SIAM J. SCI. COMPUT.
Vol. 25, No. 2, pp. 591–603

ALTERNATIVES TO THE RAYLEIGH QUOTIENT FOR THE QUADRATIC EIGENVALUE PROBLEM*

MICHIEL E. HOCHSTENBACH† AND HENK A. VAN DER VORST†

Abstract. We consider the quadratic eigenvalue problem $\lambda^2 Ax + \lambda Bx + Cx = 0$. Suppose that u is an approximation to an eigenvector x (for instance, obtained by a subspace method) and that we want to determine an approximation to the corresponding eigenvalue λ. The usual approach is to impose the Galerkin condition $r(\theta, u) = (\theta^2 A + \theta B + C)u \perp u$, from which it follows that θ must be one of the two solutions to the quadratic equation $(u^*Au)\theta^2 + (u^*Bu)\theta + (u^*Cu) = 0$. An unnatural aspect is that if $u = x$, the second solution has in general no meaning. When u is not very accurate, it may not be clear which solution is the best. Moreover, when the discriminant of the equation is small, the solutions may be very sensitive to perturbations in u.

In this paper we therefore examine alternative approximations to λ. We compare the approaches theoretically and by numerical experiments. The methods are extended to approximations from subspaces and to the polynomial eigenvalue problem.

Key words. quadratic eigenvalue problem, Rayleigh quotient, Galerkin, minimum residual, subspace method, polynomial eigenvalue problem, backward error, refining a Ritz pair

AMS subject classifications. 65F15, 65F50

DOI. 10.1137/S1064827502406403

1. Introduction.

We first consider the eigenvalue problem $Ax = \lambda x$, with A a real symmetric $n \times n$ matrix, where $n \geq 2$. Suppose that we have an approximate eigenvector u with unit norm. The usual approximation to the corresponding eigenvalue is given by the *Rayleigh quotient* of u,

$$\rho = \rho(u) := \frac{u^*Au}{u^*u}. \tag{1.1}$$

This Rayleigh quotient has the following attractive properties:

1. ρ satisfies the *Ritz–Galerkin condition* on the *residual* $r(\theta, u)$,
$$r(\rho, u) := Au - \rho u \perp u. \tag{1.2}$$
2. ρ satisfies the *minimum residual condition* on the residual,
$$\rho = \operatorname*{argmin}_{\theta \in \mathbb{R}} \|Au - \theta u\|. \tag{1.3}$$
(Here and elsewhere in the paper, $\|\cdot\|$ stands for $\|\cdot\|_2$.)
3. The function $\rho(u)$ has as its *stationary points* exactly the n eigenvectors x_i and even
$$\frac{d\rho}{du}(x_i) = 0. \tag{1.4}$$
(Recall that *stationary* means that all directional derivatives are zero.) This implies that a first order perturbation of the eigenvector only gives a second order perturbation of the Rayleigh quotient: $\rho(x_i + h) = \lambda_i + \mathcal{O}(\|h\|^2)$.

*Received by the editors April 25, 2002; accepted for publication (in revised form) March 12, 2003; published electronically November 11, 2003.
http://www.siam.org/journals/sisc/25-2/40640.html
†Mathematical Institute, Utrecht University, P.O. Box 80.010, NL-3508 TA Utrecht, The Netherlands (hochstenbach@math.uu.nl, www.math.uu.nl/people/hochstenbach; vorst@math.uu.nl, www.math.uu.nl/people/vorst).

Remark 1.1. When A is normal, (1.2) and (1.3) hold, and the eigenvectors are still stationary points, although the Rayleigh quotient is in general not differentiable. When A is nonnormal, (1.2) and (1.3) also hold, but the eigenvectors are in general not stationary points. One can show that instead of this, the *two-sided Rayleigh quotient* $\rho(u,v) := \frac{v^*Au}{v^*u}$ has as its stationary points exactly the right/left eigenvector combinations (x_i, y_i) (provided that $y_i^* x_i \neq 0$). This suggests replacing the Ritz–Galerkin condition (1.2) with the *Petrov–Galerkin condition*,

$$r(\theta, u) = Au - \theta u \perp v,$$

which is used in two-sided methods such as two-sided Lanczos [5] and two-sided Jacobi–Davidson [3]. However, we use no information about the left eigenvector in this paper.

Now consider the quadratic eigenvalue problem

$$Q(\lambda)x := (\lambda^2 A + \lambda B + C)x = 0, \tag{1.5}$$

where A, B, and C are (complex) $n \times n$ matrices. In this paper, we examine generalizations of the properties (1.2)–(1.4) for the quadratic eigenvalue problem to derive different eigenvalue approximations. See [8] for an overview of the quadratic eigenvalue problem. For an eigenvector x we have one of the following properties:

- Ax and Bx are dependent; hence Cx is also dependent, and there are two eigenvalues (counting multiplicities) corresponding to x.
- Ax and Bx are independent; hence Cx lies in the span of Ax and Bx, and the corresponding eigenvalue λ is unique.

We will assume in the remainder of the paper that x has the second property. For a motivation see Remark 2.3 at the end of subsection 2.4.

Now let u be an approximation to an eigenvector x, for instance, one obtained by a subspace method. We will also assume that Au and Bu are independent, which is not unnatural in view of the assumptions that Ax and Bx are independent, and the assumption that $u \approx x$; see also Remark 2.3. We study ways to determine an approximation θ to the eigenvalue λ given an approximate eigenvector u. In subsection 2.1 we discuss the "classical" one-dimensional Galerkin method, while in subsections 2.2, 2.3, and 2.4 we introduce new approaches. The methods are extended to subspaces of dimensions larger than one and to the polynomial eigenvalue problem in section 3. Numerical experiments and a conclusion can be found in sections 4 and 5.

