ELSEVIER Linear Algebra and its Applications 309 (2000) 339–361

LINEAR ALGEBRA AND ITS APPLICATIONS

Backward error and condition of polynomial eigenvalue problems

Françoise Tisseur [*]

Department of Mathematics, University of Manchester, Manchester M13 9PL, UK

Received 18 September 1998; accepted 13 March 1999

Submitted by J.L. Barlow

Abstract

We develop normwise backward errors and condition numbers for the polynomial eigenvalue problem. The standard way of dealing with this problem is to reformulate it as a generalized eigenvalue problem (GEP). For the special case of the quadratic eigenvalue problem (QEP), we show that solving the QEP by applying the QZ algorithm to a corresponding GEP can be backward unstable. The QEP can be reformulated as a GEP in many ways. We investigate the sensitivity of a given eigenvalue to perturbations in each of the GEP formulations and identify which formulations are to be preferred for large and small eigenvalues, respectively. © 2000 Elsevier Science Inc. All rights reserved.

Keywords: Polynomial eigenvalue problem; Quadratic eigenvalue problem; Generalized eigenvalue problem; Backward error; Condition number

1. Introduction

We are concerned with backward error analysis and conditioning for the nonlinear eigenvalue problem

$$P(\lambda)x = 0, \tag{1.1}$$

where $P(\lambda)$ is a matrix whose elements are polynomials in a scalar λ. We write P in the form

[*] E-mail: ftisseur@ma.man.ac.uk

PII: S 0 0 2 4 - 3 7 9 5 (9 9) 0 0 0 6 3 - 4

$$P(\lambda) = \lambda^m A_m + \lambda^{m-1} A_{m-1} + \cdots + A_0,$$

where $A_l \in \mathbb{C}^{n\times n}$, $l = 0 : m$ and we refer to P as a λ-matrix. If $x \neq 0$ then λ is an eigenvalue and x the corresponding right eigenvector; $y \neq 0$ is a left eigenvector if

$$y^* P(\lambda) = 0. \tag{1.2}$$

The importance of backward errors for investigating the stability and quality of numerical algorithms and condition numbers for characterizing the sensitivity of solutions to problems is widely appreciated. The forward error, condition number and backward error are related by the inequality (correct to first order in the backward error)

$$\text{forward error} \leqslant \text{condition number} \times \text{backward error}. \tag{1.3}$$

Perturbation and backward error theory is well developed for linear systems, least squares problems, the standard eigenvalue problem and more recently the generalized eigenvalue problem (GEP). Condition estimation algorithms are now used in most of the major mathematical program libraries including LAPACK, NAG and IMSL as well as much commercial scientific software. However, practically oriented analysis of backward errors and condition numbers of the polynomial eigenvalue problem for $m \geqslant 2$ has not been done.

Few direct numerical methods are available for solving the polynomial eigenvalue problem (PEP). When m is small, the common practice is to transform the PEP (1.1) into a GEP

$$\mathscr{A}\xi = \lambda \mathscr{B}\xi \tag{1.4}$$

of order mn, where $\mathscr{A}$ and $\mathscr{B}$ can be defined by

$$\mathscr{A} = \begin{pmatrix} 0 & I & 0 & \cdots & 0 \\ 0 & 0 & I & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ & & & & I \\ -A_0 & -A_1 & -A_2 & \cdots & -A_{m-1} \end{pmatrix}, \tag{1.5}$$

$$\mathscr{B} = \begin{pmatrix} I & & & & \\ & I & & & \\ & & \ddots & & \\ & & & I & \\ & & & & A_m \end{pmatrix}.$$

Then the QZ algorithm is used if all the eigenpairs are desired, or an Arnoldi or nonsymmetric Lanczos-type method if only a few of them are required.

This work has three main contributions. First, for the PEP, we give computable expressions for backward errors and condition numbers of simple eigenvalues by extending the work in [6,11] concerning the GEP.

There are many ways to reformulate the PEP as a GEP. Our second contribution is to study the stability of these transformations in the case $m = 2$. We show that solving the QEP by applying the QZ algorithm to the GEP can be backward unstable for the QEP. Finally, we investigate the sensitivity of a given eigenvalue of the QEP to perturbations in some GEP formulations. We show that there can be great variation in sensitivity and we identify which formulations are preferred for the large and the small eigenvalues, respectively.

2. Backward error and condition numbers

2.1. Preliminaries

When A_m is nonsingular, $P(\lambda)$ is said to be *regular* and has mn finite eigenvalues. When $\operatorname{rank}(A_m) < n$, $P(\lambda)$ may have infinite eigenvalues. In this paper, we make no assumption on $P(\lambda)$ for the backward error analysis but for the definition and derivation of condition numbers we restrict our attention to regular λ-matrices whose eigenvalues are simple. For a good survey of λ-matrices we refer to Lancaster [16].

Throughout the paper, the matrices E_l, $l = 0 : m$ are arbitrary and represent tolerances against which the perturbations ΔA_l to A_l will be measured. For notational convenience, we define

$$\Delta P(\lambda) = \lambda^m \Delta A_m + \lambda^{r-1} \Delta A_{m-1} + \cdots + \Delta A_0, \tag{2.1}$$

and, for a complex λ,

$$\operatorname{sign}(\lambda) = \begin{cases} \bar{\lambda}/\lambda, & \lambda \neq 0, \\ 0, & \lambda = 0. \end{cases}$$

We use the 2-norm, defined by $\|x\|_2 = (x^* x)^{1/2}$, $\|A\|_2 = \max\left\{\|Ax\|_2 : \|x\|_2 = 1\right\}$.

2.2. Backward errors

A natural definition of the normwise backward error of an approximate eigenpair $(\tilde{x}, \tilde{\lambda})$ of (1.1) is

$$\eta(\tilde{x},\tilde{\lambda}) := \min\{\epsilon : (P(\tilde{\lambda}) + \Delta P(\tilde{\lambda}))\tilde{x} = 0,\ \|\Delta A_l\|_2 \leqslant \epsilon \|E_l\|_2,\ l = 0 : m\}. \tag{2.2}$$

Our first result gives an explicit expression for $\eta(\tilde{x},\tilde{\lambda})$ and makes precise the intuitive feeling that if the residual $r = P(\tilde{\lambda})\tilde{x}$ is small, then we have a "good" approximate eigenpair. It is a straightforward modification of a result of Rigal and Gaches on the normwise backward error for a linear system [20] and a generalization of the backward errors given in [6, Lemma 2.1] and [11, Theorem 2.1].

Theorem 1. *The normwise backward error* $\eta(\tilde{x},\tilde{\lambda})$ *is given by*

$$\eta(\tilde{x},\tilde{\lambda}) = \frac{\|r\|_2}{\tilde{\alpha}\|\tilde{x}\|_2}, \tag{2.3}$$

where $r = P(\tilde{\lambda})\tilde{x}$ *and* $\tilde{\alpha} = \sum_{l=0}^{m} |\tilde{\lambda}|^l \|E_l\|_2$.

Proof. It is straightforward to show that the right-hand side of (2.3) is a lower bound for $\eta(\tilde{x},\tilde{\lambda})$. This lower bound is attained for the perturbations

$$\Delta A_l = -\frac{1}{\tilde{\alpha}}\operatorname{sign}(\tilde{\lambda}^l)\|E_l\|_2 r\tilde{x}^*/\|\tilde{x}\|_2^2, \quad l = 0 : m. \qquad \square$$

When all the A_l are Hermitian, it is of interest to consider a backward error in which the perturbations ΔA_l respect the Hermitian structure in the A_l. Therefore, we define the backward error

$$\begin{aligned}\eta_H(\tilde{x},\tilde{\lambda}) := \min\{\epsilon : {}& P(\tilde{\lambda})\tilde{x} + \Delta P(\tilde{\lambda})\tilde{x} = 0,\\ & \Delta A_l = \Delta A_l^*,\ \|\Delta A_l\|_2 \leqslant \epsilon\|E_l\|_2,\ l = 0 : m\}.\end{aligned} \tag{2.4}$$

It is clear that $\eta_H(\tilde{x},\tilde{\lambda}) \geqslant \eta(\tilde{x},\tilde{\lambda})$ and that the optimal perturbations in (2.2) are not Hermitian in general. The next theorem shows that requiring the perturbations to respect the Hermitian structure in the A_l has no effect on the backward error, provided that $\tilde{\lambda}$ is real.

