SIAM J. MATRIX ANAL. APPL.
Vol. 26, No. 1, pp. 252–256

NORMWISE SCALING OF SECOND ORDER POLYNOMIAL MATRICES*

HUNG-YUAN FAN†, WEN-WEI LIN†, AND PAUL VAN DOOREN‡

Abstract. We propose a minimax scaling procedure for second order polynomial matrices that aims to minimize the backward errors incurred in solving a particular linearized generalized eigenvalue problem. We give numerical examples to illustrate that it can significantly improve the backward errors of the computed eigenvalue-eigenvector pairs.

Key words. generalized eigenvalues, QZ algorithm, balancing

AMS subject classifications. 15A18, 15A22, 65F15, 65F35

DOI. 10.1137/S0895479803434914

1. Introduction. The quadratic eigenvalue problem (QEP) is the calculation of the roots of the determinant of the polynomial matrix

$$P(\lambda) = \lambda^2 P_2 + \lambda P_1 + P_0, \tag{1.1}$$

where $P_2, P_1, P_0 \in \mathbb{C}^{n\times n}$. A recommended method to solve it, is to reduce it to a generalized eigenvalue problem (GEP), which is the calculation of the roots of the determinant of the following pencil:

$$\lambda B - A = \lambda \begin{bmatrix} I & 0 \\ 0 & P_2 \end{bmatrix} - \begin{bmatrix} 0 & I \\ -P_0 & -P_1 \end{bmatrix}. \tag{1.2}$$

Indeed, one easily verifies that

$$\det(P(\lambda)) \equiv \det(\lambda B - A).$$

But if the matrices P_i, $i = 0, 1, 2$, have norms

$$\gamma_2 := \|P_2\|_2, \quad \gamma_1 := \|P_1\|_2, \quad \gamma_0 := \|P_0\|_2$$

that differ a lot in order of magnitude, then it was shown in [3, Table 5.1] that the QZ algorithm applied to (1.2) may yield very poor backward errors in the coefficients of the polynomial matrix (1.1).

In this note we relate this to the scaling problem of the polynomial matrix (1.1) and we indicate that the computed eigenpairs of $P(\lambda)$ gain a lot in accuracy when using the QZ algorithm on an appropriate scaling of the pencil (1.2).

*Received by the editors September 17, 2003; accepted for publication (in revised form) by N. J. Higham February 9, 2004; published electronically September 14, 2004. This research was supported by the Belgian Programme on Inter-university Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture.
http://www.siam.org/journals/simax/26-1/43491.html

†Department of Mathematics, National Tsing Hua University, Hsinchu, 300 Taiwan (d887206@am.nthu.edu.tw, wwlin@am.nthu.edu.tw).

‡Department of Mathematical Engineering, Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium (vdooren@csam.ucl.ac.be).

2. Scaling of second order polynomial matrices. In section 3 of [2], the author considers the scaled QEP defined by

$$\hat{P}(\mu)x \equiv (\mu^2\hat{P}_2 + \mu\hat{P}_1 + \hat{P}_0)x = 0 \tag{2.1}$$

with $\mu = \lambda/\alpha$, $\hat{P}_2 = \alpha^2 P_2$, $\hat{P}_1 = \alpha P_1$ and $\hat{P}_0 = P_0$, where α is a scaling factor, and investigates the possibility of using this scaling of the QEP (1.1) to improve the backward error of the solution obtained via the GEP formulation (1.2). The paper [2] does not solve this scaling problem (the Conclusions section mentions it as an open problem) but instead it derives a sufficient condition to verify the backward stability for the QEP. Here we restate this theorem without a proof (see Theorem 7 in [2]).