2. Approximations for the quadratic eigenvalue problem.

2.1. One-dimensional Galerkin. For an approximate eigenpair $(\theta, u) \approx (\lambda, x)$ we define the *residual* $r(\theta, u)$ by

$$r(\theta, u) := Q(\theta)u = (\theta^2 A + \theta B + C)u.$$

The usual approach to deriving an approximate eigenvalue θ from the approximate eigenvector u is to impose the Galerkin condition $r(\theta, u) \perp u$. Then it follows that $\theta = \theta(u)$ must be one of the two solutions to the quadratic equation

$$\alpha\theta^2 + \beta\theta + \gamma = 0, \tag{2.1}$$

where $\alpha = \alpha(u) = u^*Au$, $\beta = \beta(u) = u^*Bu$, and $\gamma = \gamma(u) = u^*Cu$. An unnatural aspect is that if $u = x$, the second solution of (2.1) has in general no meaning. If u is

close to x, we will be able to decide which one is best by looking at the norms of the residuals. But if u is not very accurate, it may not be clear which solution is the best. For instance, this may happen when we try to solve (1.5) by a subspace method; in the beginning of the process, the search space may not contain good approximations to an eigenvector. This problem is also mentioned in [1, p. 282].

A second, related problem is the subject of the rest of this subsection. A nice property that an approximate eigenvalue can (or should) have is that it is close to the eigenvalue if the corresponding approximate eigenvector is close to the eigenvector. In other words, we like the situation where

$$|\theta(x+h) - \lambda| = |\theta(x+h) - \theta(x)| \quad \text{is small}$$

for small $\|h\|$. When θ is differentiable with respect to u in the point x this is equivalent to the condition

$$\left\| \frac{\partial \theta}{\partial u}(x) \right\| \quad \text{is small.} \tag{2.2}$$

The one-dimensional Galerkin approach (2.1) defines θ implicitly as a function of α, β, and γ, say $f(\theta, \alpha, \beta, \gamma) = 0$, with $f(\lambda, \alpha(x), \beta(x), \gamma(x)) = 0$. Define the discriminant δ by

$$\delta = \delta(u) := \beta^2 - 4\alpha\gamma. \tag{2.3}$$

When $\delta(x) \neq 0$, the implicit function theorem states that locally θ is a function of α, β, and γ, say $\theta = \varphi(\alpha, \beta, \gamma)$, and that

$$\begin{aligned} D\varphi(\alpha(x), \beta(x), \gamma(x)) &= -\left((D_\theta f)^{-1} D_{(\alpha,\beta,\gamma)} f \right) (\lambda, \alpha(x), \beta(x), \gamma(x)) \\ &= \pm \frac{1}{\sqrt{\delta(x)}} \cdot (\lambda^2, \lambda, 1). \end{aligned}$$

So when δ is small, which means that (2.1) has two roots that are close, then the solutions of (2.1) may be very sensitive to perturbations in u. (Note that in general, the coefficients α, β, and γ are not differentiable with respect to u.) Therefore, we may expect that $|\theta - \lambda|$ is large for small perturbations of x; see also the numerical experiments in section 4. Thus, the second solution of (2.1) not only is useless, but may also hinder the accuracy of the solution that is of interest!

We therefore examine alternative ways to approximate λ. We generalize the Galerkin property (1.2) and minimum residual property (1.3) for the quadratic eigenvalue problem in the following three subsections.

Remark 2.1. As in the standard eigenvalue problem, $\theta = \theta(u, v)$ as solution of

$$(v^*Au)\theta^2 + (v^*Bu)\theta + (v^*Cu) = 0$$

is stationary in the right/left eigenvector combinations (x_i, y_i). However, we use no information about the (approximate) left eigenvector in this paper.

2.2. Two-dimensional Galerkin. In the standard eigenvalue problem, we deal with two vectors u and Au, which are asymptotically (by which we mean when $u \to x$) dependent. Therefore it is natural to take the length of the projection of Au onto the span of u as an approximation to the eigenvalue, which is exactly what the Rayleigh

quotient $\rho(u)$ (see (1.1)) does. For the generalized eigenvalue problem we have a similar situation.

In the quadratic eigenvalue problem, however, we deal with three vectors Au, Bu, and Cu, which asymptotically lie in a plane. Therefore it is natural to consider the projection of these three vectors onto a certain plane, spanned by two independent vectors p and q. To generalize the approach of (1.2), define the *generalized residual* $r(\mu, \nu, u)$ by

$$r(\mu, \nu, u) := (\mu A + \nu B + C)u. \tag{2.4}$$

The idea behind this is that we want to impose conditions on r such that μ forms an approximation to λ^2, and ν forms an approximation to λ. Then both μ/ν and ν may be good approximations to the eigenvalue λ. A generalization of (1.2) is obtained by imposing two Galerkin conditions $r(\mu, \nu, u) \perp p$ and $r(\mu, \nu, u) \perp q$ for specific independent vectors p, q. This leads to the system

$$W^*Z \begin{bmatrix} \mu \\ \nu \end{bmatrix} = -W^*Cu, \quad \text{where} \quad W = [\, p \;\; q \,], \quad Z = [\, Au \;\; Bu \,]. \tag{2.5}$$

When W^*Z is nonsingular, (2.5) defines unique μ and ν. A logical choice for p and q is any linear combination of Au, Bu, and Cu. Specifically, one could take the "least-squares" plane such that

$$\|(I - \Pi)Au\|^2 + \|(I - \Pi)Bu\|^2 + \|(I - \Pi)Cu\|^2$$

is minimal, where Π is the orthogonal projection onto the plane. An advantage of this least-squares plane is that it takes the norm of the vectors Au, Bu, and Cu into account.

Let z be the normal in $\text{span}(Au, Bu, Cu)$ of the sought-after plane; then one may verify that $\|(I - \Pi)Au\|^2 = \|(z^*Au)z\|^2 = |z^*Au|^2$. If D denotes the $n \times 3$ matrix with Au, Bu, and Cu as its columns, then z is the vector of unit length such that $\|z^*D\|^2$ is minimal. So we conclude that z is the minimal left singular vector of D, and for p and q we can take the two "largest" left singular vectors. Another choice for p and q, as well as its meaning, are discussed in subsection 2.4.

