IMA Journal of Numerical Analysis (2020) **00**, 1–14
doi:10.1093/imanum/draa049

On QZ steps with perfect shifts and computing the index of a differential-algebraic equation

Nicola Mastronardi*
*Istituto per le Applicazioni del Calcolo "M. Picone", sede di Bari,
Consiglio Nazionale delle Ricerche, Italy*
*Corresponding author. Email: nicola.mastronardi@cnr.it

and

Paul Van Dooren
Department of Mathematical Engineering, Catholic University of Louvain, Louvain-la-Neuve, Belgium
paul.vandooren@uclouvain.be

[Received on 21 November 2019; revised on 15 May 2020]

In this paper we revisit the problem of performing a *QZ* step with a so-called 'perfect shift', which is an 'exact' eigenvalue of a given regular pencil $\lambda B - A$ in unreduced Hessenberg triangular form. In exact arithmetic, the *QZ* step moves that eigenvalue to the bottom of the pencil, while the rest of the pencil is maintained in Hessenberg triangular form, which then yields a deflation of the given eigenvalue. But in finite precision the *QZ* step gets 'blurred' and precludes the deflation of the given eigenvalue. In this paper we show that when we first compute the corresponding eigenvector to sufficient accuracy, then the *QZ* step can be constructed using this eigenvector, so that the deflation is also obtained in finite precision. An important application of this technique is the computation of the index of a system of differential algebraic equations, since an exact deflation of the infinite eigenvalues is needed to impose correctly the algebraic constraints of such differential equations.

Keywords: QZ algorithm; eigenvalues; perfect shifts; index.

1. Introduction

The solution of systems of *implicit* differential equations

$$B\dot{x}(t) = Ax(t), \quad x(0) = x_0, \quad A, B \in \mathbb{R}^{n\times n} \tag{1.1}$$

is a standard problem occurring in the analysis of linear time-invariant dynamical systems. When the pencil $\lambda B - A$ is regular (meaning that $\det(\lambda B - A) \neq 0$ for at least one value of λ), then its finite eigenvalues are the roots of the polynomial $\det(\lambda B - A)$. If B is invertible, then these are also the eigenvalues of $B^{-1}A$ and its Jordan form yields additional information (such as the geometric and algebraic multiplicites of the eigenvalues), which describes the complete set of solutions of (1.1). In the case B is singular (and the pencil is regular), the Jordan form is replaced by the so-called Weierstrass form, which also defines the structure of the infinite eigenvalues. The structure of the eigenvalue at ∞ corresponds to the so-called *impulsive* solutions of (1.1) and the so-called *index* is the size of the corresponding largest Jordan block. This is particularly important when using numerical solvers for differential-algebraic equations (DAEs), since failing to recover the infinite structure exactly will lead to very large 'spurious' eigenvalues that will completely perturb the numerical simulation of the DAE (see Gantmacher, 1959; Mehrmann, 1991; Kunkel & Mehrmann, 2006, for further details). In the case of discrete-time dynamical systems, the structure at infinity is equally important since it defines compatibility constraints for the initial conditions of the system of difference equations.

The standard eigenvalue problem and the generalized eigenvalue problem are of course intimately related, and the problem of perfect shifts was already analyzed in Mastronardi & Van Dooren (2018) in the context of QR steps applied to unreduced Hessenberg matrices. In this paper we revisit the corresponding problem of perfect shifts in the QZ step, applied to a regular pencil $\lambda B - A$ in Hessenberg triangular form. In exact arithmetic, the QZ step applied to an unreduced Hessenberg triangular pencil moves the eigenvalue to the bottom of the pencil, while the rest of the pencil is maintained in Hessenberg triangular form, which then yields a deflation of the given eigenvalue. But in finite precision the QZ step gets 'blurred' and precludes the deflation of the given shift. We will show that when we first compute the corresponding eigenvector to sufficient accuracy, the QZ step can be constructed using this eigenvector, so that the deflation is obtained also in finite precision. We made the choice of using the right eigenvector in our analysis, which will move the deflated shift to the top of the pencil rather than to the bottom, but this also results in a deflation.

For the sake of simplicity, we consider only real matrix pencils and the deflation of their real eigenvalues. The application of finding the index of a set of DAEs with real coefficients, falls into this category since it amounts to deflating the infinite eigenvalues of a pencil $\lambda B - A$ or, as shown in Section 2, deflating the zero eigenvalues of a transformed pencil $\mu A - B$. The extension to complex pencils and the deflation of their complex eigenvalues is straightforward and is therefore not treated here. We will use the following notations. Matrices and submatrices are denoted by capital letters, i.e., A, B, H. The entry (i,j) of the matrix A is denoted by the lower case letter $a_{i,j}$. Vectors are denoted by bold letters, i.e., $\mathbf{a}$, $\mathbf{b}$, ... The identity matrix of order n is denoted by I_n and its ith column by $\mathbf{e}_i^{(n)}$ or, if there is no ambiguity, simply by I and $\mathbf{e}_i$, respectively. Generic entries different from zero in matrices or vectors are denote d by '×'. The machine precision is denoted by ε_M. We denote a Givens rotation by

$$G_i = \begin{bmatrix} I_{i-1} & & & \\ & c & -s & \\ & s & c & \\ & & & I_{n-i-1} \end{bmatrix}, \quad \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} c & -s \\ s & c \end{bmatrix}^T = I_2.$$

The paper is organized as follows. We recall the Weierstrass form in Section 2 and the Hessenberg triangular form in Section 3. We recall the derivation of a QZ step for a real shift in Section 4. In Section 5 we propose a definition of what should be a perfect shift QZ step, and in Section 6 we give sufficient conditions for achieving this. A scaling procedure to obtain these conditions is given in Section 7. A numerical example illustrating the accuracy of our method is described in Section 8. We end with some concluding remarks in Section 9.