THEOREM 2.1 (see [2]). *If $\|P_i\|_2 = 1$, $i = 0, 1, 2$, then solving the GEP (1.2) with a backward stable algorithm (e.g., the QZ algorithm) for the GEP is backward stable for the QEP: there exist perturbations Δ_i, $i = 0, 1, 2$ with norms of the order of the machine precision ϵ, such that $[\lambda^2(P_2 + \Delta_2) + \lambda(P_1 + \Delta_1) + (P_0 + \Delta_0)]\xi = 0$ for every computed eigenpair (λ, ξ).*

Remark. Theorem 2.1 is similar to a result given in an earlier paper [4], for a pencil (1.1) with $\|A\|_2 \approx \|B\|_2 \approx 1$. It is shown there that for any perturbations $\|\delta A\|_2 \approx \|\delta B\|_2 \approx \epsilon$ (e.g., the backward errors resulting from the QZ algorithm) there exist transformations $S := I + E$ and $T := I + F$ such that

$$S[\lambda(A + \delta A) - (B + \delta B)]T = \lambda \begin{bmatrix} I & 0 \\ 0 & (P_2 + \Delta_2) \end{bmatrix} - \begin{bmatrix} 0 & I \\ -(P_0 + \Delta_0) & -(P_1 + \Delta_1) \end{bmatrix},$$

where $\Delta_0, \Delta_1, \Delta_2, E$, and F have norms of the order of the machine precision. □

The above results suggest that a good scaling strategy for the QEP (1.1) is to scale P_2, P_1, and P_0 so that their 2-norms are all close to 1. Consider modifying the polynomial matrix $P(\lambda) = \lambda^2 P_2 + \lambda P_1 + P_0$ as follows:

$$\mu\alpha = \lambda; \quad \tilde{P}(\mu) \equiv P(\lambda)\beta = \mu^2\left(P_2\alpha^2\beta\right) + \mu\left(P_1\alpha\beta\right) + \left(P_0\beta\right) \tag{2.2}$$

which yields a corresponding matrix pencil

$$\mu\tilde{B} - \tilde{A} = \mu \begin{bmatrix} I & 0 \\ 0 & \tilde{P}_2 \end{bmatrix} - \begin{bmatrix} 0 & I \\ -\tilde{P}_0 & -\tilde{P}_1 \end{bmatrix} \tag{2.3}$$

with coefficient matrices $\tilde{P}_2 = P_2\alpha^2\beta$, $\tilde{P}_1 = P_1\alpha\beta$, $\tilde{P}_0 = P_0\beta$ of respective 2-norms $\tilde{\gamma}_2 = \gamma_2\alpha^2\beta$, $\tilde{\gamma}_1 = \gamma_1\alpha\beta$, $\tilde{\gamma}_0 = \gamma_0\beta$. One should thus try to minimize the maximum distance

$$\min_{\alpha,\beta} \max\left\{|\beta\alpha^2\gamma_2 - 1|,\ |\beta\alpha\gamma_1 - 1|,\ |\beta\gamma_0 - 1|\right\}. \tag{2.4}$$

If we substitute $\hat{\alpha} := \alpha\sqrt{\gamma_2/\gamma_0}$, $\hat{\beta} := \beta\gamma_0$, and $\hat{\gamma} := \gamma_1/\sqrt{\gamma_2\gamma_0}$, then this reduces to

$$\min_{\hat{\alpha},\hat{\beta}} \max\left\{|\hat{\beta}\hat{\alpha}^2 - 1|,\ |\hat{\beta}\hat{\alpha}\hat{\gamma} - 1|,\ |\hat{\beta} - 1|\right\}.$$

At the optimum, all three quantities will be equal since otherwise we can decrease the maximum by adapting $\hat{\beta}$ and $\hat{\alpha}$. Hence we must have

$$|\hat{\beta}\hat{\alpha}^2 - 1| = |\hat{\beta}\hat{\alpha}\hat{\gamma} - 1| = |\hat{\beta} - 1|.$$

Since at least two of the quantities inside $|\cdot|$ must also have equal signs, one of the following three relations must hold at the optimum:

$$\hat{\alpha}^2 = 1, \quad \text{or} \quad \hat{\alpha}\hat{\gamma} = 1, \quad \text{or} \quad \hat{\alpha} = \hat{\gamma}.$$