Theorem 2. *If the matrices* A_l, $l = 0 : m$ *are Hermitian and* $\tilde{\lambda}$ *is real then*

$$\eta_H(\tilde{x},\tilde{\lambda}) = \eta(\tilde{x},\tilde{\lambda}).$$

Proof. Let $r = P(\tilde{\lambda})\tilde{x}$ be the residual of the pair $(\tilde{x},\tilde{\lambda})$. We first find a Hermitian matrix S that satisfies the first constraint in (2.4), $S\tilde{x} = -r$. We take $S = (\|r\|_2/\|\tilde{x}\|_2)I$ if r is a negative multiple of $\tilde{x}$; otherwise we take $S = (\|r\|_2/\|\tilde{x}\|_2)H$, where H is a suitably chosen Householder matrix. Such an H exists if $\tilde{x}^*S\tilde{x} = -\tilde{x}^*r$ is real. But $\tilde{x}^*r = \tilde{x}^*P(\tilde{\lambda})\tilde{x}$, which is real since $\tilde{\lambda}$ is real.

Let ΔA_l be Hermitian matrices defined by

$$\Delta A_l = \frac{1}{\tilde{\alpha}} \operatorname{sign}(\tilde{\lambda}^l) \|E_l\|_2, \quad l = 0 : m, \tag{2.5}$$

where $\tilde{\alpha} = \sum_{l=0}^{m} |\tilde{\lambda}|^l \|E_l\|_2$, so that $\Delta P(\tilde{\lambda}) = S$. Using (2.3), we get

$$\|S\|_2 = \|r\|_2 / \|\tilde{x}\|_2 = \eta(\tilde{x}, \tilde{\lambda})\tilde{\alpha}$$

and then from (2.5) we deduce $\eta_H(\tilde{x}, \tilde{\lambda}) \leqslant \eta(\tilde{x}, \tilde{\lambda})$; since $\eta_H(\tilde{x}, \tilde{\lambda}) \geqslant \eta(\tilde{x}, \tilde{\lambda})$, equality must hold. □

In general, even if the A_l are Hermitian, the eigenvalues are complex. However, for the case $m = 1$, if the pencil $(\mathscr{A}, \mathscr{B})$ is definite, that is, if it satisfies

$$\min \left\{ \left((z^* \mathscr{A} z)^2 + (z^* \mathscr{B} z)^2 \right)^{1/2} : z \in \mathbb{C}^n, \ \|z\|_2 = 1 \right\} > 0, \tag{2.6}$$

then its eigenvalues are real. For a proof see [24, Chapter 6]. An analogous result exists for the case $m = 2$, that is, for the QEP. Let (λ, x) be an eigenpair for $(\lambda^2 A + \lambda B + C)x = 0$ with Hermitian A, B and C that satisfy

$$(z^* B z)^2 - 4(z^* A z)(z^* C z) > 0 \quad \text{for all } z \in \mathbb{C}^n. \tag{2.7}$$

Then λ is a root of

$$\lambda^2 x^* A x + \lambda x^* B x + x^* C x = 0$$

and so is real. Inequality (2.7) is usually called the *overdamping condition* as it corresponds to an overdamped physical system [5; 16, Chapter 7].

When eigenvectors are not computed, a more appropriate measure of the backward error for an approximate eigenvalue may be

$$\eta(\tilde{\lambda}) := \min_{\tilde{x} \neq 0} \eta(\tilde{x}, \tilde{\lambda}). \tag{2.8}$$

Lemma 3. *If $\tilde{\lambda}$ is not an eigenvalue of $P(\lambda)$ then*

$$\eta(\tilde{\lambda}) = \frac{1}{\tilde{\alpha} \|[P(\tilde{\lambda})]^{-1}\|_2}, \tag{2.9}$$

where $\tilde{\alpha} = \sum_{l=0}^{m} |\tilde{\lambda}|^l \|E_l\|_2$.

Proof. The result follows from Theorem 1 on using the equality for a nonsingular matrix $S \in \mathbb{C}^{n \times n}$, $\min_{x \neq 0} \|Sx\|_2 / \|x\|_2 = 1/\|S^{-1}\|_2$. □

In a similar way, we can define the backward error for an approximate eigenvector by

$$\eta(\tilde{x}) := \min_{\tilde{\lambda}} \eta(\tilde{x}, \tilde{\lambda}). \tag{2.10}$$

In general, this minimization problem is unsolved. For $m = 1$ with $A_1 = I$ (the standard eigenvalue problem) and $E_1 = 0$, $\eta(\tilde{x}) = \tilde{x}^* A\tilde{x}/\tilde{x}^*\tilde{x}$. For the generalized eigenvalue problem, Higham and Higham [11] obtained an upper bound by maximizing the numerator.

We define the backward error of a triple $(\tilde{x}, \tilde{y}, \tilde{\lambda})$, where $\tilde{y}$ is an approximate left eigenvector, by

$$\begin{aligned}\eta(\tilde{x}, \tilde{y}, \tilde{\lambda}) := \min\{\epsilon : P(\tilde{\lambda})\tilde{x} + \Delta P(\tilde{\lambda})\tilde{x} = 0,\ \tilde{y}^* P(\tilde{\lambda}) + \tilde{y}^* \Delta P(\tilde{\lambda}) = 0,\\ \|\Delta A_l\|_2 \leqslant \epsilon \|E_l\|_2\ l = 0 : m\}.\end{aligned} \tag{2.11}$$

Theorem 4. *We have*

$$\eta(\tilde{x}, \tilde{y}, \tilde{\lambda}) = \frac{1}{\tilde{\alpha}} \max\left\{ \frac{\|r\|_2}{\|\tilde{x}\|_2}, \frac{\|s\|_2}{\|\tilde{y}\|_2} \right\}, \tag{2.12}$$

where $r = P(\tilde{\lambda})\tilde{x}$, $s^* = \tilde{y}^* P(\tilde{\lambda})$ *and* $\tilde{\alpha} = \sum_{l=0}^{m} |\tilde{\lambda}|^l \|E_l\|_2$.

Proof. By taking the 2-norms of r and s in the equation $r = -\Delta P(\tilde{\lambda})\tilde{x}$ and $s^* = -y^* \Delta P(\tilde{\lambda})$, we find that

$$\eta(\tilde{x}, \tilde{y}, \tilde{\lambda}) \geqslant \frac{1}{\tilde{\alpha}} \max\left\{ \frac{\|r\|_2}{\|\tilde{x}\|_2}, \frac{\|s\|_2}{\|\tilde{y}\|_2} \right\}.$$

To show that this bound is attained, we use a result of Kahan et al. [13, Theorem 2] that states that

$$\min\left\{ \|H\|_2 : H\tilde{x} = r,\ \tilde{y}^* H = s^* \right\} = \max\left\{ \frac{\|r\|_2}{\|\tilde{x}\|_2}, \frac{\|s\|_2}{\|\tilde{y}\|_2} \right\}.$$

Let $H_{\min}$ be a matrix that achieves this minimum and define

$$\Delta A_l = \frac{-\operatorname{sign}(\tilde{\lambda}^l)\|E_l\|_2}{\tilde{\alpha}} H_{\min}, \quad l = 0 : m. \tag{2.13}$$

Then $\Delta P = -H_{\min}$, showing that the ΔA_l are feasible perturbations, and

$$\|\Delta A_l\|_2 = \frac{\|E_l\|_2}{\tilde{\alpha}} \max\left\{ \frac{\|r\|_2}{\|\tilde{x}\|_2}, \frac{\|s\|_2}{\|\tilde{y}\|_2} \right\},$$

so that the lower bound for $\eta(\tilde{x}, \tilde{y}, \tilde{\lambda})$ is attained. □

Note that Theorem 4 shows that $\eta(\tilde{x},\tilde{y},\tilde{\lambda}) = \max(\eta(\tilde{x},\tilde{\lambda}),\eta(\tilde{y},\tilde{\lambda}))$, that is, the backward error of the triple is the maximum of the backward errors of the left and right eigenvectors.