2. Weierstrass form of a regular pencil

A pencil $\lambda B - A$ with coefficients $A, B \in \mathbb{R}^{n \times n}$ is said to be regular if its determinant is not identically zero for all λ. Such pencils have a canonical form under the group of invertible transformations on rows and columns, which is known as the Weierstrass canonical form. If $\lambda B - A$ is a real regular pencil then there exist (complex) invertible transformations S and T such that

$$S(\lambda B - A)T = \mathrm{diag}\{(\lambda - \lambda_1)I_{n_1} - N_{n_1}, \dots, (\lambda - \lambda_\ell)I_{n_\ell} - N_{n_\ell}, I_{n_\infty} - \lambda N_{n_\infty}\}, \tag{2.1}$$

where the $\lambda_i, i = 1, \dots, \ell$ are the distinct finite eigenvalues of $\lambda B - A$ and N_{n_i} are nilpotent bidiagonal matrices of the form

$$N_{n_i} = \mathrm{diag}\{J_{\nu_{(i,1)}}, \dots, J_{\nu_{(i,k_i)}}\}, \quad J_\nu = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} \in \mathbb{R}^{\nu \times \nu}.$$

This form gives the complete Jordan structure of the finite eigenvalues $\lambda_i, i = 1, \dots, \ell$ via the Jordan block sizes $\nu_{(i,j)}, j = 1, \dots, k_i$. In particular, the Jordan block sizes of the infinite eigenvalue $\lambda = \infty$ are given by the index set $\nu_{(\infty,j)}, j = 1, \dots, k_\infty$. The geometric multiplicity of the infinite eigenvalue is k_∞ (i.e. the number of blocks) and the algebraic multiplicity is just the sum of the indices $\sum_{j=1}^{k_\infty} \nu_{(\infty,j)}$. Finally, the index k of the system of DAEs (1.1) is the largest Jordan block at infinity, i.e. $k = \max_j \nu_{(\infty,j)}$.

We are interested here in finding the Jordan structure of a real and finite eigenvalue λ_0 of a real regular pencil $\lambda B - A$. An important special application of this is to find the index k of a system of DAEs. In order to reduce that to a real and finite eigenvalue problem, we make a change of variable $\mu = 1/\lambda$, then the index k of (1.1) is the size of the largest Jordan block of the eigenvalue $\mu_0 = 0$ in the Weierstrass form of the pencil $\mu A - B$. We will see that the problem is also simpler to analyze if the coefficient of μ is invertible. This can always be ensured on the diagonal sub-blocks of the Hessenberg triangular form (see Golub & Van Loan, 2013, Section 7.7.5) or by the change of variable $\mu = 1/(\lambda - c)$, which now reduces the problem to that of finding the Jordan structure of the eigenvalue $\mu_0 = 0$ of the pencil $\mu A_c - B$, where the matrix $A_c := A - cB$ is real and invertible. Such a value c is easy to find since, if the original pencil $\lambda B - A$ is regular, then the matrix $A_c := A - cB$ satisfies those conditions for any real value c that is not an eigenvalue of the pencil $\lambda B - A$. In the sequel we will only consider real and finite eigenvalues of a real and regular pencil $\lambda B - A$.

3. Preliminary reduction to Hessenberg triangular form

The standard procedure of the QZ approach to solve the generalized eigenvalue problem of a regular pencil $\lambda B - A$, is to reduce it first to its Hessenberg triangular form. Throughout this section we will assume that the matrix transformed to triangular form is nonsingular and we will point out how to deal with the general case later. The Hessenberg triangular form of a regular pencil $\lambda B - A$ can be obtained using orthogonal transformations U and V such that the transformed pencil $\lambda B_T - A_H := V^T(\lambda B - A)U$ has the form

$$\lambda B_T - A_H := \lambda \begin{bmatrix} b_{1,1} & \cdots & \cdots & b_{1,n} \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ & & & b_{n,n} \end{bmatrix} - \begin{bmatrix} a_{1,1} & \cdots & \cdots & a_{1,n} \\ a_{2,1} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & a_{n,n-1} & a_{n,n} \end{bmatrix}, \tag{3.1}$$

where A_H is Hessenberg and B_T is upper-triangular. If B is nonsingular, the same also holds for B_T and it follows then that

$$H := B_T^{-1} A_H = U^T(B^{-1}A)U = \begin{bmatrix} h_{1,1} & \cdots & \cdots & h_{1,n} \\ h_{2,1} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & h_{n,n-1} & h_{n,n} \end{bmatrix}$$

is in Hessenberg form as well. It is also clear from this relation that the off-diagonal elements $a_{i,i-1}$ are zero if and only if the off-diagonal elements $h_{i,i-1}$ are zero, and hence that A_H is unreduced if and only if H is. The procedure to reduce a regular pencil to Hessenberg triangular form using orthogonal transformations Q and Z is detailed in Golub & Van Loan (2013, Section 7.7.4). It shows that in general B_T may be singular and A_H may not be unreduced, but if so, the pencil is partitioned in block triangular form where each diagonal block is of the above type, where B_T is triangular and invertible and A_H is Hessenberg and unreduced.

4. Deflating a real eigenvalue

Let us assume that we are already given a pencil $\lambda B - A$ in Hessenberg triangular form and that A is unreduced. If not, the operations described below can be applied to each unreduced sub-pencil of a general Hessenberg triangular pencil. We point out that we chose here to use so-called 'backward' QZ steps, which implies that the matrix Q will appear as a right transformation and Z as a left transformation. This is linked to the fact that we will use left eigenvectors in our construction of the QZ step and make the link to the 'implicit Q theorem' (see Remark 4.1).

In exact arithmetic, if λ_0 is a real and finite eigenvalue of the unreduced matrix Hessenberg triangular pencil $\lambda B - A$ and we perform one backward QZ step with shift λ_0, the pencil $\lambda\tilde{B} - \tilde{A} = \lambda Z^T BQ - Z^T AQ$ is still in Hessenberg triangular form with its first column proportional to $(\lambda - \lambda_0)\mathbf{e}_1$, and Q is an unreduced Hessenberg matrix formed by the product of $n-1$ Givens rotations $G_{n-i}^{(r)}$, $i = 1, \ldots, n-1$. Unfortunately, this may not be the case anymore in finite precision because of the phenomenon known as 'blurring' (Watkins, 1996) or because of the ill-conditioning of the eigenvalue λ_0.