This two-dimensional Galerkin method yields three approximations to the eigenvalue. Besides the possibilities μ/ν and ν already mentioned, we can determine a third approximation by solving for $\theta \in \mathbb{C}$ such that

$$\left\| \begin{bmatrix} \theta^2 \\ \theta \end{bmatrix} - \begin{bmatrix} \mu \\ \nu \end{bmatrix} \right\|^2 \tag{2.6}$$

is minimal. We will indicate this solution as the "argmin" solution. Differentiating (2.6) with respect to $\text{Re}(\theta)$ and $\text{Im}(\theta)$ gives two mixed equations of degree 3 in $\text{Re}(\theta)$ and $\text{Im}(\theta)$. We may try to solve these equations by modern algorithms for systems of polynomials; see, for instance, [9, 10]. Another approach to solving the two coupled cubic equations is to form an equation, the so-called *resultant*, in only $\text{Re}(\theta)$ or $\text{Im}(\theta)$ (see section 4). It appears that in this case, the degree of the resultant is (only) 5. Since we use this equation to find $\text{Re}(\theta)$ and $\text{Im}(\theta)$, only the real solutions are of interest. The approach via the resultant may be numerically somewhat less stable, but an advantage of it is that one can use widely available mathematical packages such as MAPLE, as is done for the experiments in section 4. We will summarize the

two-dimensional Galerkin method and the two-dimensional minimum residual method (to be discussed in subsection 2.4) in Algorithm 2.1.

We now come to the sensitivity of the approximations derived by the two-dimensional Galerkin method (2.5). As seen, we have three possible approximations to the eigenvalue: μ/ν, ν, and the "argmin" solution. Asymptotically, if we take $u = x$, then we may assume $W = Z = [\, Ax \;\; Bx \,]$. When we "freeze" $W = Z = [\, Ax \;\; Bx \,]$ and differentiate (2.5), it can be seen that the sensitivity of μ and ν is related to $\kappa(Z)$, the condition number of Z (cf. [2, section 5.3.7]). For comparison, we will give $\kappa(Z)$ in the experiments.

When μ and ν are differentiable with respect to u (as is the case for the quasi-hyperbolic quadratic eigenvalue problem; see below) we have that

$$\frac{\partial(\mu/\nu)}{\partial u}(x) = \frac{1}{\lambda} \cdot \frac{\partial \mu}{\partial u}(x) - \frac{\partial \nu}{\partial u}(x). \tag{2.7}$$

This suggests that μ/ν might give inaccurate approximations for small λ, which is confirmed by numerical experiments; see Experiment 4.1. In general, μ and ν are not differentiable with respect to u, but μ/ν still will be sensitive for small λ. The sensitivity of the "argmin" solution will depend on the coefficients of the two defining cubic equations; we omit a (difficult) analysis.

2.3. One-dimensional minimum residual. Another approach generalizes the minimum residual approach (1.3). We try to minimize the norm of the residual with respect to θ:

$$\min_{\theta \in \mathbb{C}} \|(\theta^2 A + \theta B + C)u\|^2. \tag{2.8}$$

For complex θ, differentiating (2.8) with respect to $\mathrm{Re}(\theta)$ and $\mathrm{Im}(\theta)$ gives two mixed equations of degree 3 in $\mathrm{Re}(\theta)$ and $\mathrm{Im}(\theta)$. As in the previous subsection, we may use available algorithms [9, 10] or form the resultant, which in this case has degree 9 (in only $\mathrm{Re}(\theta)$ or $\mathrm{Im}(\theta)$); see also section 4.

In the special case in which we know that λ is real, we would like to have a real approximation θ. Then differentiating (2.8) with respect to θ gives the cubic equation with real coefficients,

$$4\|Au\|^2\theta^3 + 6\,\mathrm{Re}((Au)^*Bu)\theta^2 + 2(\|Bu\|^2 + 2\,\mathrm{Re}((Cu)^*Au))\theta + 2\,\mathrm{Re}((Cu)^*Bu) = 0, \tag{2.9}$$

which may be solved analytically. For instance, this is the case for the important class of *quasi-hyperbolic quadratic eigenvalue problems*; see the following definition.

DEFINITION 2.2 (cf. [8, p. 257]). *A quadratic eigenvalue problem $Q(\lambda)x = 0$ is called* quasi hyperbolic *if A is Hermitian positive definite, B and C are Hermitian, and for all eigenvectors of $Q(\lambda)$ we have*

$$(x^*Bx)^2 > 4(x^*Ax)(x^*Cx).$$

It is easy to see that all eigenvalues of quasi-hyperbolic quadratic eigenvalue problems are real.

We would like to stress that the one-dimensional minimum residual approach may also suffer from the same difficulties as the one-dimensional Galerkin method. In some cases, there may be more than one solution to choose from, and the irrelevant solutions may affect the accuracy of the relevant solution (cf. the discussion in subsection 2.1). When there is more than one real solution, we take the one that minimizes the norm

of the residual (just as for the one-dimensional Galerkin method). However, in most of the numerical experiments of section 4, the resultant has a unique real solution, making it unnecessary to choose. Moreover, the one-dimensional minimum residual method often gives (much) better results than the one-dimensional Galerkin method; see section 4.

2.4. Two-dimensional minimum residual. Another idea is to minimize the norm of the generalized residual (2.4) with respect to μ, ν:

$$\operatorname*{argmin}_{(\mu,\nu)\in\mathbb{C}^2} \|(\mu A + \nu B + C)u\|. \tag{2.10}$$

To solve this, consider the corresponding overdetermined $n \times 2$ linear system

$$Z \begin{bmatrix} \mu \\ \nu \end{bmatrix} = -Cu,$$

with Z as in (2.5). By assumption, Au and Bu are independent, so μ and ν are uniquely determined by

$$\begin{bmatrix} \mu \\ \nu \end{bmatrix} := -Z^+Cu = -\left(Z^*Z\right)^{-1} Z^*Cu,$$

where Z^+ denotes the pseudoinverse of Z. We see that (2.10) is a special case of (2.5), namely, the case in which we choose $p = Au$ and $q = Bu$, so $W = Z$.