2.3. Condition number

Let λ be a nonzero simple eigenvalue of a regular PEP with corresponding right eigenvector x and left eigenvector y. A normwise condition number of λ can be defined by

$$\kappa(\lambda,P) = \limsup_{\epsilon\to 0}\left\{ \frac{|\Delta\lambda|}{\epsilon|\lambda|} : (P(\lambda+\Delta\lambda)+\Delta P(\lambda+\Delta\lambda))(x+\Delta x)= 0, \right. \\ \left. \|\Delta A_l\|_2 \leqslant \epsilon\|E_l\|_2 \ l=0:m \right\}. \tag{2.14}$$

Theorem 5. *The normwise condition number $\kappa(\lambda,P)$ is given by*

$$\kappa(\lambda,P) = \frac{\alpha\|y\|_2\|x\|_2}{|\lambda||y^*P'(\lambda)x|}, \tag{2.15}$$

where $\alpha = \sum_{l=0}^{m} |\lambda|^l \, \|E_l\|_2$.

Proof. By expanding the first constraint in (2.14) and keeping only the first order terms, we get

$$\Delta\lambda P'(\lambda)x + P(\lambda)\Delta x + \Delta P(\lambda)x = \mathrm{O}(\epsilon^2).$$

Premultiplying by y^* leads to

$$\Delta\lambda y^*P'(\lambda)x + y^*\Delta P(\lambda)x = \mathrm{O}(\epsilon^2).$$

Since λ is a simple eigenvalue, $y^*P'(\lambda)x \neq 0$ [1, Theorem 3.2]. Thus

$$\Delta\lambda = -\frac{y^*\Delta P(\lambda)x}{y^*P'(\lambda)x} + \mathrm{O}(\epsilon^2)$$

and so

$$\frac{|\Delta\lambda|}{\epsilon|\lambda|} \leqslant \frac{\alpha\|y\|_2\|x\|_2}{|\lambda||y^*P'(\lambda)x|} + \mathrm{O}(\epsilon).$$

Hence the expression in (2.15) is an upper bound for the condition number. To show that this bound can be attained we consider the matrix $H = yx^*/(\|y\|_2\|x\|_2)$, for which

$$\|H\|_2 = 1, \quad y^*Hx = \|x\|_2\|y\|_2.$$

Let

$$\Delta A_l = -\text{sign}(\lambda^l)\epsilon\|E_l\|_2 H, \quad l = 0 : m.$$

Then all the norm inequalities in (2.14) are satisfied as equalities and

$$|y^* \Delta P(\lambda)x| = \epsilon\alpha\|y\|_2\|x\|_2.$$

Dividing by $\epsilon|\lambda||y^*P'(\lambda)x|$ and taking the limit as $\epsilon \to 0$ gives the desired equality. □

As for the backward error, if the A_l are Hermitian, it is natural to restrict the perturbations ΔA_l in (2.14) to be Hermitian.

Lemma 6. *Let A_l, $l = 0 : m$ be Hermitian matrices and λ be a real eigenvalue. Let $\kappa^H(\lambda, P)$ denote the condition number defined as in* (2.14) *but with the additional requirement that the ΔA_l are Hermitian. Then $\kappa^H(\lambda, P) = \kappa(\lambda, P)$.*

Proof. We can take $y = x$, so in the proof of Theorem 5, $H = \|x\|_2^{-2}xx^*$ which is Hermitian. It follows that the perturbations for which the bound is attained are also Hermitian. □

2.4. Comments

For our analysis we have used the 2-norm. However, our results can easily be extended to the mixed subordinate matrix norm $\|A\|_{\alpha,\beta}$ on $\mathbb{C}^{n\times n}$ defined by

$$\|A\|_{\alpha,\beta} = \max_{x\neq 0} \frac{\|Ax\|_\beta}{\|x\|_\alpha},$$

as used in [11]. Furthermore, it is straightforward to derive componentwise backward errors and condition numbers, as in [11], but we do not consider them here.

As the PEP can be reformulated as a highly structured GEP, we could use the GEP backward error and condition number that respect linear structure of the matrices as developed by Higham and Higham [11]. However, the results obtained using this approach are harder to interpret and more difficult to compute than the ones we have presented.

3. The quadratic eigenproblem

In this section, we consider the case $m = 2$, corresponding to the important quadratic eigenvalue problem (QEP)

$$Q(\lambda)x = (\lambda^2 A + \lambda B + C)x = 0. \tag{3.1}$$

This problem arises in many applications, including the finite element analysis of automobile brakes [14], earthquake engineering [4] and the analysis of conservative and non-conservative structural systems [23,28]. The QEP also arises when solving least squares problems with quadratic constraints [7].

3.1. From quadratic to generalized forms

Few algorithms work directly on the QEP; for small dense problems most of the ones that do are based on Newton iterations [15,19]. For a good review of such methods, we refer to Ruhe [21]. More recently, Guillaume [10] developed a new method based on the derivative of the function $x(\lambda) = Q(\lambda)^{-1}b$ where b is a given vector. For large sparse problems, Jacobi–Davidson techniques have been investigated [22].

The usual way of dealing with the QEP (3.1) is to transform it into a GEP of twice the order. There are several possible ways to carry out such a transformation. The most commonly used transformation is to companion form, given by

$$\text{GEP}_1: \quad \begin{pmatrix} -B & -C \\ I & 0 \end{pmatrix}\xi = \lambda\begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}\xi, \tag{3.2}$$

with

$$\xi = \begin{pmatrix} \lambda x \\ x \end{pmatrix}.$$

In many applications [16,18,23], the matrices A, B and C are Hermitian. Then the following reformulations of (3.1) are Hermitian GEPs:

$$\text{GEP}_2: \quad \begin{pmatrix} A & 0 \\ 0 & -C \end{pmatrix}\xi = \lambda\begin{pmatrix} 0 & A \\ A & B \end{pmatrix}\xi, \tag{3.3}$$

$$\text{GEP}_3: \quad \begin{pmatrix} B & C \\ C & 0 \end{pmatrix}\xi = \lambda\begin{pmatrix} -A & 0 \\ 0 & C \end{pmatrix}\xi. \tag{3.4}$$

These three are not the only possible formulations (see, e.g. [14]), but we will restrict our analysis to them as they are the most common in the literature. The analysis below is easily adapted for other formulations.

In practice, the choice of the GEP formulation depends on the properties of the matrices A, B and C. When A is Hermitian positive definite then the second matrix of the GEP (3.2) is Hermitian positive definite, too, and (3.2) can be transformed to a standard eigenvalue problem:

$$\begin{pmatrix} -A^{-1}B & -A^{-1}C \\ I & 0 \end{pmatrix} v = \lambda \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} v. \tag{3.5}$$

A similar approach can be taken when C is Hermitian positive definite by considering the GEP

$$L(\mu)x = 0, \tag{3.6}$$

where $L(\mu) = \mu^2 C + \mu B + A$ and $\mu = 1/\lambda$. When A, B, C are all real symmetric positive definite, Parlett and Chen [18] recommend the use of the GEP$_2$ formulation of (3.6) in the context of their pseudosymmetric Lanczos procedure.

For the special case $A = I$, Veselić [26] considers a class of transformations of the the QEP into standard eigenvalue problems that have the smallest Henrici departure from normality. Such transformations may decrease the number of iterations of some numerical diagonalization methods.