Therefore, we need to consider alternative constructions of the QZ step, for which we recall the following theorem. Since we want to relate the rotations used in these different constructions, we will make them unique by choosing the sign of s always positive when $s \neq 0$, and to choose $c = 1$ when $s = 0$. Also, since λ_0 is assumed to be finite, we can represent it as the ratio $\lambda_0 = \alpha_0/\beta_0$ with $\beta_0 \neq 0$ and $\alpha_0^2 + \beta_0^2 = 1$. The results of this theorem are known (see Mastronardi & Van Dooren, 2018 for the equivalent result for the standard eigenvalue problem), but since we rephrase them for the backward *QZ* step, we repeat it here.

THEOREM 4.1 Let $\lambda B - A$ be a real unreduced Hessenberg triangular pencil with real and finite eigenvalue $\lambda_0 = \alpha_0/\beta_0$ with normalization $\alpha_0^2 + \beta_0^2 = 1$, and define the Hessenberg matrix $H := (\alpha_0 B - \beta_0 A)$. Then,

1. the pencil $\lambda B - A$ has a normalized eigenvector $\mathbf{x}$ corresponding to $\lambda_0 = \alpha_0/\beta_0$:

$$(\alpha_0 B - \beta_0 A)\mathbf{x} = H\mathbf{x} = 0, \quad \|\mathbf{x}\|_2 = 1,$$

which is unique up to a scale factor ± 1, and has its last component x_n nonzero; therefore, there is an 'essentially unique' orthogonal transformation $Q = G_1^{(r)}, \dots, G_{n-1}^{(r)}$ that transforms $\mathbf{x}$ to $Q\mathbf{x} = \pm\mathbf{e}_1$.

2. The unreduced Hessenberg matrix $H := (\alpha_0 B - \beta_0 A)$ is transformed to upper triangular form R with $r_{1,1} = 0$ by the 'essentially unique' orthogonal transformation $Q = G_1^{(r)}, \dots, G_{n-1}^{(r)}$, yielding the factorization

$$H = \alpha_0 B - \beta_0 A = RQ. \tag{4.1}$$

3. There are two 'essentially unique' sequences of Givens rotations $G_{n-1}^{(r)}, \dots, G_1^{(r)}$ and $G_{n-1}^{(\ell)}, \dots, G_1^{(\ell)}$ whose products

$$Q := G_1^{(r)} G_2^{(r)} \cdots G_{n-1}^{(r)}, \quad Z := G_1^{(\ell)} G_2^{(\ell)} \cdots G_{n-1}^{(\ell)}, \tag{4.2}$$

are both Hessenberg and transform the triple $(A, B, \mathbf{x})$ to an equivalent one

$$(\tilde{A}, \tilde{B}, \tilde{\mathbf{x}}) := (ZAQ^T, ZBQ^T, Q\mathbf{x}),$$

where

$$\tilde{\mathbf{x}} = \pm\mathbf{e}_1, \quad (\alpha_0 \tilde{B} - \beta_0 \tilde{A})\mathbf{e}_1 = 0, \quad \lambda\tilde{B} - \tilde{A} \text{ is in Hessenberg triangular form.}$$

Proof. To prove item 1, we point out that the normalized eigenvector $\mathbf{x}$ is unique (up to a scaling factor ± 1) because it is the solution of $H\mathbf{x} = 0$, where H has rank $n-1$ since it is unreduced and Hessenberg. For the same reason its last component x_n is nonzero, since otherwise the whole vector $\mathbf{x}$ would be zero. The reduction of $\mathbf{x}$ to $\tilde{\mathbf{x}} = Q\mathbf{x} = \pm\mathbf{e}_1$ then requires a sequence of Givens rotations

$$G_{i-1}^{(r)} \in \mathbb{R}^{n \times n}, \quad i = n, n-1, \dots, 2,$$

in order to eliminate the entries x_i, $i = n, n-1, \dots, 2$ of the vector $\mathbf{x}$. By choosing the sign of s in these Givens rotations positive, we make them unique.

To prove item 2, we use that $\left[h_{n,n-1}, h_{n,n}\right] \begin{bmatrix} x_{n-1} \\ x_n \end{bmatrix} = 0$ follows from $H\mathbf{x} = 0$. The orthogonality of these two nonzero vectors implies then that the Givens rotation $G_{n-1}^{(r)}$ eliminating x_n in the product

$G^{(r)}_{n-1}\mathbf{x}$ is the transpose of the rotation that eliminates $h_{n,n-1}$ in the product $HG^{(r)T}_{n-1}$. We then obtain the expression

$$\left(HG^{(r)T}_{n-1}\right)\left(G^{(r)}_{n-1}\mathbf{x}\right) = \left[\begin{array}{cccc|c} \times & \times & \dots & \times & \times \\ \times & \times & \dots & \times & \times \\ & \ddots & \ddots & \vdots & \vdots \\ & & \times & \times & \times \\ \hline & & & 0 & \hat{\times} \end{array}\right]\left[\begin{array}{c} \times \\ \vdots \\ \times \\ \hat{\times} \\ \hline 0 \end{array}\right] = 0,$$

where the elements $\hat{\times}$ are nonzero. Deflating the last row and column in this expression yields a smaller ‘deflated’ unreduced Hessenberg matrix and a corresponding null vector. We can thus follow the same reasoning by induction, to show that the rotations transforming the vector $Q\mathbf{x}$ to $\pm\mathbf{e}_1$ are the same rotations transforming HQ^T to triangular form:

$$HQ^T = \left[\,\diagdown\!\!\!\urcorner\,\right] = R.$$

This thus shows that the upper Hessenberg transformation Q transforming the eigenvector $\mathbf{x}$ to $Q\mathbf{x} = \pm\mathbf{e}_1$ is essentially the same as the one implementing an explicit QZ-step.

