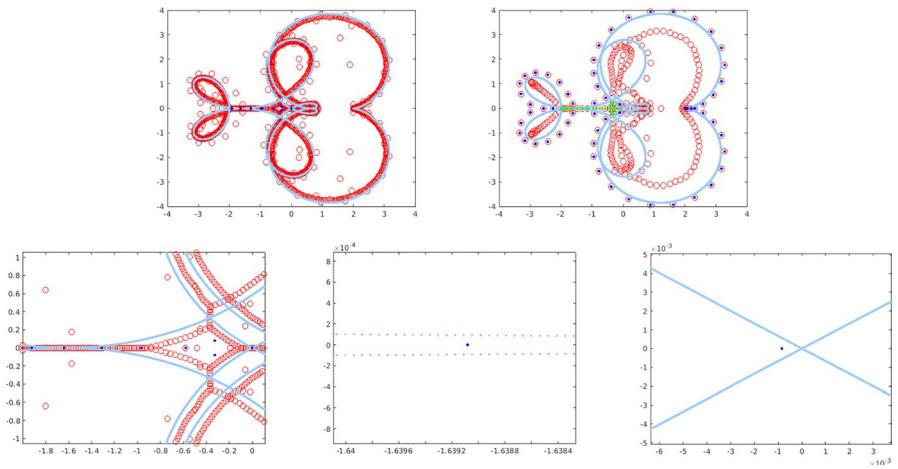

Fig. 6 Test 2, In the first line the eigenvalues of Case 1 (left) and Case 2 (right) are displayed with a blue dot. In the second line, some portion of the domain where the eigenvalues of Case 1 are located are displayed; more specifically, from the left, the area with all the eigenvalues, the second leftmost eigenvalue, and the last rightmost eigenvalue are zoomed, respectively

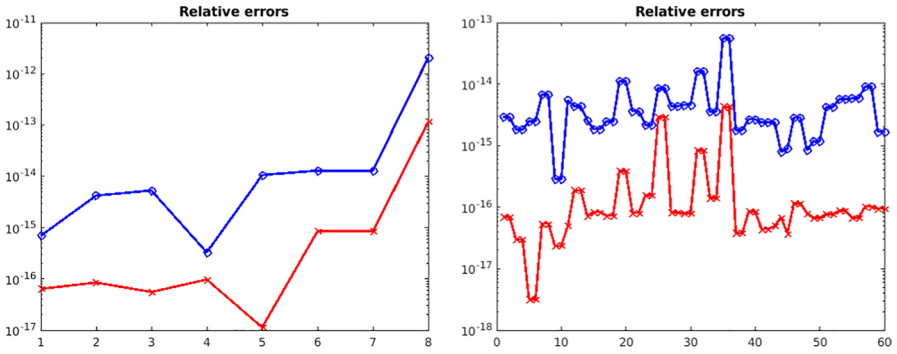

Fig. 7 Relative errors in each eigenvalue computed with Algorithm V (blue circle) and with Algorithm F (red cross). Case 1 and case 2 on the left and on the right, respectively. Eigenvalues are sorted with respect to the real part

Concerning the accuracy of approximation, Fig. 5 shows the relative errors of approximating the eigenvalues of A with the Vandermonde approach (blue circle) and with the Frobenius approach (red cross) in the two cases of Test 1. Here, the eigenvalues have been sorted according to the real part. The relative errors have been obtained by comparing the eigenvalues computed in the double precision floating point arithmetic with those computed in the quadruple precision relying on Advanpix. Observe that the results obtained by the Frobenius version are generally more accurate than the ones obtained with the Vandermonde version.

Finally, concerning the CPU time, the two algorithms have similar performances even though, for this test, Algorithm F generally requires a double time.