3.2. Backward error of the GEP solution of the QEP

We assume that we have a backward stable algorithm, such as the QZ algorithm, for computing a solution $(\tilde{\lambda}, \tilde{\xi})$ of a GEP

$$\mathscr{A}\xi = \lambda\mathscr{B}\xi.$$

This means that $(\tilde{\lambda}, \tilde{\xi})$ is the exact solution of a slightly perturbed pencil $(\tilde{\mathscr{A}}, \tilde{\mathscr{B}})$ with

$$\|\mathscr{A} - \tilde{\mathscr{A}}\|_2 \leqslant p_{\mathscr{A}} \|\mathscr{A}\|_2 u, \quad \|\mathscr{B} - \tilde{\mathscr{B}}\|_2 \leqslant p_{\mathscr{B}} \|\mathscr{B}\|_2 u, \tag{3.7}$$

where $p_{\mathscr{A}}$ and $p_{\mathscr{B}}$ are polynomial expressions in n and u is the machine precision. If the pencil $(\tilde{\mathscr{A}}, \tilde{\mathscr{B}})$ comes from a GEP formulation of a QEP, then, certainly, the perturbed matrices $(\tilde{\mathscr{A}}, \tilde{\mathscr{B}})$ will in general have lost their specific structure (see for example (3.2)–(3.4)), so that $\tilde{\mathscr{A}}\,\tilde{\xi} - \tilde{\lambda}\tilde{\mathscr{B}}\tilde{\xi}$ does not correspond to a GEP formulation of the QEP (3.1). Van Dooren and Dewilde [25] show that solving the PEP (1.1) with the companion formulation (1.5) and the QZ algorithm is backward stable, where for the backward error they used the weaker definition

$$\eta_{vdd} := \min\{\epsilon : (P(\tilde{\lambda}) + \Delta P(\tilde{\lambda}))\tilde{x} = 0,\ \|[\Delta A_m \ldots \Delta A_0]\|_F \leqslant \epsilon \|[A_m \ldots A_0]\|_F\}, \tag{3.8}$$

where the Frobenius norm $\|A\|_F = \mathrm{trace}(A^* A)^{1/2}$.

With our definition of backward error (2.2), in which each perturbation ΔA_i is measured relative to the matrix A_i that it perturbs, we need stronger assumptions on the norm of the coefficient matrices to get backward stability.

Theorem 7. *If* $\|A\|_2 = \|B\|_2 = \|C\|_2 = 1$ *then solving* GEP$_1$ *with a backward stable algorithm for the* GEP *is backward stable for the* QEP.

Proof. The proof is similar to the one given in [25]. Let $(\tilde{\mathscr{A}}, \tilde{\mathscr{B}})$ be the perturbed pencil in (3.7). Using block Gaussian elimination and block scaling of the pivots, we can construct nonsingular matrices G_1 and G_2 such that

$$\bar{\mathscr{A}} - \tilde{\lambda}\bar{\mathscr{B}} = G_1(\tilde{\mathscr{A}} - \tilde{\lambda}\tilde{\mathscr{B}})G_2$$

with

$$\bar{\mathscr{A}} = \begin{pmatrix} -\bar{B} & -\bar{C} \\ I & 0 \end{pmatrix}, \qquad \bar{\mathscr{B}} = \begin{pmatrix} \bar{A} & 0 \\ 0 & I \end{pmatrix},$$

that is, $\bar{\mathscr{A}} - \tilde{\lambda}\bar{\mathscr{B}}$ is again a GEP_1 formulation of a QEP. After some calculations, we find that

$$\|A - \bar{A}\|_2 \leqslant [\|A\|_2\|\mathscr{A}\|_2 + \|\mathscr{B}\|_2]u,$$

$$\|B - \bar{B}\|_2 \leqslant [(1 + \|B\|_2 + \|A\|_2)\|\mathscr{A}\|_2 + (1 + \|C\|_2)\|\mathscr{B}\|_2]u,$$

$$\|C - \bar{C}\|_2 \leqslant [(1 + \|B\|_2)\|\mathscr{A}\|_2 + \|C\|_2\|\mathscr{B}\|_2]u.$$

Backward stability is therefore assured for $\|A\|_2 = \|B\|_2 = \|C\|_2 = 1$. □

Note that this result holds for the GEP_1 formulation only. To illustrate, we carried out some experiments in Matlab, for which the unit roundoff is $u = 2^{-53} \approx 1.1 \times 10^{-16}$. We used the direct search maximization routine `mdsmax` of the Matlab Test Matrix Toolbox [12] and we applied it to the function $f(A, B, C) = \eta(\tilde{x}, \tilde{\lambda})$, where the eigenpair $(\tilde{\lambda}, \tilde{x})$ is computed using the QZ algorithm. It is easy to generate matrices A, B and C where $\|A\|_2 = \|B\|_2 = \|C\|_2 = 1$ and for which the backward error associated with the GEP_2 or GEP_3 formulation is large. As an illustration, we report in Table 1 the backward error for the smallest eigenvalue in absolute value of a 2 × 2 symmetric QEP for which A, B and C are given to three significant digits by

$$A = \begin{pmatrix} 9.88\mathrm{e}-1 & 1.49\mathrm{e}-1 \\ 1.49\mathrm{e}-1 & -8.04\mathrm{e}-1 \end{pmatrix}, \quad B = \begin{pmatrix} 9.70\mathrm{e}-1 & -7.77\mathrm{e}-2 \\ -7.77\mathrm{e}-2 & 7.97\mathrm{e}-1 \end{pmatrix},$$

$$C = \begin{pmatrix} 2.29\mathrm{e}-7 & 4.79\mathrm{e}-4 \\ 4.79\mathrm{e}-4 & 1 \end{pmatrix}. \tag{3.9}$$

Table 1
Backward errors $\eta(\tilde{x}, \tilde{\lambda})$ for the QEP with data (3.9)

GEP_2	GEP_3
2.8×10^{-16}	1.6×10^{-8}

We could not generate examples where the backward errors of the GEP$_2$ and GEP$_3$ formulations were simultaneously large. As expected from Theorem 7, all the QEPs we generated this way had good backward error when solved with the GEP$_1$ formulation. However, this is no longer true when A, B and C vary widely in norm.

As for some problems such as the solution of the Riccati equation [9], a scaling of the QEP could improve the backward error. We define the scaled QEP by

$$\mu^2 A_\alpha x + \mu B_\alpha x + Cx = 0, \tag{3.10}$$

with $\mu = \lambda/\alpha, A_\alpha = \alpha^2 A$ and $B_\alpha = \alpha B$, where α is the scaling factor. Note that the backward error is scale independent: if $\tilde{\mu} = \tilde{\lambda}/\alpha$, $\eta(\tilde{x}, \tilde{\lambda})$ for the original problem equals $\eta(\tilde{x}, \tilde{\mu})$ for (3.10). We generated a QEP where $\|A\|_2 = 10^3$, $\|B\|_2 = 10^2$ and $\|C\|_2 = 10^{-4}$ and we used the GEP$_1$ formulation with the QZ algorithm to solve it. The backward error associated with the computation of the smallest eigenvalue in absolute value without scaling was seven orders of magnitude larger than the machine precision. We plot in Fig. 1 the influence of a scaling α in the range $[0, 1]$ on the backward error and on the condition number. We also

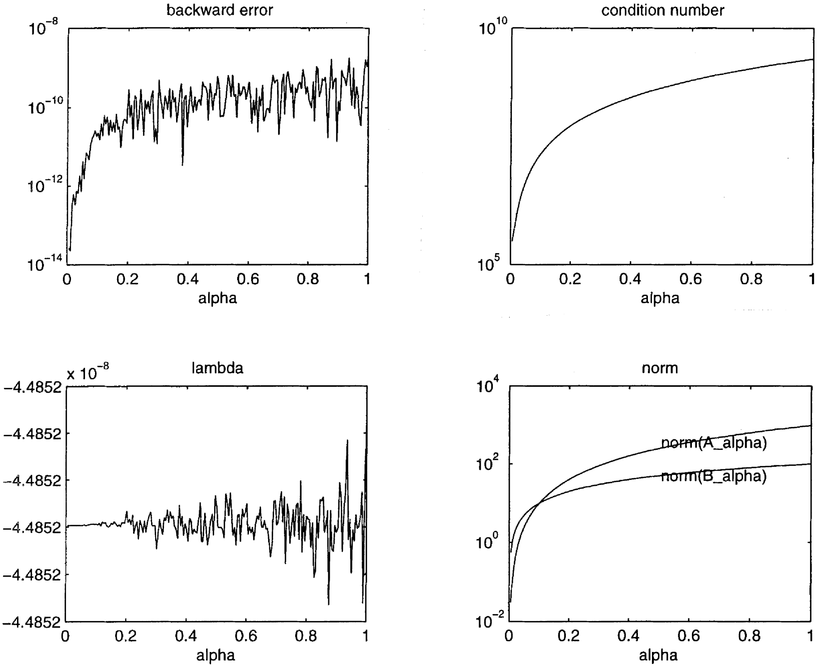

Fig. 1. Effect of varying scale parameter α.

plot the variation in the computed eigenvalue and the norms of A_α, B_α. In this particular example, the backward error is improved for sufficiently small values of α. The variation in the computed eigenvalue decreases as the backward error decreases. It seems that best results are obtained when $\|A_\alpha\|_2 \approx \|B_\alpha\|_2$. Theorem 7 suggests choosing α such that $\|A_\alpha\|_2 = \|B_\alpha\|_2 = \|C\|_2$ but we cannot achieve this with only one parameter at our disposal. It is not clear how to choose an α that will ensure a good backward error.