For point 3, we point out that the matrix $\tilde{B} := ZBQ^T$ is constrained to be upper triangular. Therefore, each rotation $G^{(\ell)}_i$ has to annihilate element $(i+1, i)$ in the matrix $G^{(\ell)}_{i-1}\cdots G^{(\ell)}_{n-1}BG^{(r)T}_{n-1}\cdots G^{(r)T}_i$, in order to restore its triangular form. That makes the rotation $G^{(\ell)}_i$ ‘essentially’ unique. Since the product $Z := G^{(\ell)}_1 G^{(\ell)}_2 \cdots G^{(\ell)}_{n-1}$ is upper Hessenberg, the matrix $\tilde{H} := ZHQ^T = ZR$ is also upper Hessenberg. But $\tilde{H} = \alpha_0\tilde{B} - \beta_0\tilde{A}$, which then implies that also $\tilde{A}$ must be upper Hessenberg. Finally, since $\mathbf{x} = \pm Q^T\mathbf{e}_1$, we also have $R\mathbf{e}_1 = r_{1,1}\mathbf{e}_1 = 0$, which implies $\tilde{H}\mathbf{e}_1 = 0$. □

Remark 4.1 The implicit Q theorem for regular pencils is closely related to Theorem 4.1. It implies that the transformations Q and Z can also be determined from the first rotation $G^{(r)}_{n-1}$ that computes

$$\left[h_{n,n-1},\ h_{n,n}\right] G^{(r)T}_{n-1} = \left[\ 0\ \ \times\ \right] \tag{4.3}$$

and from the fact that (ZAQ^T, ZBQ^T) is still Hessenberg triangular. This is known as ‘chasing the bulge’ (Watkins, 2007) in the QZ algorithm.

Theorem 4.1 also says that there are three alternative ways to determine the sequence of right Givens rotations $Q := G^{(r)}_1 G^{(r)}_2 \cdots G^{(r)}_{n-1}$:

1. determine Q from $Q\mathbf{x} = \pm\mathbf{e}_1$,
2. determine Q from $H = RQ$,
3. determine $G^{(r)}_{n-1}$ from (4.3) and the rest of Q and Z from the Hessenberg triangular form of $(\tilde{A}, \tilde{B})$.

The left transformations $Z = G^{(\ell)}_1 G^{(\ell)}_2 \cdots G^{(\ell)}_{n-1}$ are always obtained from restoring the triangular form of $\tilde{B}$.

Although these three different approaches are equivalent under exact arithmetic, their numerical implementations are different. For the standard eigenvalue problem, examples were given in Mastronardi & Van Dooren (2018) that the eigenvector approach is the most reliable method. We will show here that this is also the case for the generalized eigenvalue problem.

5. Defining a perfect shift QZ step

In general, a computer implementation of a numerical algorithm yields only an approximation $\mathbf{v}$ of an eigenvector $\mathbf{x}$, corresponding to a presumed eigenvalue λ_0, which is also only an approximation of a true eigenvalue of the pencil $\lambda B - A$. In order to define what we mean by a QZ step corresponding to a 'perfect shift', we first need to define the arithmetic model. Here we will assume that we are working on a machine using floating point arithmetic with unit roundoff u. We then follow a reasoning developed in Mastronardi & Van Dooren (2018), that was based on the explicit version of the QR-algorithm. In the QZ version of this algorithm we construct the Hessenberg matrix $H := \alpha_0 B - \beta_0 A$ and then proceed as follows:

- find an upper Hessenberg Q such that $H = RQ$ with R upper triangular,
- find an upper Hessenberg Z such that $\tilde{B} := ZBQ^T$ is upper triangular,
- construct $\tilde{A} := ZAQ^T$.

It then follows from the upper Hessenberg form of $\tilde{H} := ZHQ^T$ and the upper triangular form of $\tilde{B}$ that $\tilde{A}$ is upper Hessenberg. We now look at the results for the corresponding computed matrices, when using inexact arithmetic on a machine with unit roundoff u and $\gamma_n := \frac{nu}{1-nu}$, as described in Higham (2002), and we will ignore the effects of gradual underflow.

Theorem 5.1 Let H be the unreduced Hessenberg matrix $H := \alpha_0 B - \beta_0 A$, where $\lambda_0 := \alpha_0/\beta_0$ is an arbitrary shift. Let the sequence of Givens transformations $G_i^{(r)}$ be constructed from the explicit factorization $H = RQ$, and the sequence of Givens transformations $G_i^{(\ell)}$ be constructed from the triangular product $\tilde{B} = Z(BQ^T)$, then in inexact arithmetic, the computed Hessenberg matrix $\tilde{H}$ satisfies

$$Z(H + \Delta_H)Q^T = \tilde{H} + \Delta_{\tilde{H}}, \tag{5.1}$$

where

$$Q := \tilde{G}_1^{(r)} \tilde{G}_2^{(r)} \cdots \tilde{G}_{n-1}^{(r)}, \quad Z := \tilde{G}_1^{(\ell)} \tilde{G}_2^{(\ell)} \cdots \tilde{G}_{n-1}^{(\ell)},$$

are the products of exactly orthogonal Givens rotations contructed to eliminate appropriate elements in the matrix transformations $H = RQ$ and $\tilde{B} = Z(BQ^T)$, and

$$\|\Delta_H\|_F \leq \gamma_{cn} \|H\|_F, \quad \|\Delta_{\tilde{H}}\|_F \leq \gamma_{cn} \|\tilde{H}\|_F.$$

Moreover, the perturbations Δ_H and $\Delta_{\tilde{H}}$ are Hessenberg as well and c is a moderate constant of the order of 1, provided the rotation parameters are computed via the standard construction.

Proof. This theorem is very similar to Theorem 3.1 of Mastronardi & Van Dooren (2018), which is proven in its Appendix A on p. 1613, and is based on the application of Givens transformations to go from Hessenberg form to triangular form and to go from triangular form back to Hessenberg form. The errors incurred during the triangularization $H = RQ$ are mapped to Δ_H and the errors incurred during the left transformation $\tilde{H} = ZR$ are mapped to $\Delta_{\tilde{H}}$. The details for the bounds are given in Appendix A of Mastronardi & Van Dooren (2018). □

Remark 5.1 Note that the backward errors Δ_A and Δ_B are not Hessenberg, but only their linear combination $\Delta_H = (\alpha_0\Delta_B - \alpha_0\Delta_A)$ is. We also point out here that Theorem 5.1 does not apply to the implicit QZ step. For this, one can prove the weaker result that $Z(H + \Delta_H)Q^T = \tilde{H}$ where the backward error Δ_H satisfies $\|\Delta_H\|_F \leq 2\gamma_{cn}\|H\|_F$, but without the constraint that Δ_H is Hessenberg.