Test 2, Case 1, points out in a more evident manner that the eigenvalues of A which are close to the curve $a(\mathbb{T})$ can be hardly approximated by the eigenvalues of a finite section A_N of A, unless N is extremely large. In fact, as clearly shown in Fig. 6 and in the zoomed areas, out of the 8 eigenvalues of A, there is a group of few eigenvalues that lie very close to the light blue curve. In particular, the second (from the left) eigenvalue and the last one. The distances of these eigenvalues to the closest eigenvalue of A_N for different values of N are reported in Table 2. It turns out that in order to approximate such eigenvalues within the machine precision u without applying Newton's iteration,

Table 2 Test 2, Case 1: Distances of the real eigenvalues of A from the closest eigenvalue of A_N for different values of N. A "–" denotes a value below 1.e-15

$\lambda \setminus N$	400	1600	6400	25600	102400	409600	1638400
–1.9	1.4e-01	8.3e-02	4.1e-05	–			
–1.6	7.6e-01	2.3e-01	9.8e-02	5.8e-02	2.9e-03	2.6e-03	1.4e-07
–1.3	4.2e-01	1.2e-01	1.5e-04	–			
–9.6e-01	1.6e-01	5.2e-06	–				
–5.8e-01	3.0e-03	7.3e-11	–				
–8.5e-04	6.8e-02	1.0e-01	9.2e-02	7.8e-03	2.9e-04	5.3e-13	–

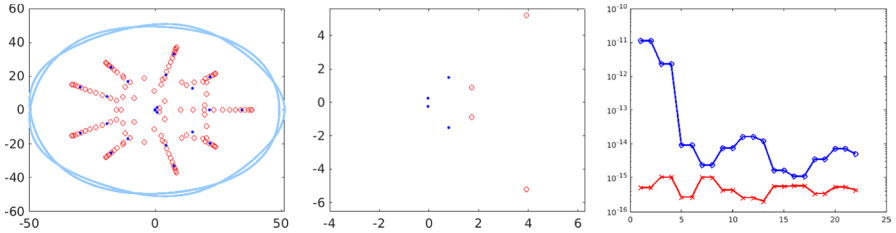

Fig. 8 Test 3. From the left: Geometry of the eigenvalues with a zoom of the cluster computed by Algorithm F; relative errors for each eigenvalue computed by Algorithm V (blue circle) and by Algorithm F (red cross), where eigenvalues are sorted by increasing modulus

one would need truncation levels larger than 1.6 millions, whereas Newton's iteration converges quickly just starting from the eigenvalues of A_N, with $N = 3200$.

Also in this test, the number of iterations required by Algorithm V and Algorithm F is the same. Namely, for Case 1 it ranges between 5 and 12 with average value 7.25, for Case 2 it ranges between 2 and 4 with average value 3.0. Algorithm F turns out to be more accurate than Algorithm V as shown in Fig. 7.

Concerning Test 3, the matrix A has a set $\mathcal{S}$ of 22 eigenvalues, shown in Fig. 8, that can be grouped into 3 subsets $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$. The subset $\mathcal{S}_1$ is formed by 4 entries of modulus in the range [0.25, 1.7], while $\mathcal{S}_2$ and $\mathcal{S}_3$ are formed by 9 entries of modulus roughly 20, and 30, respectively. Recall that the Mignotte-like polynomial $z^m a(z)$ has a tight cluster formed by three ill-conditioned zeros. For $\lambda \in \mathcal{S}$ the polynomial $z^m(a(z) - \lambda)$ still has a cluster of ill-conditioned zeros, where the cluster is tighter and consequently the zeros are more ill-conditioned the smaller is $|\lambda|$. This explains why the errors of the algorithm based on the Vandermonde formulation are much higher in the leftmost part of the graph shown in Fig. 8.

7 Conclusions and open problems

We have reformulated the problem of computing the eigenvalues of a QT matrix as a nonlinear eigenvalue problem, for which Newton's method has been analyzed, both in the Vandermonde and in the Frobenius version. We use the eigenvalues of the truncated matrix A_N for a moderate N as initial approximation for starting the iteration. Our approach is shown to be effective by numerical tests, while approximating all the eigenvalues to the machine precision directly from the eigenvalues of the truncated matrix A_N, without using Newton's iteration, is shown to be infeasible due to the huge values needed for N. The algorithm based on the Frobenius formulation turned out to be more accurate even though slightly slower. A Matlab implementation of the algorithm has been provided and the software has been included in the CQT-toolbox of [9].