Eigenvectors for the QEP (3.1) can be recovered as $x_1 = \tilde{\xi}(1:n)/\lambda$ or $x_2 = \tilde{\xi}(n+1:2n)$. In our experiments, we noticed that the computed $\tilde{x}_1$ and $\tilde{x}_2$ can have greatly different backward errors. Consider the GEP (3.2) corresponding to the QEP (3.1) with

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1, \end{pmatrix} \quad B = \begin{pmatrix} 1 & 1 \\ 0 & 1, \end{pmatrix} \quad C = \begin{pmatrix} -2t & 1 \\ 0 & 4t^2 \end{pmatrix} \tag{3.11}$$

and $t = 10^{-5}$. We report in Table 2 the backward error for the eigenvalue of smallest absolute value, $\lambda = -4 \times 10^{-10}$. While $\eta(\tilde{x}_2, \tilde{\lambda})$ reveals good stability, the backward error for $(\tilde{x}_1, \tilde{\lambda})$ is six orders of magnitude larger than that for $(\tilde{x}_2, \tilde{\lambda})$. The reason is that $\tilde{x}_1$ is determined from small components of ξ and these small components are computed relatively inaccurately. In examples where λ is large, we find the converse situation: $\tilde{x}_1$ gives a small backward error but $\tilde{x}_2$ does not. We conclude that one should determine the QEP eigenvector x using whichever is the larger of x_1 and x_2 (which we did in the experiments earlier in this section). This observation does not seem to have appeared in the literature before, although it may be known to practitioners.

3.3. Condition numbers for the GEP formulations

Condition numbers and backward error are related to the accuracy of the solutions by the inequality (1.3). Most of the algorithms applied to a GEP form of the QEP do not preserve the structure. Hence it is the condition number of the GEP form which is relevant. From Theorem 5 the normwise condition number for λ of the pair $(\mathscr{A}, \mathscr{B})$ is given by

$$\kappa(\lambda, \mathscr{A}, \mathscr{B}) = \frac{\|\chi\|_2\|\xi\|_2(\|\mathscr{E}_{\mathscr{A}}\|_2 + |\lambda|\|\mathscr{E}_{\mathscr{B}}\|_2)}{|\lambda||\chi^*\mathscr{B}\xi|}, \tag{3.12}$$

where χ and ξ are corresponding left and right eigenvectors with $\xi = [\lambda x^{\mathrm{T}}\, x^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathscr{E}_{\mathscr{A}}, \mathscr{E}_{\mathscr{B}}$ are the matrices of tolerances against which the perturbations to $\mathscr{A}$

Table 2
Backward errors for the QEP with data (3.11)

$\eta(\tilde{x}_1, \tilde{\lambda})$	$\eta(\tilde{x}_2, \tilde{\lambda})$
1×10^{-11}	2×10^{-17}

and $\mathscr{B}$ are measured. The left eigenvectors of the GEPs are also related to those of the QEP, though not in the same way in each case.

Lemma 8. *Let λ be an eigenvalue of $Q(\lambda)$ and let y be a corresponding left eigenvector. Then*

$$\chi = \begin{pmatrix} \bar{\lambda} y \\ y \end{pmatrix}$$

is a left eigenvector of GEP_2 *and* GEP_3 *with corresponding eigenvalue λ and*

$$\chi_1 = \begin{pmatrix} \bar{\lambda} y \\ -C^* y \end{pmatrix}$$

is a left eigenvector of GEP_1 *with corresponding eigenvalue λ.*

We define $(\mathscr{A}_i, \mathscr{B}_i)$ to be the pairs of matrices involved in the GEP_i formulations of the QEP (3.1), for $i = 1 : 3$. The next theorem gives an expression for the condition numbers $\kappa(\lambda, \mathrm{GEP}_i)$, with $\mathscr{E}_{\mathscr{A}_i} = \mathscr{A}_i$ and $\mathscr{E}_{\mathscr{B}_i} = \mathscr{B}_i$. It is now convenient to impose a normalization on the eigenvectors of the QEP.

Theorem 9. *Let λ be a simple eigenvalue of $Q(\lambda)$ and x, y be corresponding right and left eigenvectors normalized so that $y^* Q'(\lambda) x = 1$. Define*

$$f(z) = \frac{\|C^* z\|_2}{\|z\|_2}.$$

Then

$$\kappa(\lambda, \mathrm{GEP}_1) = \frac{\sqrt{(1 + |\lambda|^2)(|f(y)|^2 + |\lambda|^2)}}{|\lambda|^2} (\|\mathscr{A}_1\|_2 + |\lambda| \|\mathscr{B}_1\|_2) \|x\|_2 \|y\|_2,$$
$$\kappa(\lambda, \mathrm{GEP}_2) = \frac{1 + |\lambda|^2}{|\lambda|} (\|\mathscr{A}_2\|_2 + |\lambda| \|\mathscr{B}_2\|_2) \|x\|_2 \|y\|_2,$$
$$\kappa(\lambda, \mathrm{GEP}_3) = \frac{1 + |\lambda|^2}{|\lambda|^2} (\|\mathscr{A}_3\|_2 + |\lambda| \|\mathscr{B}_3\|_2) \|x\|_2 \|y\|_2,$$

Proof. From Lemma 8 we have

$$\|\chi\|_2 = \sqrt{1 + |\lambda^2|}\, \|y\|_2, \quad \|\xi\|_2 = \sqrt{1 + |\lambda^2|}\, \|x\|_2,$$
$$\|\chi_1\|_2^2 = |\lambda|^2 \|y\|_2^2 + \|C^* y\|_2^2 = (|f(y)|^2 + |\lambda|^2) \|y\|_2^2.$$

Using $y^*Q'(\lambda)x = 1$, we get for the GEP$_2$ formulation

$$|\chi^*\mathscr{B}_2\xi| = |2\lambda y^*Ax + y^*Bx| = 1.$$

For GEP$_1$ and GEP$_3$ we have

$$\begin{aligned}|\chi_1^*\mathscr{B}_1\xi| = |\chi^*\mathscr{B}_3\xi| &= |-\lambda^2 y^*Ax + y^*Cx| \\ &= |-2\lambda^2 y^*Ax - \lambda y^*Bx| \\ &= |-2\lambda^2 y^*Ax - \lambda + 2\lambda^2 y^*Ax| = |\lambda|.\end{aligned}$$

The result follows on substituting into (3.12). □

The next lemma shows that the condition number for λ in the GEP$_2$ formulation (respectively GEP$_3$ formulation) of the QEP (3.1) is the condition number for $\mu = 1/\lambda$ in the GEP$_3$ (respectively GEP$_2$ formulation) of the QEP (3.6).