Proof. This remark is very similar to Remark 3.1 of Mastronardi & Van Dooren (2018), which is also proven in its Appendix A on p. 1613, and is again based on the application of left and right Givens transformations, but without the Hessenberg constraint. The details for the bounds are given in Appendix A of Mastronardi & Van Dooren (2018). □

We will then say that the QZ step with shift $\lambda_0 = \alpha_0/\beta_0$ is 'perfect' provided $\lambda\tilde{B} - \tilde{A}$ is in Hessenberg triangular form with the first column, satisfying

$$(\tilde{H} + \Delta_{\tilde{H}})\mathbf{e}_1 := \alpha_0(\tilde{B} + \Delta_{\tilde{B}})\mathbf{e}_1 - \beta_0(\tilde{A} + \Delta_{\tilde{A}})\mathbf{e}_1 = 0,$$

where $\Delta_{\tilde{A}}$ and $\Delta_{\tilde{B}}$ satisfy $\Delta_{\tilde{H}} = \alpha_0\Delta_{\tilde{B}} - \beta_0\Delta_{\tilde{A}}$ and are of the order of $\varepsilon_M\|H\|_F$. Equivalently, one would have that

$$\alpha_0\tilde{B}\mathbf{e}_1 - \beta_0\tilde{A}\mathbf{e}_1 \approx 0.$$

This implies that $\mathbf{x} := Q^T\mathbf{e}_1$ is an *exact* null vector of the perturbed Hessenberg matrix $H + \Delta_H$ corresponding to the *exact* 'shift' $\lambda_0 = \alpha_0/\beta_0$ and that $(\lambda_0, \mathbf{x})$ is an exact eigenvalue/eigenvector pair of the slightly perturbed pencil $\lambda(B + \Delta B) - (A + \Delta_A)$. Notice that the use of the forward error $\Delta_{\tilde{H}}$ is needed for this interpretation. Usually, a tolerance τ is specified for the errors Δ_H and $\Delta_{\tilde{H}}$ in (5.1) that is of the order $\varepsilon_M\|H\|_F$ and compatible with the bound of Theorem 5.1 or Remark 5.1, i.e. $\tau \geq \gamma_{cn}\|H\|_F$. In the sequel, we will insist that the backward error Δ_H is Hessenberg, because we will be able to construct such a perturbation.

Definition 5.2 A (backward) QZ step with shift $\lambda_0 = \alpha_0/\beta_0$ is 'perfect' if it corresponds to a perturbed Hessenberg matrix $H + \Delta_H$ with $\|\Delta_H\|_F \leq \tau$ for which the (backward) QZ step satisfies (5.1) *exactly* and for which $(\lambda_0, \mathbf{x})$ is an *exact* eigenvalue/eigenvector pair. Moreover, the property that λ_0 is an exact eigenvalue of the transformed matrix $\tilde{H}$ is made possible by a perturbation $\Delta_{\tilde{H}} = \alpha_0\Delta_{\tilde{B}} - \beta_0\Delta_{\tilde{A}}$ of norm $\|\Delta_{\tilde{H}}\|_F \leq \tau$, by choosing a minimum norm solution for the matrices $\Delta_{\tilde{B}}$ and $\Delta_{\tilde{A}}$.

Remark 5.2 If we can guarantee that the backward errors (Δ_A, Δ_B) and the forward errors $(\Delta_{\tilde{A}}, \Delta_{\tilde{B}})$ have the same structure as the corresponding data pairs (A, B) and $(\tilde{A}, \tilde{B})$, then Δ_H and $\Delta_{\tilde{H}}$ will be Hessenberg. This can be viewed as a form of mixed stability on the transformation $Z(\lambda B - A)Q^T = \lambda\tilde{B} - \tilde{A}$.

Notice that this error analysis does not say if, for a given matrix H, a shift λ_0 will be 'perfect' and show up in the $(1, 1)$ position of the computed matrix $\tilde{H}$, since we do not know what backward errors correspond to it and these can affect the forward errors a lot. One would think that it suffices to have the property

$$\|H\mathbf{x}\|_2 \approx \varepsilon_M\|H\|_2, \tag{5.2}$$

where $\mathbf{x}$ is the presumed eigenvector since it yields a small residual, and where ε_M is the machine epsilon of the computer used. This would imply that $\tilde{H}\mathbf{e}_1$ is of the order of $\varepsilon_M\|H\|_2$ and hence $\mathbf{e}_1$ is an eigenvector of a pencil close to $\lambda\tilde{B} - \tilde{A}$. Simple examples were given in Mastronardi & Van Dooren (2018) to indicate that this is in general not correct for the case of standard eigenvalue problems, so obviously, this is not the case for the generalized eigenvalue problem either. We now look at sufficient conditions for guaranteeing a perfect shift *QZ* step.

6. Sufficient conditions for a perfect shift *QZ* step

We consider $H := \alpha_0 B - \beta_0 A$ where $\lambda_0 := \alpha_0/\beta_0$ is a presumed eigenvalue, and $\alpha_0^2 + \beta_0^2 = 1$. Let us assume that H is nearly singular in the sense that its smallest singular value $\underline{\sigma} := \sigma_{\min}(H)$ is equal to $\varepsilon\|H\|_2$, with ε of the order of the machine accuracy ε_M. This suggests that λ_0 might be a good choice for a perfect shift. Let us then choose as approximate eigenvector the vector $\mathbf{v}$ minimizing the residual

$$\min_{\mathbf{v}} \|H\mathbf{v}\|_2, \quad \|\mathbf{v}\|_2 = 1. \tag{6.1}$$

An optimal solution $\mathbf{v}$ to this problem is given by the right singular vector of H:

$$H\mathbf{v} = \mathbf{u}, \quad \|\mathbf{u}\|_2 = \underline{\sigma}. \tag{6.2}$$