As in subsection 2.2, we can form the following three possible approximations to the eigenvalue from the computed μ and ν: μ/ν, ν, and $\operatorname{argmin} \|(\theta^2, \theta) - (\mu, \nu)\|^2$. From subsection 2.2, it follows that the approximations derived by the two-dimensional methods depend on the plane of projection. The plane of the two-dimensional Galerkin method is contained in $\operatorname{span}\{Au, Bu, Cu\}$ (see subsection 2.2), while the plane for the two-dimensional minimum residual method is $\operatorname{span}\{Au, Bu\}$. Since $\operatorname{span}\{Ax, Bx, Cx\} = \operatorname{span}\{Ax, Bx\}$, we conclude that when $u = x$, both two-dimensional methods yield the same approximations. As a consequence, the sensitivity of the approximations is also the same for both two-dimensional methods.

The two-dimensional methods are summarized in Algorithm 2.1.

Input: an approximate vector u
Output: three approximate eigenvalues

1. Perform one of the following two actions:
   (a) Compute the two "largest" left singular vectors p and q of $[\, Au \;\; Bu \;\; Cu \,]$ **(Galerkin)**,
   (b) or take $p = Au$ and $q = Bu$ **(minres)**.
2. Compute $(\mu, \nu) = -(W^*Z)^{-1}W^*Cu$, where $W = [\, p \;\; q \,]$ and $Z = [\, Au \;\; Bu \,]$.
3. Approximate λ by one of the following:
   (a) μ/ν
   (b) ν
   (c) $\operatorname{argmin}_\theta \|(\theta^2, \theta) - (\mu, \nu)\|^2$

ALGORITHM 2.1
The two-dimensional Galerkin and two-dimensional minimum residual methods.

The following remark explains why we assumed in section 1 that both of the pairs Ax and Bx and Au and Bu are independent.

Remark 2.3. When Au and Bu are dependent, then the one-dimensional minimum residual approach reduces to the one-dimensional Galerkin approach, while

the two-dimensional methods are not uniquely determined. When Ax and Bx are dependent, then, although the approaches may be uniquely determined, the results may be bad. (In this case $\sigma_{\min}([\, Ax \;\; Bx\,]) = 0$, so μ and ν are "infinitely sensitive"; see subsection 2.2.) For example, the matrix Z in the two-dimensional methods is ill-conditioned if u is a good approximation to x. Of course, for Au and Bu to be independent, we must have $n \geq 2$.

3. Extensions.

3.1. Approximations from subspaces. We can also use the techniques described in section 2 for approximations to eigenpairs from subspaces of dimension larger than one. Let $\mathcal{U}$ be a k-dimensional subspace, where for subspace methods one typically has $k \ll n$, and let the columns of U form a basis for $\mathcal{U}$. The *Ritz–Galerkin condition*

$$\theta^2 Au + \theta Bu + Cu \perp \mathcal{U}, \qquad u \in \mathcal{U},$$

leads, with the substitution $u = Us$, to the projected quadratic eigenvalue problem

$$(\theta^2 U^*AU + \theta U^*BU + U^*CU)s = 0, \tag{3.1}$$

which in general yields $2k$ Ritz pairs (θ, u). For a specific pair, one can, as a first step, "refine" the value θ by the methods of section 2. Although it is not guaranteed that the new $\widetilde{\theta}$ is better, that often seems to be the case; see the numerical experiments. Moreover, we can monitor the *backward error*.

DEFINITION 3.1 (cf. [7]). *The backward error of an approximate eigenpair (θ, u) of Q is defined as*

$$\begin{aligned}\eta(\theta, u) := \min\{\varepsilon : \quad & (\theta^2(A + \Delta A) + \theta(B + \Delta B) + (C + \Delta C))u = 0, \\ & \|\Delta A\| \leq \varepsilon\zeta_1,\ \|\Delta B\| \leq \varepsilon\zeta_2,\ \|\Delta C\| \leq \varepsilon\zeta_3\}.\end{aligned}$$

The backward error of an approximate eigenvalue θ of Q is defined as

$$\eta(\theta) := \min_{\|u\|=1} \eta(\theta, u).$$

In [7, Theorems 1 and 2], the following results are proved:

$$\eta(\theta, u) = \frac{\|r\|}{\zeta_1|\theta|^2 + \zeta_2 \cdot |\theta| + \zeta_3}, \qquad \eta(\theta) = \frac{\sigma_{\min}(Q(\theta))}{\zeta_1|\theta|^2 + \zeta_2 \cdot |\theta| + \zeta_3}. \tag{3.2}$$

In the numerical experiments we therefore examine the quality of the computed θ by examining $\|r\|$ and $\sigma_{\min}(Q(\theta))$, which, for convenience, are also called backward errors. Note that the backward errors are related: $\sigma_{\min}(Q(\theta)) \leq \|r\|$.

Then, as a second step after refining the θ, one can "refine" the vector u by taking $\widetilde{u} = U\widetilde{s}$, where

$$\widetilde{s} = \text{the "smallest" right singular vector of } \quad \widetilde{\theta}^2 AU + \widetilde{\theta} BU + CU.$$

(For the Arnoldi method for the standard eigenvalue problem, a similar refinement of a Ritz vector has been proposed in [4].) This step is relatively cheap, because all matrices are "skinny." Given $\widetilde{\theta}$, the vector $\widetilde{u}$ minimizes the backward error $\eta(\widetilde{\theta}, u)$; see (3.2). It is also possible to repeat these two steps to get better and better approximations, leading to Algorithm 3.1.