In order to make the software more robust and effective we plan to provide an optimized implementation of polynomial spectral factorization relying on the algorithms of [16] and [17]. Another important issue is to find theoretical estimates of

the truncation parameter N that guarantees the approximation to all the eigenvalues of A, starting from those of A_N. Other approaches to solving the nonlinear eigenvalue problem, say the ones based on rational approximation, could be the subject of subsequent research.

Acknowledgements The first author wishes to thank Matthew Colbrook and Mark Embree for helpful conversations and comments. The second author would like to thank Dimitri Breda for useful discussions.

Funding This work has been partially supported by University of Pisa's project PRA_2020_61, and by GNCS of INdAM.

Data availability The data used in this paper are available from the corresponding author under request.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Barnett, S.: Polynomials and linear control systems, volume 77 of Monographs and Textbooks in Pure and Applied Mathematics. Marcel Dekker, Inc., New York (1983)
2. Bini, D.A., Fiorentino, G., Gemignani, L., Meini, B.: Effective fast algorithms for polynomial spectral factorization. Numer. Algorithms **34**(2–4), 217–227 (2003)
3. Bini, D.A., Iannazzo, B., Meng, J.: Algorithms for approximating means of semi-infinite quasi-Toeplitz matrices. In: Nielsen, B.F. (ed) Geometric Science of Information, GSI 2021, volume 12829 of Lecture Notes in Computer Science, pp. 405–414. Springer
4. Bini, D.A., Iannazzo, B., Meng, J.: Geometric means of quasi-Toeplitz matrices. arXiv preprint. (2021)
5. Bini, D.A., Latouche, G., Meini, B.: Numerical methods for structured Markov chains. Numerical Mathematics and Scientific Computation. Oxford University Press, New York (2005)
6. Bini, D.A., Massei, S., Meini, B.: Semi-infinite quasi-Toeplitz matrices with applications to QBD stochastic processes. Math. Comp. **87**(314), 2811–2830 (2018)
7. Bini, D.A., Massei, S., Meini, B., Robol, L.: On quadratic matrix equations with infinite size coefficients encountered in QBD stochastic processes. Numer. Linear Algebra Appl. **25**(6), 2128, 12 (2018)
8. Bini, D.A., Massei, S., Meini, B., Robol, L.: A computational framework for two-dimensional random walks with restarts. SIAM J. Sci. Comput. **42**(4), A2108–A2133 (2020)
9. Bini, D.A., Massei, S., Robol, L.: Quasi-Toeplitz matrix arithmetic: a MATLAB toolbox. Numerical Algorithms **81**(2), 741–769 (2019)
10. Bini, D.A., Meini, B.: On the exponential of semi-infinite quasi-Toeplitz matrices. Numer. Math. **141**(2), 319–351 (2019)
11. Bini, D.A., Meini, B., Meng, J.: Solving quadratic matrix equations arising in random walks in the quarter plane. SIAM J. Matrix Anal. Appl. **41**(2), 691–714 (2020)
12. Böttcher, A., Embree, M., Sokolov, V.I.: Infinite Toeplitz and Laurent matrices with localized impurities. Linear Algebra Appl. **343—-344,** 101–118 (2002)
13. Böttcher, A., Embree, M., Sokolov, V.I.: On large Toeplitz band matrices with an uncertain block. Linear Algebra Appl **366**, 87–97 (2003)
14. Böttcher, A., Grudsky, S.M.: Toeplitz matrices, asymptotic linear algebra, and functional analysis. Birkhäuser Verlag, Basel (2000)
15. Böttcher, A., Grudsky, S.M.: Spectral properties of banded Toeplitz matrices. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2005)
16. Böttcher, A., Halwass, M.: A Newton method for canonical Wiener-Hopf and spectral factorization of matrix polynomials. Electron. J. Linear Algebra **26**, 873–897 (2013)