Lemma 10. *Let λ be a simple and nonzero eigenvalue of $Q(\lambda)$ and let $\mu = 1/\lambda$ be an eigenvalue of $L(\mu) = \mu^2C + \mu B + A$. Then*

$$\kappa(\mu, \mathrm{GEP}_2) = \kappa(\lambda, \mathrm{GEP}_3), \qquad \kappa(\mu, \mathrm{GEP}_3) = \kappa(\lambda, \mathrm{GEP}_2).$$

Proof. Let x, y be the left and right eigenvectors of $Q(\lambda)$ and $L(\mu)$ normalized so that $y^*Q'(\lambda)x = 1$. We define

$$\mathscr{A}'_2 = \begin{pmatrix} C & 0 \\ 0 & -A \end{pmatrix}, \qquad \mathscr{B}'_2 = \begin{pmatrix} 0 & C \\ C & B \end{pmatrix},$$

so that the GEP$_2$ formulation of $L(\mu)x = 0$ is

$$\mathscr{A}'_2\xi' = \mu\mathscr{B}'_2\xi', \text{ with } \xi' = \begin{pmatrix} \mu x \\ x \end{pmatrix}.$$

Let $\chi' = (\bar{\mu}y^{\mathrm{T}}\ y^{\mathrm{T}})^{\mathrm{T}}$. From the normalization condition $y^*Q'(\lambda)x = 1$ and the fact that $\mu = 1/\lambda$ and $y^*Cx = -\lambda^2 y^*Ax - \lambda y^*Bx$, we have

$$|\chi'^*\mathscr{B}'_2\xi'| = |2\mu y^*Cx + y^*Bx| = 1.$$

Then,

$$\begin{aligned}\kappa(\mu, \mathrm{GEP}_2) &= \frac{\|\chi'\|_2\|\xi'\|_2(\|\mathscr{A}'_2\|_2 + |\mu|\|\mathscr{B}'_2\|_2)}{|\mu||\chi'^*\mathscr{B}'_2\xi'|} \\ &= \frac{1+|\lambda|^2}{|\lambda|^2}(|\lambda|\|\mathscr{A}'_2\|_2 + \|\mathscr{B}'_2\|_2)\|x\|_2\|y\|_2.\end{aligned}$$

But, $\|\mathscr{A}_2'\|_2 = \|\mathscr{B}_3\|_2$ and $\|\mathscr{B}_2'\|_2 = \|\mathscr{A}_3\|_2$ and from Theorem 9 we get that $\kappa(\mu, \mathrm{GEP}_2) = \kappa(\lambda, \mathrm{GEP}_3)$. The proof of $\kappa(\mu, \mathrm{GEP}_3) = \kappa(\lambda, \mathrm{GEP}_2)$ is similar. □

These three condition numbers are quite different, but given some information on $\|A\|_2, \|B\|_2$ and $\|C\|_2$ it is possible to compare them. As an illustration, we consider in the next corollary the case where all the matrices have unit norms.

Corollary 11. *If* $\|A\|_2 = \|B\|_2 = \|C\|_2 = 1$, *then, under the assumptions of Theorem* 9, *we have*

$$\begin{aligned}
&\frac{\sqrt{(1+|\lambda|^2)(f(y)^2+|\lambda|^2)}}{|\lambda|^2}(1+|\lambda|)\\
&\quad\leqslant \frac{\kappa(\lambda, \mathrm{GEP}_1)}{\|x\|_2\|y\|_2} \leqslant \frac{\sqrt{(1+|\lambda|^2)(f(y)^2+|\lambda|^2)}}{|\lambda|^2}(2+|\lambda|),\\
&\frac{(1+|\lambda|^2)}{|\lambda|}(1+|\lambda|) \leqslant \frac{\kappa(\lambda, \mathrm{GEP}_2)}{\|x\|_2\|y\|_2} \leqslant \frac{(1+|\lambda|^2)}{|\lambda|}(1+2|\lambda|),\\
&\frac{(1+|\lambda|^2)}{|\lambda|^2}(1+|\lambda|) \leqslant \frac{\kappa(\lambda, \mathrm{GEP}_3)}{\|x\|_2\|y\|_2} \leqslant \frac{(1+|\lambda|^2)}{|\lambda|^2}(2+|\lambda|).
\end{aligned}$$

Proof. Note that

$$\|\mathscr{A}_2\|_2 = \|\mathscr{B}_1\|_2 = \|\mathscr{B}_3\|_2 = 1. \tag{3.13}$$

For $\mathscr{A}_1$ we have

$$\|\mathscr{A}_1\|_2 \leqslant \left\|\begin{pmatrix} -B & 0 \\ 0 & 0 \end{pmatrix}\right\|_2 + \left\|\begin{pmatrix} 0 & C \\ I & 0 \end{pmatrix}\right\|_2 = 2$$

and

$$\|\mathscr{A}_1\|_2 \geqslant \left\|\begin{pmatrix} 0 & C \\ 0 & 0 \end{pmatrix}\right\|_2 = 1.$$

The same argument works for $\|\mathscr{A}_3\|_2$ and $\|\mathscr{B}_2\|_2$ so that

$$1 \leqslant \|\mathscr{A}_1\|_2 \leqslant 2, \qquad 1 \leqslant \|\mathscr{A}_3\|_2 \leqslant 2, \qquad 1 \leqslant \|\mathscr{B}_2\|_2 \leqslant 2. \tag{3.14}$$

Substituing (3.13) and (3.14) into Theorem 9 gives the result. □

We can now compare the condition numbers according to the magnitude of $|\lambda|$.

Corollary 12. *If* $\|A\|_2 = \|B\|_2 = \|C\|_2 = 1$, *then, under the assumptions of Theorem* 9, *we have*

$$\begin{aligned}
&\textit{for } |\lambda| \geqslant \sqrt{2}, \quad \kappa(\lambda, \mathrm{GEP}_1) \leqslant \kappa(\lambda, \mathrm{GEP}_2), \quad \kappa(\lambda, \mathrm{GEP}_3) \leqslant \kappa(\lambda, \mathrm{GEP}_2),\\
&\textit{for } |\lambda| \leqslant 2^{-1/2}, \quad \kappa(\lambda, \mathrm{GEP}_2) \leqslant \kappa(\lambda, \mathrm{GEP}_3),\\
&\textit{for } |\lambda| \ll 1, \quad \kappa(\lambda, \mathrm{GEP}_2) \ll \kappa(\lambda, \mathrm{GEP}_3),\\
&\textit{for } |\lambda| \gg 1, \quad \kappa(\lambda, \mathrm{GEP}_2) \gg \kappa(\lambda, \mathrm{GEP}_3).
\end{aligned}$$

Moreover, if $|\lambda| \ll 1$ *and* $f(y) \ll 1$ *then*

$$\kappa(\lambda, \mathrm{GEP}_1) \ll \kappa(\lambda, \mathrm{GEP}_3).$$

The theorem and the corollary show that an eigenvalue of the QEP may be much more or less sensitive to perturbations in the different GEP formulations. Of course, the perturbations allowed in the definitions of $\kappa(\lambda, \mathrm{GEP}_i), i = 1:3$, do not preserve the structure of the problems; if they did, then these condition numbers would be equal to the condition number $\kappa(\lambda, Q)$ for the QEP in (2.15). The practical relevance of our observation is that the standard algorithm for solving the QEP, the QZ algorithm, [8,17], does not preserve the structure of the GEP formulations of the QEP in its backward error results.

We can also compare the condition numbers of the QEP $Q(\lambda)$ with that of the "reversed form" $L(\mu)$.