Input: a search space $\mathcal{U}$
Output: an approximate eigenpair (θ, u) with $u \in \mathcal{U}$

1. Compute an approximate eigenpair (θ, u) according to the standard Ritz–Galerkin method
   for $k = 1, 2, \ldots$.
2. Compute a new θ_k choosing one of the methods of section 2.
3. Compute the "smallest" singular vector s_k of $\theta_k^2 AU + \theta_k BU + CU$.
4. $u_k = U s_k$.

ALGORITHM 3.1
Refinement of an approximate eigenpair for the quadratic eigenproblem.

During this algorithm, we do not know the (forward) error $|\theta_k - \lambda|$, but the backward errors $\|r\|$ and $\sigma_{\min}(\theta_k^2 AU + \theta_k BU + CU)$ are cheaply available; they can be used to decide whether or not to continue the algorithm. When we take the one-dimensional minimum residual method in each step, we are certain that the backward error $\|r\|$ decreases monotonically. In Experiment 4.3 we use the two-dimensional Galerkin approach in every step.

Remark 3.2. For the symmetric eigenvalue problem, the possibility of an iterative procedure for minimizing $\|Au - \rho(u)u\|$ over the subspace $\mathcal{U}$ is mentioned in [6] in the context of finding inclusion intervals for eigenvalues. Moreover, a relation between the minimization of $\|Au - \rho(u)u\|$ and the smallest possible Lehmann interval is given.

3.2. The polynomial eigenvalue problem. Consider the polynomial eigenvalue problem

$$(\lambda^l A_l + \lambda^{l-1} A_{l-1} + \cdots + \lambda A_1 + A_0)x = 0,$$

where the size n of the matrices satisfies $n \geq l$. Define the *generalized residual* as

$$r(\mu_1, \ldots, \mu_l, u) := (\mu_l A_l + \mu_{l-1} A_{l-1} + \cdots + \mu_1 A_1 + A_0)u.$$

Both the l-dimensional Galerkin method

$$r(\mu_1, \ldots, \mu_l, u) \perp \{p_1, \ldots, p_l\}$$

and the l-dimensional minimum residual method

$$\min_{\mu_1, \ldots, \mu_l} \|r(\mu_1, \ldots, \mu_l, u)\|$$

lead to a system of the form

$$W^* Z \begin{bmatrix} \mu_l \\ \vdots \\ \mu_1 \end{bmatrix} = -W^* A_0 u, \tag{3.3}$$

where $Z = [A_l u \ \cdots \ A_1 u]$. For the l-dimensional minimum residual method we have $W = Z$; for the l-dimensional Galerkin approach with "least-squares" l-dimensional plane, W consists of the l largest left singular vectors of $[\, Z \ A_0 u \,]$. Assuming that the vectors $A_1 u, \ldots, A_l u$ are independent, we see that (3.3) has a unique solution. In principle we can try every quotient μ_l/μ_{l-1}, μ_{l-1}/μ_{l-2}, …, μ_2/μ_1, μ_1, and also some other combinations like $\mu_l/(\mu_{l-2}\mu_1)$, as an approximation to λ. When λ is small, μ_1 will probably be the best solution. In principle, an "argmin" solution is also possible, although the degree of the associated polynomials will become larger quickly.

The one-dimensional minimum residual approach may be less attractive for the polynomial eigenvalue problem, as the degree of the associated polynomials (cf. (2.9) and (4.1)) increases quickly. This results in more irrelevant solutions, while the relevant solution will likely be more sensitive to perturbations in the (approximate) eigenvector.

4. Numerical experiments. The experiments are carried out in MATLAB and MAPLE. First, we offer a word on solving (2.8) and (2.6) for the one-dimensional minimum residual approach, and the "argmin" solution of the two-dimensional Galerkin and minimum residual approach, respectively. Write $\theta = \theta_1 + i\theta_2$, $\mu = \mu_1 + i\mu_2$, and $\nu = \nu_1 + i\nu_2$. Differentiating (2.8) with respect to θ_1 and θ_2 leads to two mixed equations (in θ_1 and θ_2) of degree 3. With MAPLE the equations are manipulated so that we have two equations of degree 9 in only θ_1 or only θ_2, which are each called a resultant. When we know that λ is real, then we get the cubic equation (2.9).

Differentiation of (2.6) with respect to θ_1 and θ_2 leads to

$$\begin{aligned} \theta_1^3 + (\theta_2^2 - \mu_1 + \tfrac{1}{2})\theta_1 - \mu_2\theta_2 - \tfrac{1}{2}\nu_1 = 0, \\ \theta_2^3 + (\theta_1^2 + \mu_1 + \tfrac{1}{2})\theta_2 - \mu_2\theta_1 - \tfrac{1}{2}\nu_2 = 0. \end{aligned} \tag{4.1}$$

Because of the missing θ_1^2 and θ_2^2 terms, in the first and second equation, respectively, the corresponding resultants have a degree of only 5. All equations were solved numerically by a MAPLE command of the form

`solve(resultant(`*equation*$_1$`(x, y),`*equation*$_2$`(x, y), y), x).`

Of course, we only have to solve one resultant, say for $\mathrm{Re}(\theta)$, then $\mathrm{Im}(\theta)$ can be solved from a cubic equation. In our experiments, many equations have a unique real solution, making it unnecessary to choose. When there is more than one real solution, we take the one that minimizes the norm of the residual.