By mere comparison, one then finds that the optimum $\hat{\alpha}^*$ is given by the first choice, which finally yields

$$\hat{\alpha}^* = 1, \quad \hat{\beta}^* = 2/(1+\hat{\gamma}).$$

In terms of the original variables we thus have

$$\alpha^* = \sqrt{\gamma_0/\gamma_2}, \quad \beta^* = 2/(\gamma_0 + \gamma_1\sqrt{\gamma_0/\gamma_2})$$

and the new values for the scaled norms are

$$\tilde{\gamma}_0 = \tilde{\gamma}_2 = 2/(1+\hat{\gamma}), \quad \tilde{\gamma}_1 = 2\hat{\gamma}/(1+\hat{\gamma}),$$

while

$$\max\{|\tilde{\gamma}_2 - 1|, |\tilde{\gamma}_1 - 1|, |\tilde{\gamma}_0 - 1|\} = |(1-\hat{\gamma})/(1+\hat{\gamma})|.$$

We point out that bounding (2.4) also implies bounding the normwise backward error of the matrices $\tilde{P}_i$. Indeed, one easily checks that $\|\tilde{A}\|_2 \leq 2$ and $\|\tilde{B}\|_2 \leq \sqrt{5}$. When running the QZ algorithm on $\mu\tilde{B} - \tilde{A}$ we will have—according to the above remark—equivalent absolute backward errors $\tilde{\Delta}_i$, $i = 0, 1, 2$ with norms of the order of the machine precision ϵ. The *structured relative backward errors* will therefore be of the order of

$$\|\tilde{\Delta}_0\|_2/\|\tilde{P}_0\|_2 \approx \|\tilde{\Delta}_2\|_2/\|\tilde{P}_2\|_2 \approx \epsilon(1+\hat{\gamma}), \quad \|\tilde{\Delta}_1\|_2/\|\tilde{P}_1\|_2 \approx \epsilon(1+\hat{\gamma})/\hat{\gamma}, \tag{2.5}$$

and $\max\left\{1+\hat{\gamma}, 1+\hat{\gamma}^{-1}\right\}$ can thus be seen as a *growth factor* between unstructured relative backward errors on the pencil $\mu\hat{B} - \hat{A}$ and structured relative backward errors on the second order polynomial matrix $\hat{P}(\mu) = \mu^2\hat{P}_2 + \mu\hat{P}_1 + \hat{P}_0$. In the numerical examples section we indeed show that the backward error of an approximate eigenpair (ξ, λ) computed with this optimal scaling strategy improves a lot. Moreover, if $\hat{\gamma} = 1$ (this is, when $\gamma_1^2 = \gamma_0\gamma_2$) then the normwise backward error will be of the order of the machine precision according to Theorem 2.1.

Remark. One could consider a more general type of scaling

$$\mu\tilde{B} - \tilde{A} = \begin{bmatrix} \ell_1 I & 0 \\ 0 & \ell_2 I \end{bmatrix} (\alpha\mu B - A) \begin{bmatrix} r_1 I & 0 \\ 0 & r_2 I \end{bmatrix} \tag{2.6}$$

involving 5 parameters, ℓ_1, ℓ_2, r_1, r_2, and α, but this is in fact the same problem. Dividing ℓ_1, ℓ_2 and multiplying r_1, r_2 by a common factor yields the same solution, so we can choose $r_1 = 1$. Moreover, setting one block norm equal to 1 in both $\tilde{B}$ and $\tilde{A}$ does not modify relative block norms in each individual block, so we can set $\alpha\ell_1 r_1 = 1$ and $\ell_1 r_2 = 1$. This then yields the parametrization $\ell_1 = 1/\alpha$, $\ell_2 = \beta$, $r_1 = 1$, $r_2 = \alpha$, which is exactly the problem we studied above.