Lemma 13. *Let* λ *be a nonzero eigenvalue of* $Q(\lambda)$, *and let* $\mu = 1/\lambda$ *be an eigenvalue of* $L(\mu) = \mu^2 C + \mu B + A$. *If* $\|A\|_2 = \|B\|_2 = \|C\|_2 = 1$ *then, for* $|\lambda| \ll 1$ *and* $|f(y)| > |\lambda|$,

$$\kappa_L(\mu, \mathrm{GEP}_1) < \kappa_Q(\lambda, \mathrm{GEP}_1)$$

and for $|\lambda| \gg 1$,

$$\kappa_L(\mu, \mathrm{GEP}_1) > \kappa_Q(\lambda, \mathrm{GEP}_1)$$

Proof. Let $(\mathscr{A}_1', \mathscr{B}_1')$ the pair of matrices corresponding to the GEP_1 formulation of the QEP $L(\mu)x = 0$. The corresponding left and right eigenvectors with associated eigenvalue $\mu = 1/\lambda$ are given by

$$\chi_1' = \begin{pmatrix} \bar{\mu} y \\ -A^* y \end{pmatrix}, \quad \xi' = \begin{pmatrix} \mu x \\ x \end{pmatrix}.$$

It is easy to show that

$$\kappa(\mu, \mathrm{GEP}_1) = \frac{\sqrt{(1+|\lambda|^2)(|\lambda|^2|g(y)|^2+1)}}{|\lambda|}(|\lambda|\|\mathscr{A}_1'\|_2 + \|\mathscr{B}_1'\|_2)\|x\|_2\|y\|_2,$$

where $g(y) = \|A^*y\|_2/\|y\|_2$. As $\|\mathscr{B}_1'\|_2 = 1$ and $1 \leqslant \|\mathscr{A}_1'\|_2 \leqslant 2$, the first inequality of the lemma follows. □

In practice, when an iterative method, such as the Arnoldi method, is used to find a few low-frequency modes λ (small λ, that is, large μ), the GEP_1 formulation of $L(\mu)$ seems to be preferred to the GEP_1 formulation of $Q(\lambda)$ [3]. Fortunately, the previous theorem states that for a small eigenvalue λ, the GEP_1 formulation of $L(\mu)$ leads to a better condition number than the GEP_1 formulation of $Q(\lambda)$.

In our experiments, we used the implementation of the QZ algorithm in the LAPACK library (routine xGEGV) to compute the solution of the GEP. Unlike the qz function of Matlab, which is based on a routine from EISPACK and does not compute left eigenvectors, this routine computes both left and right eigenvectors. In the current release of LAPACK (verion 2.0) the routine xGEGV performs by default "full balancing" on the matrices $\mathscr{A}$ and $\mathscr{B}$. This involves permutations together with a diagonal similarity transformation ("scaling") to make rows and columns as close in norm as possible. Full balancing is an attempt to reduce the 1-norms of the matrices and to improve the accuracy of the computed eigenvalues and eigenvectors of the GEP [27] but it has no influence on the conditioning of the GEP. The rest of our computations were carried out with Matlab. The approximate eigenvalue $\tilde{\lambda}$ was computed in single precision (unit roundoff $u_\mathrm{s} = 2^{-24} \simeq 5.9 \times 10^{-8}$). For the computation of the relative error, the condition number and the backward error, we took as exact eigenpair the eigenvalue and eigenvector computed in double precision ($u_\mathrm{d} = 2^{-53} \simeq 1.1 \times 10^{-16}$). In our experiments we computed these measures both with and without scaling and did not notice any major differences.

Example 1. To verify the results of Corollary 12, we generated random matrices of 2-norm approximately 1. As an illustration, consider the following matrices for which we give only two significant digits:

$$\begin{aligned} A &= \begin{pmatrix} 0.81 & -0.38 \\ -0.41 & 0.19 \end{pmatrix}, \quad B = \begin{pmatrix} -0.24 & 0.97 \\ -0.021 & 0.086 \end{pmatrix}, \\ C &= \begin{pmatrix} -0.11 & -0.057 \\ -0.88 & -0.46 \end{pmatrix}. \end{aligned} \tag{3.15}$$

The associated QEP has one large eigenvalue $\lambda_1 = 2.16 \times 10^2$, one small eigenvalue $\lambda_2 = -4.61 \times 10^{-4}$ and a complex conjugate pair $\lambda_{3,4} = 0.911 \pm 1.22i$.

Table 3
Relative errors and condition numbers for the QEP (3.15)

		Type of formulation		
		GEP$_1$	GEP$_2$	GEP$_3$
$\|\lambda_1\|$	Relative error	6.6e − 5	5.8e − 3	2.14e − 5
2.1×10^2	Condition number	5.3e + 2	1.7e + 5	5.3e + 2
$\|\lambda_2\|$	Relative error	8.3e − 5	1.1e − 5	6.4e − 2
4.6×10^{-4}	Condition number	4.5e + 3	2.3e + 3	5.3e + 6
$\|\lambda_{3,4}\|$	Relative error	1.8e − 7	4.7e − 7	1.8e − 7
1.5246	Condition number	6.6	11.5	6.3

We report in Table 3 the relative errors and condition numbers. As expected, for the large eigenvalue,

$$\kappa(\lambda, \mathrm{GEP}_2) \gg \kappa(\lambda, \mathrm{GEP}_3), \qquad \kappa(\lambda, \mathrm{GEP}_2) > \kappa(\lambda, \mathrm{GEP}_1).$$

For the small eigenvalue,

$$\kappa(\lambda, \mathrm{GEP}_2) \ll \kappa(\lambda, \mathrm{GEP}_3), \qquad \kappa(\lambda, \mathrm{GEP}_1) \ll \kappa(\lambda, \mathrm{GEP}_3)$$

with $f(y) = 4.5 \times 10^{-4} \ll 1$ as in the assumption of Corollary 12. Note that the relative errors of the computed eigenvalues reflect the conditioning of the GEP, confirming that the accuracy of the computed eigenvalues depends on the choice of GEP formulation.

Example 2. We now consider the connected damped mass-spring system illustrated in Fig. 2. The ith mass of weight m_i is connected to its $(i+1)$st neighbour by a spring and a damper with constants k_i and d_i, respectively. The ith mass is also connected to the ground by a spring and a damper with constants κ_i and τ_i, respectively. The vibration of this system is governed by a second-order differential equation

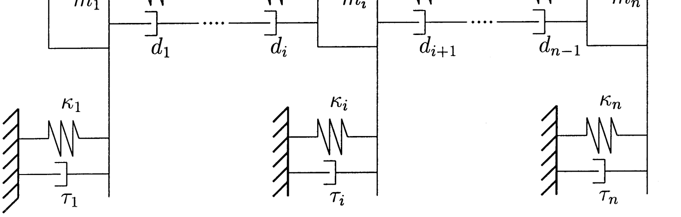

Fig. 2. An n degree of freedom damped mass-spring system.

$$M\frac{\mathrm{d}^2}{\mathrm{d}t^2}x + D\frac{\mathrm{d}}{\mathrm{d}t}x + Kx = 0,$$

where the mass matrix $M = \mathrm{diag}(m_1, \ldots, m_n)$ is diagonal, and the damping matrix D and stiffness matrix K are symmetric tridiagonal and are defined by

$$D = P\,\mathrm{diag}(d_1, \ldots, d_{n-1}, 0)P^{\mathrm{T}} + \mathrm{diag}(\tau_1, \ldots, \tau_n),$$
$$K = P\,\mathrm{diag}(k_1, \ldots, k_{n-1}, 0)P^{\mathrm{T}} + \mathrm{diag}(\kappa_1, \ldots, \kappa_n),$$

with $P = (\delta_{ij} - \delta_{i,j+1})$, where δ_{ij} is the Kronecker delta.

Let $m = \max_{1 \leqslant i \leqslant n} m_i$, $d = \|D\|_2$ and $k = \|K\|_2$. Then we have

$$\|\mathscr{A}_2\|_2 = \max(m, k), \qquad k \leqslant \|\mathscr{A}_3\|_2 \leqslant d + k,$$
$$m \leqslant \|\mathscr{B}_2\|_2 \leqslant m + d, \qquad \|\mathscr{B}_3\|_2 = \max(m, k).$$

Using Theorem 9, we conclude that

$$\kappa(\lambda, \mathrm{GEP}_2) \geqslant \kappa(\lambda, \mathrm{GEP}_3) \quad \text{for} \quad |\lambda| \geqslant \left(\frac{d+k}{m}\right)^{1/2},$$
$$\kappa(\lambda, \mathrm{GEP}_2) \leqslant \kappa(\lambda, \mathrm{GEP}_3) \quad \text{for} \quad |\lambda| \leqslant \left(\frac{k}{m+d}\right)^{1/2}.$$

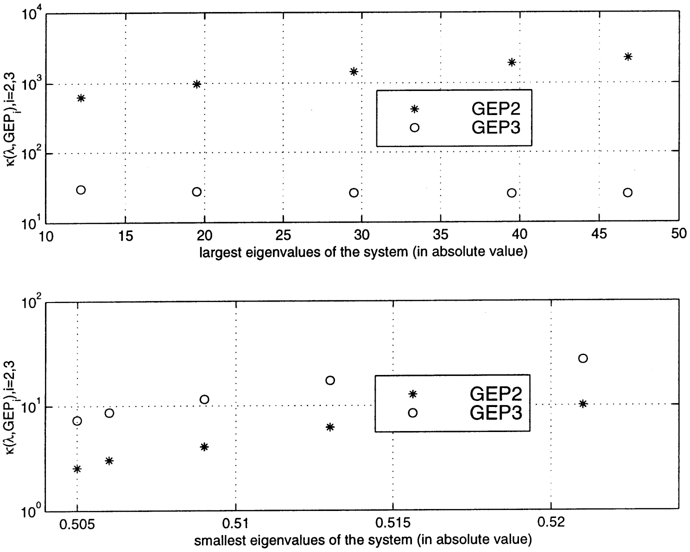

Fig. 3. Conditioning of the eigenvalues of a 5-degree of freedom damped mass-spring system.