EXPERIMENT 4.1. Our first example is taken from [8, p. 250]:

$$A = \begin{bmatrix} 0 & 6 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & -6 & 0 \\ 2 & -7 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad C = I_3.$$

This problem has two eigenvectors with the property that there are two eigenvalues to each eigenvector: $[\,1\ 1\ 0\,]^T$ corresponds to $\lambda = 1/2$ and $\lambda = 1/3$, while $[\,0\ 0\ 1\,]^T$ corresponds to $\lambda = \pm i$. In line with our assumptions, we do not consider these. Instead, we focus on the other eigenpairs $(\lambda, x) = \left(1, [\,0\ 1\ 0\,]^T\right)$ and $(\lambda, x) = \left(\infty, [\,1\ 0\ 0\,]^T\right)$. For the latter pair we consider the problem for $\lambda^{-1} = 0$. We simulate the situation of having a good approximation $u \approx x$ by adding a random (complex) perturbation to x:

$$u := (x + \varepsilon \cdot w) \,/\, \|x + \varepsilon \cdot w\|, \tag{4.2}$$

where w is a normalized vector of the form `rand(3,1)` $+ i \cdot$ `rand(3,1)`. (For all experiments, we take "`seed=0`" so that our results are reproducible.) Table 4.2 gives the results of the four approaches for $\varepsilon = 0.01$. The first row of the two-dimensional Galerkin (Gal-2) and two-dimensional minimum residual (MR-2) approaches represents μ/ν, the second row gives ν, while the third row indicates the "argmin" solution as an approximate eigenvalue. For clarity, the meaning of the different rows is first summarized in Table 4.1.

TABLE 4.1
The rows of Tables 4.2 *to* 4.4*, with their meanings.*

Row no.	Label	Meaning
1	Gal-1	best approximation (of the two) of the one-dimensional Galerkin method
2	Gal-2	μ/ν approximation of the two-dimensional Galerkin method
3		ν approximation of the two-dimensional Galerkin method
4		"argmin" approximation of the two-dimensional Galerkin method
5	MR-1	best approximation of the one-dimensional minimum residual method
6	MR-2	μ/ν approximation of the one-dimensional minimum residual method
7		ν approximation of the one-dimensional minimum residual method
8		"argmin" approximation of the two-dimensional minimum residual method

TABLE 4.2
*The approximations of the one-dimensional Galerkin (Gal-*1*), two-dimensional Galerkin (Gal-*2*: μ/ν, ν, and* "argmin"*), one-dimensional minimum residual (MR-*1*), and two-dimensional minimum residual (MR-*2*: μ/ν, ν, and* "argmin"*) approaches for $\lambda = 1$ and $\lambda^{-1} = 0$. Columns* 3, 4, 5, 7, 8, *and* 9 *give the (forward) error $|\theta - \lambda|$, and $\|r\|$ and $\sigma_{\min}(Q(\theta))$ for the backward errors.*

Method	Appr. for $\lambda = 1$	Error	$\|r\|$	$\sigma_{\min}$	Appr. for $\lambda^{-1} = 0$	Error	$\|r\|$	$\sigma_{\min}$
Gal-1	$0.99936 - 0.00192i$	0.00202	0.0112	0.00143	$0.00118 - 0.03957i$	0.03958	0.0399	0.0280
Gal-2	$0.99948 - 0.00160i$	0.00168	0.0112	0.00119	$1.00009 - 0.00229i$	1.00009	2.8285	0.0016
	1.00004	0.00004	0.0181	0.00003	$-0.00069 - 0.01986i$	0.01987	0.0206	0.0141
	$0.99981 - 0.00064i$	0.00067	0.0143	0.00047	$0.00010 - 0.01988i$	0.01988	0.0206	0.0141
MR-1	$0.99942 - 0.00173i$	0.00182	0.0111	0.00128	$-0.00036 - 0.02384i$	0.02384	0.0186	0.0168
MR-2	$0.99946 - 0.00160i$	0.00169	0.0112	0.00119	$1.00005 - 0.00229i$	1.00006	2.8284	0.0016
	0.99987	0.00013	0.0178	0.00009	$-0.00069 - 0.01986i$	0.01987	0.0206	0.0141
	$0.99971 - 0.00064i$	0.00070	0.0142	0.00050	$0.00010 - 0.01988i$	0.01988	0.0206	0.0141

For $\lambda = 1$, all other approaches (Gal-2, MR-1, and MR-2) give a smaller (forward) error than the classical one-dimensional Galerkin method (Gal-1). The "ν" approximation of the two-dimensional approaches Gal-2 (row 3) and MR-2 (row 7) is particularly good. For the sensitivity of the ν-solution for the two-dimensional approaches, the modest value $\kappa([\, Ax \;\; Bx \,]) \approx 26$ already indicates this. Of all approximations, the MR-1 solution has the smallest backward error $\|r\|$, as expected, but not the smallest forward error. For the discriminant (2.3) we have $\delta = 25$.

For $\lambda^{-1} = 0$, the "μ/ν" approximations (rows 2 and 6) are bad, which was already predicted by (2.7); $\kappa([\, Ax \;\; Bx \,]) \approx 2.6$ is again modest, and for the discriminant we have $\delta = 1$.

EXPERIMENT 4.2. For the second example we construct matrices such that the discriminant δ is small, and hence the zeros of (2.1) are close. For small $\zeta > 0$ define

$$A = I_3, \qquad B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & -2 & 2 \\ 0 & 0 & 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & -1-\sqrt{\zeta} & 0 \\ 0 & 1-\zeta & 2 \\ 0 & 0 & 1 \end{bmatrix}.$$

One may check that $x = [\, 0 \;\; 1 \;\; 0 \,]^T$ is an eigenvector with corresponding eigenvalue $1 + \sqrt{\zeta}$. (The second solution $1 - \sqrt{\zeta}$ to (2.1) is close to the eigenvalue but has no meaning.) The discriminant is equal to 4ζ. We take $\zeta = 10^{-4}$, so $\lambda = 1.01$. We test the approaches for $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-3}$; see Table 4.3.

For the sensitivity of μ and ν of the two-dimensional methods Gal-2 and MR-2 we note that $\kappa([\, Ax \;\; Bx \,]) \approx 5.8$. Because the discriminant $\delta = 4 \cdot 10^{-4}$ is small, and the sensitivity is modest, it is no surprise that all other approximations are much better (measured in forward or backward error) than Gal-1.