From this, one also finds the minimum norm perturbation $\varDelta = -\mathbf{u}\mathbf{v}^T$ of 2-norm $\underline{\sigma}$ ensuring that $\mathbf{v}$ is a true null vector of $H + \varDelta$:

$$(H + \varDelta)\mathbf{v} = \mathbf{0},$$

but this solution is not Hessenberg in general. The next Lemma, proven in Mastronardi & Van Dooren (2018) gives the minimum norm perturbation while imposing this Hessenberg structure, starting from an arbitrary pair of vectors $(\mathbf{u}, \mathbf{v})$, satisfying

$$\|\mathbf{v}\|_2 = 1, \quad \mathbf{u} = H\mathbf{v}. \tag{6.3}$$

Lemma 6.1 (Mastronardi & Van Dooren (2018)). The minimum Frobenius norm solution $\varDelta_H$ of Hessenberg form

$$\varDelta_H = \begin{bmatrix} \delta h_{1,1} & \delta h_{2,1} & \cdots & \delta h_{1,n} \\ \delta h_{2,1} & \delta h_{2,2} & \cdots & \delta h_{2,n} \\ & \ddots & \cdots & \vdots \\ & & \delta h_{n,n-1} & \delta h_{n,n} \end{bmatrix}$$

to the system

$$(H + \varDelta_H)\mathbf{v} = \mathbf{0}, \quad H\mathbf{v} = \mathbf{u}, \tag{6.4}$$

where $\|\mathbf{v}\|_2 = 1$ and $v_n \neq 0$ has Frobenius norm equal to

$$\|\varDelta_H\|_F = \|u_1/v_1, u_2/v_2, \ldots, u_n/v_n\|_2, \tag{6.5}$$

where

$$\nu_1 = 1, \quad \nu_i = \|[\nu_{i-1}, \nu_i, \ldots, \nu_{n-1}, \nu_n]\|_2, \; i = 2, \ldots, n.$$

Remark 6.1 It is easy to see that the successive vector norms ν_i satisfy the inequalities

$$\nu_n \leq \ldots \leq \nu_3 \leq \nu_2 = \nu_1 = 1,$$

where $\nu_n = \| \left[\; \nu_{n-1} \quad \nu_n \; \right] \|_2$. Therefore, the Frobenius norm for Δ_H is bounded by

$$\|\Delta_H\|_F \leq \frac{\|\mathbf{u}\|_2}{\nu_n}$$

and hence the Hessenberg perturbation Δ_H is then of the same order as the unstructured perturbation Δ if $\nu_n \approx 1$.

A way to guarantee a bound for Δ_H that is of the same order as the unstructured error Δ, is to compute the approximate null vector $\mathbf{x}$ in such a way that the residual vector (which we now denote by $\mathbf{r} = H\mathbf{x}$) satisfies stricter conditions. This is shown in the next theorem, also proven in Mastronardi & Van Dooren (2018).

Theorem 6.1 Let H be an unreduced Hessenberg matrix and let us have an estimate of a null vector $\mathbf{x}$, satisfying

$$\mathbf{r} := H\mathbf{x}, \quad \|\mathbf{x}\|_2 = 1,$$

where

$$\mathbf{x}_i^T := [x_{i-1}, x_i, \ldots, x_{n-1}, x_n], \quad \nu_i := \|\mathbf{x}_i\|_2, i = 2, ..., n,$$
$$\nu_1 = 1, \quad \hat{\varepsilon}_i := r_i/\nu_i \quad \text{and} \quad \|[\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n]\|_2 \leq \varepsilon_M \|H\|_F \tag{6.6}$$

and let the Givens rotations G_i be computed to annihilate element x_{i+1} for $i = n-1, \ldots, 1$ of the approximate eigenvector $\mathbf{x}$ and transforming it to $\mathbf{e}_1$. Then the product

$$Q := \tilde{G}_1 \tilde{G}_2 \cdots \tilde{G}_{n-1},$$

where each $\tilde{G}_i$ is the exactly orthogonal Givens rotation corresponding to G_i, yields a 'perfect' triangularization of the Hessenberg matrix H in the sense that there exist backward perturbations Δ_H and $\Delta_{\mathbf{x}}$ such that

$$(H + \Delta_H)Q^T = R, \quad \mathbf{x} + \Delta_{\mathbf{x}} = \mathbf{e}_1, \quad \|\Delta_H\|_F \leq c\varepsilon_M \|H\|_F, \quad \|\Delta_{\mathbf{x}}\|_2 \leq c\varepsilon_M,$$

where R is upper triangular and c is a constant of the order of 1.

Proof. The proof is essentially identical to that of Lemma 4.1 in Mastronardi & Van Dooren (2018). □

It then also follows that the matrix $\tilde{H} + \Delta_{\tilde{H}} := Z(H + \Delta_H)Q^T$ is Hessenberg, since Z is Hessenberg, and this implies that the QZ step $(\tilde{A}, \tilde{B}, \mathbf{x}) := (ZAQ^T, ZBQ^T, Q\mathbf{x})$ is a perfect shift QZ step with shift

$\lambda_0 := \alpha_0/\beta_0$. We summarize this so-called *eigenvector* method below by giving a pseudo-code.

```
function $[A, B, \mathbf{x}] =$ eigenvector_method$(A, B, \alpha, \beta, \mathbf{x}, n)$;
  $H := \alpha B - \beta A$;
  for $= i = n - 1 : -1 : 1$,
        $G_i^{(r)} =$ givens$(x_i, x_{i+1})$;
        $\mathbf{x}_{i:i+1} = G_i^{(r)} \mathbf{x}_{i:i+1}$;
        $H_{:,i:i+1} = H_{:,i:i+1} {G_i^{(r)}}^T; A_{:,i:i+1} = A_{:,i:i+1} {G_i^{(r)}}^T; B_{:,i:i+1} = B_{:,i:i+1} {G_i^{(r)}}^T$;
        $G_i^{(\ell)} =$ givens$(B_{i,i}, B_{i+1,i})$;
        $H_{i:i+1,:} = G_i^{(\ell)} H_{i:i+1,:}; A_{i:i+1,:} = G_i^{(\ell)} A_{i:i+1,:}; B_{i:i+1,:} = G_i^{(\ell)} B_{i:i+1,:}$;
  end;
```

The key point in this Theorem is of course that we need an approximate null vector $\mathbf{x}$ with a sufficiently small residual, especially in the components where each trailing sub-vector $\mathbf{x}_i$ has small norm ν_i. We explain in the next subsection how to compute such an approximation.