In our experiments, we took all the springs (respectively dampers) to have the same constant κ (respectively τ), except the first and last ones for which $\kappa_1 = \kappa_n = 2\kappa$ and $\tau_1 = \tau_n = 2\tau$. Then

$$D = \tau \,\mathrm{tridiag}(-1, 3, -1), \quad K = \kappa \,\mathrm{tridiag}(-1, 3, -1),$$

so that we have an explicit formula for $d = \|D\|_2$ and $k = \|K\|_2$ depending on the degrees of freedom of the damped mass-spring system n and the constants τ and κ.

For $n = 5,\ m = 1,\ \tau = 10,\ k = 5$, we plot in Fig. 3 the condition number of each eigenvalue for the GEP_2 and GEP_3 formulations. The theory says that

$$\kappa(\lambda, \mathrm{GEP}_2) \geqslant \kappa(\lambda, \mathrm{GEP}_3) \quad \text{for} \quad |\lambda| \geqslant 3.8730,$$

$$\kappa(\lambda, \mathrm{GEP}_2) \leqslant \kappa(\lambda, \mathrm{GEP}_3) \quad \text{for} \quad |\lambda| \leqslant 0.6742,$$

which is confirmed by the numerical results.

4. Conclusions

We have derived new computable backward errors and condition numbers for the PEP. The most common way of dealing with the PEP is to reformulate it as a GEP. We used our expressions to show that backward stable algorithms for the GEP that do not respect the special structure of the GEP formulations can be backward unstable for the QEP. We investigated the possibility of using a scaling of the QEP to improve the backward error of the solutions obtained via the GEP formulations. It is an open problem to find a scaling that optimizes the backward stability.

We analyzed the sensitivity of three GEP formulations of a QEP and showed that given some information on the norm of the coefficient matrices we can identify which formulations are preferred for the large and the small eigenvalues, respectively. These results are of practical relevance as in applications it is often only the eigenpairs corresponding to small or large eigenvalues that are of interest [2,14].

Acknowledgements

This work was supported by Engineering and Physical Sciences Research Council grant GR/L76532.

References

[1] A.L. Andrew, K.-W. Eric Chu, P. Lancaster, Derivatives of eigenvalues and eigenvectors of matrix functions, SIAM J. Matrix Anal. Appl. 14 (4) (1993) 903–926.

[2] M. Borri, P. Mantegazza, Efficient solution of quadratic eigenproblems arising in dynamic analysis of structures, Comput. Meth. Appl. Mech. Engrg. 12 (1977) 19–31.

[3] H.C. Chen, Partial eigensolution of damped structural systems by Arnoldi's method, Earthquake Engrg. Struct. Dyn. 22 (1993) 63–74.

[4] R.W. Clough, S. Mojtahedi, Earthquake response analysis considering non-proportional damping, Earthquake Engrg. Struct. Dyn. 4 (1976) 489–496.

[5] R.J. Duffin, A minimax theory for overdamped networks, J. Rat. Mech. Anal. 4 (1955) 221–233.

[6] V. Frayssé, V. Toumazou, A note on the normwise perturbation theory for the regular generalized eigenproblem, Numer. Linear Algebra Appl. 5 (1) (1998) 1–10.

[7] W. Gander, G.H. Golub, U. Von Matt, A constrained eigenvalue problem, Linear Algebra and Appl. 114/115 (1989) 815–839.

[8] G.H. Golub, C.F. Van Loan, Matrix Computations, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[9] T. Gudmundsson, C. Kenney, A.J. Laub, Scaling of the discrete-time algebraic Riccati equation to enhance stability of the Schur solution method, IEEE Trans. Automat. Control 37 (4) (1992) 513–518.

[10] P. Guillaume, Nonlinear eigenproblems, SIAM J. Matrix Anal. Appl. 20 (3) (1999) 575–595.

[11] D.J. Higham, N.J. Higham, Structured backward error and condition of generalized eigenvalue problems, SIAM J. Matrix Anal. Appl. 20 (2) (1998) 493–512.

[12] N.J. Higham, The Test Matrix Toolbox for MATLAB (version 3.0), Numerical Analysis Report No. 276, Manchester Centre for Computational Mathematics, Manchester, England, 1995.

[13] W. Kahan, B.N. Parlett, E. Jiang, Residual bounds on approximate eigensystems of nonnormal matrices, SIAM J. Numer. Anal. 19 (3) (1982) 470–484.

[14] L. Komzsik, Implicit computational solution of generalized quadratic eigenvalue problems, Manuscript, The MacNeal–Schwendler Corporation, 1998.

[15] V.N. Kublanovskaya, On an approach to the solution of the generalized latent value problem for lambda-matrices, SIAM J. Numer. Anal. 7 (1970) 532–537.

[16] P. Lancaster, Lambda-Matrices and Vibrating Systems, Pergamon Press, Oxford, 1966.

[17] C.B. Moler, G.W. Stewart, An algorithm for generalized matrix eigenvalue problems, SIAM J. Numer. Anal. 10 (2) (1973) 241–256.

[18] B.N. Parlett, H.C. Chen, Use of indefinite pencils for computing damped natural modes, Linear Algebra Appl. 140 (1990) 53–88.

[19] G. Peters, J.H. Wilkinson, $Ax = \lambda Bx$ and the generalized eigenproblem, SIAM J. Numer. Anal. 7 (4) (1970) 479–492.

[20] J.L. Rigal, J. Gaches, On the compatibility of a given solution with the data of a linear system, J. Assoc. Comput. Mach. 14 (3) (1967) 543–548.

[21] A. Ruhe, Algorithms for the nonlinear eigenvalue problem, SIAM J. Numer. Anal. 10 (1973) 674–689.

[22] G.L.G. Sleijpen, A.G.L. Booten, D.R. Fokkema, H.A. Van der Vorst, Jacobi–Davidson type methods for generalized eigenproblems and polynomial eigenproblems, BIT 36 (3) (1996) 595–633.

[23] H.A. Smith, R.K. Singh, D.C. Sorensen, Formulation and solution of the non-linear, damped eigenvalue problem for skeletal systems, Int. J. Numer. Meth. Engrg. 38 (1995) 3071–3085.

[24] G.W. Stewart, J. Sun, Matrix Perturbation Theory, Academic Press, London, 1990.

[25] P.M. Van Dooren, P. Dewilde, The eigenstructure of an arbitrary polynomial matrix: Computational aspects, Linear Algebra and Appl. 50 (1983) 545–579.

[26] K. Veselić, On optimal linearisations of a quadratic eigenvalue problem, Linear and Multilinear Algebra 8 (1980) 253–258.

[27] R.C. Ward, Balancing the generalized eigenvalue problem, SIAM J. Sci. Stat. Comput. 2 (2) (1981) 141–152.

[28] Z.C. Zheng, G.X. Ren, W.J. Wang, A reduction method for large scale unsymmetric eigenvalue problems in structural dynamics, J. Sound and Vibration 199 (2) (1997) 253–268.