EXPERIMENT 4.3. For the following example we take A, B, and C random

TABLE 4.3

The approximations of the one-dimensional Galerkin (Gal-1), two-dimensional Galerkin (Gal-2: μ/ν, ν, and "argmin"), one-dimensional minimum residual (MR-1), and two-dimensional minimum residual (MR-2: μ/ν, ν, and "argmin") approaches for $\lambda = 1.01$; for $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-3}$, respectively. Columns $3, 4, 5, 7, 8$, and 9 give the (forward) error $|\theta - \lambda|$, and $\|r\|$ and $\sigma_{\min}(Q(\theta))$ for the backward errors.

Method	Appr. ($\varepsilon = 10^{-2}$)	Error	$\|r\|$	$\sigma_{\min}$	Appr. ($\varepsilon = 10^{-3}$)	Error	$\|r\|$	$\sigma_{\min}$
Gal-1	$1.0317 - 0.1442i$	0.1459	0.1344	0.01327	$1.0117 - 0.0445i$	0.0445	0.0431	0.001326
Gal-2	$0.9861 - 0.0250i$	0.0346	0.0349	0.00052	$1.0076 - 0.0025i$	0.0035	0.0035	0.000037
	$1.0050 - 0.0143i$	0.0151	0.0274	0.00019	$1.0095 - 0.0015i$	0.0015	0.0028	0.000018
	$0.9974 - 0.0186i$	0.0225	0.0286	0.00027	$1.0088 - 0.0019i$	0.0022	0.0029	0.000026
MR-1	$1.0046 - 0.0150i$	0.0159	0.0274	0.00020	$1.0094 - 0.0014i$	0.0015	0.0027	0.000018
MR-2	$0.9857 - 0.0250i$	0.0348	0.0351	0.00053	$1.0076 - 0.0025i$	0.0035	0.0035	0.000037
	$1.0047 - 0.0143i$	0.0152	0.0274	0.00019	$1.0095 - 0.0015i$	0.0015	0.0028	0.000018
	$0.9971 - 0.0186i$	0.0226	0.0287	0.00027	$1.0088 - 0.0019i$	0.0023	0.0029	0.000026

TABLE 4.4

The approximations of the one-dimensional Galerkin (Gal-1), two-dimensional Galerkin (Gal-2: μ/ν, ν, and "argmin"), one-dimensional minimum residual (MR-1), and two-dimensional minimum residual (MR-2: μ/ν, ν, and "argmin") approaches for $\lambda \approx 7.2288 + 2.7803i$, and $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-4}$, respectively. Columns $3, 4, 5, 7, 8$, and 9 give the (forward) error $|\theta - \lambda|$, and $\|r\|$ and $\sigma_{\min}(Q(\theta))$ for the backward errors.

Method	Appr. ($\varepsilon = 10^{-3}$)	Error	$\|r\|$	$\sigma_{\min}$	Appr. ($\varepsilon = 10^{-4}$)	Error	$\|r\|$	$\sigma_{\min}$
Gal-1	$6.86+2.71i$	0.37	2.89	0.186	$7.218+2.739i$	0.0428	0.308	0.0221
Gal-2	$7.26+2.68i$	0.10	2.90	0.054	$7.231+2.769i$	0.0110	0.290	0.0057
	$6.87+3.04i$	0.44	3.16	0.234	$7.189+2.801i$	0.0446	0.330	0.0232
	$7.07+2.86i$	0.18	2.93	0.096	$7.210+2.785i$	0.0195	0.300	0.0101
MR-1	$7.04+2.61i$	0.24	2.81	0.123	$7.227+2.769i$	0.0107	0.290	0.0055
MR-2	$7.23+2.65i$	0.13	2.88	0.064	$7.231+2.769i$	0.0112	0.290	0.0058
	$3.67+1.63i$	3.74	6.34	0.709	$7.123+2.775i$	0.1057	0.437	0.0545
	$5.14+2.08i$	2.20	5.23	0.822	$7.177+2.772i$	0.0529	0.332	0.0247

symmetric matrices of size 100×100. We try to approximate the eigenvalue $\lambda \approx 7.2288 + 2.7803i$ for $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-4}$ in (4.2); see Table 4.4.

For the sensitivity for Gal-2 and MR-2 we have $\kappa([\, Ax \;\; Bx\,]) \approx 63$; $|\delta| \approx 2.4 \cdot 10^{-5}$. We see that the two "$\mu/\nu$" approximations (rows 2 and 6) are the best solutions, together with the MR-1 solution (row 5). Note that for larger matrices, the computation of $\sigma_{\min}(Q(\theta))$ is expensive. In practice, one does not compute it, but it is shown here to compare the methods.

EXPERIMENT 4.4. Next, we test Algorithm 3.1. We start with a three-dimensional subspace $\mathcal{U}$ consisting of the same vector as in the previous experiment ($\varepsilon = 10^{-3}$), completed by two random (independent) vectors. We determine six Ritz pairs according to (3.1) and refine the one with θ approximating the eigenvalue $\lambda \approx 7.2288 + 2.7803i$ by Algorithm 3.1, where in every step we choose the μ/ν-approximation of the two-dimensional Galerkin method. The results, shown in Table 4.5, reveal that both u and θ are improved four times, after which they remain fixed in the decimals shown. Note that the smallest possible angle of a vector in $\mathcal{U}$ with x is

$$\angle(\mathcal{U}, x) = \angle(x_U, x) \approx 6.2809 \cdot 10^{-4},$$

where $x_U = UU^*x/\|UU^*x\|$ is the eigenvector projected onto $\mathcal{U}$.

We see in particular that the first step of the algorithm considerably improves the approximate eigenpair. After four steps, the angle of the refined approximate

TABLE 4.5

Refinement of an approximate eigenvalue by Algorithm 3.1 for $\lambda \approx 7.2288 + 2.7803i$. The columns give the iteration number; angle between u and x; (forward) error $|\theta - \lambda|$; and $\|r\|$, $\tau_{\min} := \sigma_{\min}(\theta^2 AU + \theta BU + CU)$, and $\sigma_{\min}(Q(\theta)) = \sigma_{\min}(\theta^2 A + \theta B + C)$ for the backward errors.