17. Böttcher, A., Halwass, M.: Wiener-Hopf and spectral factorization of real polynomials by Newton's method. Linear Algebra Appl. **438**(12), 4760–4805 (2013)
18. Böttcher, A., Silbermann, B.: Introduction to large truncated Toeplitz matrices. Universitext. Springer-Verlag, New York (1999)
19. Breda, D., Liessi, D.: Approximation of Eigenvalues of Evolution Operators for Linear Renewal Equations. SIAM J. Numer. Anal. **56**(3), 1456–1481 (2018)
20. Colbrook, M..J., Roman Bogdan, B., Hansen, A.C.: How to compute spectra with error control. Phys. Rev. Lett **122**(25), 250201, 6 (2019)
21. Colbrook, M.J., Hansen, A.C.: On the infinite-dimensional QR algorithm. Numer. Math. **143**(1), 17–83 (2019)
22. D'Angelo, J.P.: Several complex variables and the geometry of real hypersurfaces. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL (1993)
23. Gander, W.: New algorithms for solving nonlinear eigenvalue problems. Comput. Math. Math. Phys. **61**(5), 761–773 (2021)
24. Garoni, C., Serra-Capizzano, S.: Generalized locally Toeplitz sequences: theory and applications, vol. I. Springer, Cham (2017)
25. Garoni, C., Serra-Capizzano, S.: Generalized locally Toeplitz sequences: theory and applications, vol. II. Springer, Cham (2018)
26. Gavin, B., Międlar, A., Polizzi, E.: FEAST eigensolver for nonlinear eigenvalue problems. J. Comput. Sci. **27**, 107–117 (2018)
27. Güttel, S., Tisseur, F.: The nonlinear eigenvalue problem. Acta Numer. **26**, 1–94 (2017)
28. Hochstenbach, M.E., Plestenjak, B.: Computing several eigenvalues of nonlinear eigenvalue problems by selection. Calcolo, 57(2), Paper No. 16, 25 (2020)
29. Jackson, J.R.: Networks of waiting lines. Operations Res. **5**, 518–521 (1957)
30. Kim, H.-M., Meng, J.: Structured perturbation analysis for an infinite size quasi-Toeplitz matrix equation with applications. BIT Numerical Mathematics **61**, 859–879 (2021)
31. Latouche, G., Ramaswami, V.: Introduction to matrix analytic methods in stochastic modeling. ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia, PA (1999)
32. Mignotte, M.: Some useful bounds. In: Computer algebra, pp. 259–263. Springer, Vienna (1983)
33. Neuts, M.F.: Matrix-geometric solutions in stochastic models: An algorithmic approach, volume 2 of Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, Md. (1981)
34. Ostrowski, A.: Recherches sur la méthode de Graeffe et les zéros des polynomes et des séries de Laurent. Acta Mathematica **72**, 99–155 (1940)
35. Ozawa, T.: Stability condition of a two-dimensional QBD process and its application to estimation of efficiency for two-queue models. Performance Evaluation **130**, 101–118 (2019)
36. Ozawa, T.: Asymptotic properties of the occupation measure in a multidimensional skip-free Markov-modulated random walk. Queueing Syst. **97**(1–2), 125–161 (2021)
37. Robol, L.: Rational Krylov and ADI iteration for infinite size quasi-Toeplitz matrix equations. Linear Algebra Appl. **604**, 210–235 (2020)
38. Schechter, M.: Basic theory of Fredholm operators. Ann. Scuola Norm. Sup. Pisa Cl. Sci. 21(3)261–280 (1967)
39. Webb, M., Olver, S.: Spectra of Jacobi operators via connection coefficient matrices. Commun. Math. Phys. **382**, 657–707 (2021)

Authors and Affiliations

D. A. Bini[1] · B. Iannazzo[2] · B. Meini[1] · J. Meng[3] · L. Robol[1]

D. A. Bini
dario.bini@unipi.it

B. Iannazzo
bruno.iannazzo@unipg.it

B. Meini
beatrice.meini@unipi.it

L. Robol
leonardo.robol@unipi.it

[1] University of Pisa, Pisa, Italy

[2] University of Perugia, Perugia, Italy

[3] Ocean University of China, Qingdao, China