7. Scaling the eigenvector

In this section we show how to compute an approximate null vector $\mathbf{x}$ of an unreduced Hessenberg matrix such that its residual $\mathbf{r}$ satisfies the conditions (6.6) requested by Theorem 6.1.

The basic idea here is to apply a diagonal scaling (with $d \geq 1$)

$$D := \mathrm{diag}(1, d, d^2, \ldots, d^{n-1}) \tag{7.1}$$

that 'balances' the entries of $\mathbf{x}$ without affecting too much the norm of H. The theorem is proven in Mastronardi & Van Dooren (2018).

THEOREM 7.1 (Mastronardi & Van Dooren (2018)). Let H be an unreduced Hessenberg matrix and let $(\lambda_0, \mathbf{x})$ be an approximate eigenvalue/eigenvector pair. Then there always exists a scaling D of the form (7.1) such that the transformed pair $(H_D, \mathbf{x}_D) := (DHD^{-1}, D\mathbf{x})$ satisfies the constraints

$$\|H_D\|_F \leq d\|H\|_F, \quad d := \max(\min[\max_{i \leq n-2}\{|x_i/x_{n-1}|^{1/n-i-1}\}, \max_{i \leq n-2}\{|x_i/x_n|^{1/n-i}\}], 1)$$

and such that the largest component of $\mathbf{x}_D$ is one of its last two components.

Using this scaling technique it was shown in Mastronardi & Van Dooren (2018) that the residual $\mathbf{r}$ typically has the required scaling in order to guarantee the bounds of Theorem 6.1.

LEMMA 7.1 (Mastronardi & Van Dooren (2018)). Let $\mathbf{x}_D$ be an approximate normalized null vector of H_D and let the residual $\mathbf{r}_D := H_D \mathbf{x}_D$ be computed with accuracy $\|\mathbf{r}_D\|_2 \leq \varepsilon \|H_D\|_2$, then:

$$r_i^{(D)} \leq \|\mathbf{r}_D\|_2 \leq \varepsilon \sqrt{n} \|H_D\|_2 \nu_i^{(D)}, \tag{7.2}$$

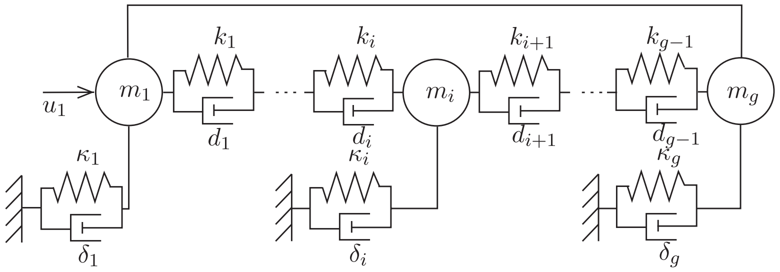

FIG. 1. Damped spring-mass model borrowed from Mehrmann, V. & Stykel, T. (2005).

and the rescaled residual $\mathbf{r} := D^{-1}\mathbf{r}_D$ satisfies the bound

$$r_i \leq \varepsilon\sqrt{n}\|H_D\|_2 \nu_i^{(D)}/d^{i-1}. \tag{7.3}$$

If we assume $d \geq 2$ and define c_ν to satisfy the bound

$$\nu_i^{(D)}/d^{i-2} \leq c_\nu \nu_i, \tag{7.4}$$

we obtain the simplified inequality

$$r_i \leq (4c_\nu\sqrt{n}/3)\varepsilon\|H\|_2\nu_i \tag{7.5}$$

in terms of the rescaled vector $\mathbf{x} := D^{-1}\mathbf{x}_D$ and its sub-norms ν_i.

REMARK 7.1 It was also pointed out in Mastronardi & Van Dooren (2018) that one often has the stronger bound $r_i \leq \varepsilon\|H\|_2\nu_i$, when $\mathbf{x}_D$ and $\mathbf{r}_D$ are both ‘balanced’ in the sense that their entries are of comparable sizes. It was also pointed out there that it is good practice to choose a scaling d that is a power of 2 in order to avoid rounding errors in the scaling of the matrix or the computation of the rescaled vector $\mathbf{x}_D$.

8. Numerical example

The example is taken from Mehrmann & Stykel (2005) and is a damped spring-mass system shown in Fig. 1 with state space model $\lambda B - A$ of dimension $2g + 1$. The masses $m_i, i = 1, \ldots, g$ are all set to 100, and the other parameters k_i, κ_i, d_i and δ_i are ranging from 2 to 10. We chose $g = 10$, which results in a pencil of dimension 21 and with the sparsity pattern indicated in Fig. 2.

This $\lambda B - A$ is known to have a triple eigenvalue at $\lambda = \infty$ that belongs to a single Jordan block. The index of this infinite eigenvalue is therefore 3 and will require three successive deflations. We proceeded as follows for this example. We first reduced the pencil to Hessenberg triangular form

$$Z^T A Q = T, \quad Z^T B Q = H,$$

where T is triangular and H is Hessenberg. In our example T is invertible and H is unreduced (but singular). A first null vector $\mathbf{x}$ was computed for $H\mathbf{x} = 0$, but the corresponding residual vector

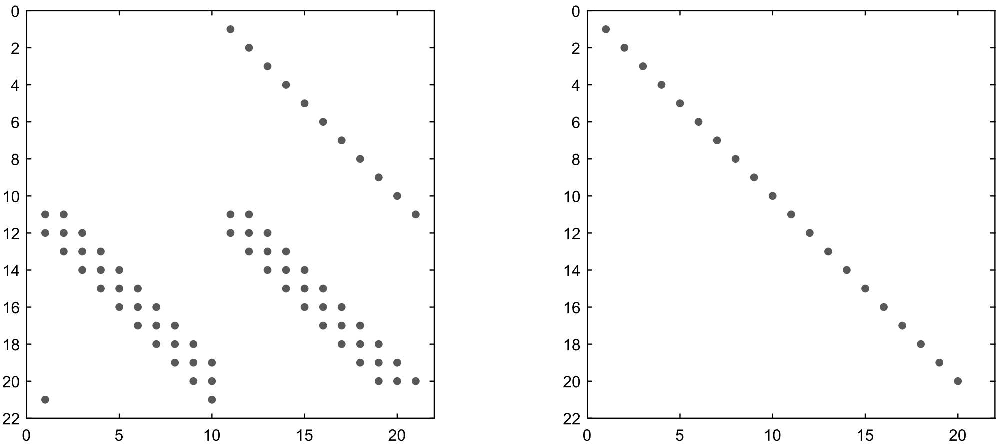

FIG. 2. Sparsity patterns of the matrices A and B of the damped spring-mass model.