Iteration	$\angle(u,x)$	θ	Error	$\|r\|$	$\tau_{\min}$	$\sigma_{\min}$
0	$7.192\ e-4$	$7.222+2.778i$	$6.112\ e-3$	$1.234\ e-1$	$1.166\ e-3$	$3.178\ e-3$
1	$6.542\ e-4$	$7.231+2.783i$	$4.113\ e-3$	$1.196\ e-1$	$1.137\ e-3$	$2.142\ e-3$
2	$6.529\ e-4$	$7.231+2.781i$	$2.627\ e-3$	$1.137\ e-1$	$1.137\ e-3$	$1.368\ e-3$
3	$6.528\ e-4$	$7.231+2.781i$	$2.597\ e-3$	$1.137\ e-1$	$1.137\ e-3$	$1.352\ e-3$
≥ 4	$6.528\ e-4$	$7.231+2.781i$	$2.596\ e-3$	$1.137\ e-1$	$1.137\ e-3$	$1.351\ e-3$

TABLE 4.6

Approximations of the one-dimensional Galerkin (Gal-1) and four-dimensional Galerkin (μ_4/μ_3, μ_3/μ_2, μ_2/μ_1, and μ_1) approaches for $\lambda \approx -2.2009 - 1.5366i$. The other columns give the (forward) error $|\theta - \lambda|$, and $\|r\|$ and $\sigma_{\min}(Q(\theta))$ for the backward errors.

Method	Approximation	Error	$\|r\|$	$\sigma_{\min}$
Gal-1	$-2.1891 - 1.5404i$	0.0123	0.0894	0.0388
μ_4/μ_3	$-2.2010 - 1.5370i$	0.0004	0.0171	0.0014
μ_3/μ_2	$-2.2057 - 1.5318i$	0.0067	0.0422	0.0215
μ_2/μ_1	$-2.2054 - 1.5238i$	0.0135	0.0863	0.0429
μ_1	$-2.1693 - 1.5346i$	0.0317	0.2121	0.0972

eigenvector with the optimal vector in $\mathcal{U}$ is less than 30% of the angle that the Ritz vector makes with the optimal vector. The error in θ is more than halved. Note again that $\sigma_{\min}(\theta^2 A + \theta B + C)$ is expensive, but $\tau_{\min} := \sigma_{\min}(\theta^2 AU + \theta BU + CU)$ is readily available in the algorithm.

EXPERIMENT 4.5. Finally, we test the ideas of subsection 3.2. Consider the polynomial eigenvalue problem of degree 4,

$$(\lambda^4 A_4 + \lambda^3 A_3 + \lambda^2 A_2 + \lambda A_1 + A_0)x = 0,$$

where the A_i are random 5×5 matrices. We try to approximate the eigenpair corresponding to $\lambda \approx -2.2009 - 1.5366i$ and take $\varepsilon = 10^{-4}$ in (4.2). The μ_i in the generalized residual

$$r(\mu_1, \mu_2, \mu_3, \mu_4, u) := (\mu_4 A_4 + \mu_3 A_3 + \mu_2 A_2 + \mu_1 A_1 + A_0)u$$

are determined by the four-dimensional Galerkin method with "least-squares" plane. The results of the μ_4/μ_3, μ_3/μ_2, μ_2/μ_1, and μ_1 approximations are summarized in Table 4.6.

Note that both the μ_4/μ_3 and μ_3/μ_2 approximations give better results than the standard approach. We mention that the results of the four-dimensional minimum residual method were roughly the same.

5. Conclusions. The usual one-dimensional Galerkin approach for the determination of an approximate eigenvalue corresponding to an approximate eigenvector may give inaccurate results, especially when the discriminant of (2.1) is small. We have proposed several alternative ways that often give better results with little extra effort (all methods require three matrix-vector multiplications Au, Bu, and Cu and, additionally, $\mathcal{O}(n)$ time). Based on our analysis and the numerical experiments, we recommend in particular the μ/ν and ν approximations of the two-dimensional approaches Gal-2 and MR-2 because they are cheap to compute and give good results.

For small eigenvalues, one should take the "ν" approximations. The MR-1 method ensures a minimum residual (backward error).

The approaches are also useful for approximations from a subspace and for polynomial eigenvalue problems of higher degree.

Acknowledgments. We thank Hans Duistermaat for discussions about (4.1) and the referees for very useful and insightful comments.

REFERENCES

[1] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000.

[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, London, 1996.

[3] M. E. Hochstenbach and G. L. G. Sleijpen, *Two-sided and alternating Jacobi–Davidson*, Linear Algebra Appl., 358 (2003), pp. 145–172.

[4] Z. Jia, *Refined iterative algorithms based on Arnoldi's process for large unsymmetric eigenproblems*, Linear Algebra Appl., 259 (1997), pp. 1–23.

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

[6] G. L. G. Sleijpen and J. van den Eshof, *On the use of harmonic Ritz pairs in approximating internal eigenpairs*, Linear Algebra Appl., 358 (2003), pp. 115–137.

[7] F. Tisseur, *Backward error and condition of polynomial eigenvalue problems*, Linear Algebra Appl., 309 (2000), pp. 339–361.

[8] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problem*, SIAM Rev., 43 (2001), pp. 235–286.

[9] J. Verschelde, *Algorithm* 795*: PHCPACK: A general-purpose solver for polynomial systems by homotopy continuation*, ACM Trans. Math. Software, 25 (1999), pp. 251–276.

[10] S. M. Wise, A. J. Sommese, and L. T. Watson, *Algorithm* 801*: POLSYS_PLP: A partitioned linear product homotopy code for solving polynomial systems of equations*, ACM Trans. Math. Software, 26 (2000), pp. 176–200.