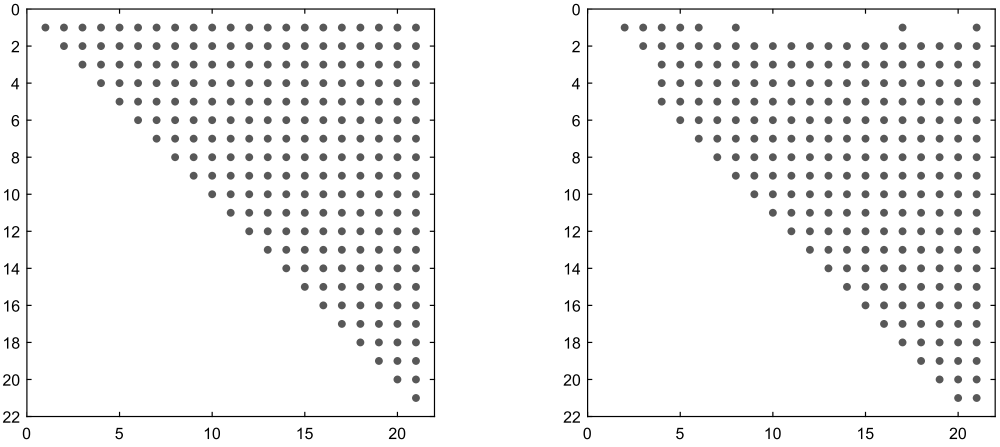

FIG. 3. Sparsity patterns of the matrices $\tilde{A}$ and $\tilde{B}$ after the orthogonal deflation.

$\hat{\varepsilon} := [\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n]$ of Theorem 6.1 had 2-norm of the order of $8.10^{-13}\|H\|_2$, which does not guarantee that the so-called perfect shift will be executed correctly. The scaling procedure was then applied and yielded a modified null vector $D^{-1}\mathbf{x}_D$, but now with a residual vector $\hat{\varepsilon}$ of the order of $4.10^{-16}\|H\|_2$, which guaranteed a perfect deflation of the first infinite eigenvalue. This was repeated two more times on the deflated pencils to finally yield the transformed pencil $\lambda\tilde{B} - \tilde{A}$ with the pattern of nonzeros indicated in Fig. 3. In all three deflations, the perfect shift could be executed successfully after making use of the scaling procedure. The residual vectors before and after scaling were each of the order of $10^{-13}\|H\|_2$ and $10^{-16}\|H\|_2$, respectively, indicating that the scaling procedure worked correctly. The so-called off-norms of matrices $\tilde{A}$ and $\tilde{B}$ (i.e. the norms of the lower triangular and lower Hessenberg parts that are dismissed in the Hessenberg triangular form) were respectively of the order of $10^{-15}\|H\|_2$ and $10^{-15}\|T\|_2$, indicating that the method was numerically stable.

9. Concluding remarks

In this paper we revisited the implementation of QZ steps with so-called perfect shifts. We restricted ourselves to the case of real shifts, but the analysis for complex shifts is quite comparable. The problem of double implicit QZ steps, on the other hand, is more complicated, as can be seen from the analysis of the double QR steps performed in Mastronardi & Van Dooren (2018) for the standard eigenvalue problem. We also gave an example of the calculation of the index of infinite eigenvalues when the matrix B is singular. This is an important application since DAE solvers suffer from serious error propagation when the index is not correctly computed. The use of the present method to retrieve the full Jordan form characteristic of a given eigenvalue is another possible application and can probably be solved using techniques similar to those of Mastronardi & Van Dooren (2017) developed for the standard eigenvalue problem.

Funding

Gruppo Nazionale Calcolo Scientifico-Istituto Nazionale di Alta Matematica (to N.M.); Consiglio Nazionale delle Ricerche Italy (Short Term Mobility Program to P.V.D.).

References

Gantmacher, F. R. (1959) *The Theory of Matrices*. New York: Chelsea.

Golub, G. H. & Van Loan, C. F. (2013) *Matrix Computations*, 4th edn. Baltimore: Johns Hopkins University Press.

Higham, N. J. (2002) *Accuracy and Stability of Numerical Algorithms*, 2nd edn. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Kunkel, P. & Mehrmann, V. (2006) *Differential-Algebraic Equations: Analysis and Numerical Solution*. Zürich, Switzerland: European Mathematical Society.

Mastronardi, N. & Van Dooren, P. (2017) Computing the Jordan structure of an eigenvalue. *SIAM J. Matrix Anal. Appl.*, **38**, 949–966.

Mastronardi, N. & Van Dooren, P. (2018) The *QR*-steps with perfect shifts. *SIAM J. Matrix Anal. Appl.*, **39**, 1591–1615.

Mehrmann, V. (1991) *The Autonomous Linear Quadratic Control Problem*. Berlin: Springer.

Mehrmann, V. & Stykel, T. (2005) Balanced truncation model reduction for large-scale systems in descriptor form *Dimension Reduction of Large-Scale Systems* (P. Benner, D. C. Sorensen & V. Mehrmann eds). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 83–115.

Watkins, D. S. (1996) The transmission of shifts and shift blurring in the QR algorithm. *Linear Algebra Appl.*, **241–243**, 877–896.

Watkins, D. S. (2007) *The Matrix Eigenvalue Problem*. Philadelphia: SIAM.