We point out that in [1] the more general problem of optimal scaling of companion pencils is considered, but the technique and results are quite different. One could also consider other GEPs with the same generalized eigenvalues as (1.2) (see [3]), but the proposed scaling would then probably have to be adapted.

3. Numerical examples. When applying the QZ algorithm to $\mu\tilde{B} - \tilde{A}$, each *computed eigenpair* μ, ξ satisfies $(\mu\tilde{B} - \tilde{A})\xi \approx 0$. Both subvectors $\xi_1 := \xi(1:n)$, $\xi_2 := \xi(n+1:2n)$ should be proportional to each other and will yield $P(\lambda)\xi_i \approx 0$, where $\lambda = \mu\alpha$. The normwise backward errors $\Delta_i, i = 0, 1, 2$, that are *compatible* with the computed eigenpair

$$[\lambda^2(P_2 + \Delta_2) + \lambda(P_1 + \Delta_1) + (P_0 + \Delta_0)]\xi_j = 0$$

can be bounded using the residuals $P(\lambda)\xi_j$. In [3] it is shown that the smallest normwise backward error satisfies

$$\max_{i=0,1,2} \|\Delta_i\|_2/\|P_i\|_2 = \eta(\xi_j, \lambda) \equiv \frac{\|P(\lambda)\xi_j\|}{(|\lambda|^2\|P_2\| + |\lambda|\|P_1\| + \|P_0\|)\|\xi_j\|} \quad \text{for } j = 1, 2.$$

In the following examples, we use these quantities as measure for the backward error for each eigenpair computed by the QZ algorithm. The quantities $\eta_s(\xi_j, \lambda)$, $j = 1, 2$, on the other hand, refer to the computed eigenvector/eigenvalue pairs obtained *after scaling*. All computations were performed using MATLAB/Version 6.0 on a Compaq/DS20 workstation. The machine precision is 1.1×10^{-16}.

Example 1. We first consider the nuclear power plant problem in [3]. The backward errors of the computed eigenpairs corresponding to the smallest and largest eigenvalues in modulus, and the corresponding *scaled* backward errors are shown in Table 3.1. In this example, the 2-norms of the matrices P_2, P_1, P_0 are of the order of 10^8, 10^{10}, and 10^{13}, respectively. After applying the optimal scaling presented in section 2, their 2-norms are reduced to $\tilde{\gamma}_0 = \tilde{\gamma}_2 \approx 1.18$, $\tilde{\gamma}_1 \approx 0.821$, respectively. For this example $\hat{\gamma} = 0.697$, which implies that the scaled backward errors should be of the order of the machine precision.

TABLE 3.1
Backward errors for Example 1.

$\lvert\lambda\rvert$	$\eta(\xi_1, \lambda)$	$\eta(\xi_2, \lambda)$	$\eta_s(\xi_1, \lambda)$	$\eta_s(\xi_2, \lambda)$
17.7	3e-5	6e-8	3e-15	1e-16
361	2e-11	2e-11	1e-18	2e-18

Example 2. Here we tested randomly generated second order polynomial matrices $P(\lambda)$ with $\|P_2\|_2 = O(10^5)$, $\|P_1\|_2 = O(10^3)$, $\|P_0\|_2 = O(10^{-3})$, and $n = 10$, respectively. The absolute values of computed eigenvalues range between $O(10^{-2})$ and $O(10^{-7})$ and in Table 3.2 we give the backward errors of the 5 eigenpairs of smallest modulus, computed without and with scaling. With the optimal scaling, the 2-norms of the scaled coefficient matrices $\tilde{P}_2$, $\tilde{P}_1$, and $\tilde{P}_0$ are reduced to $\tilde{\gamma}_0 = \tilde{\gamma}_2 \approx 2.13 \times 10^{-2}$, $\tilde{\gamma}_1 \approx 1.98$, respectively. For this example $\hat{\gamma} = 93.01$, which means that after scaling we should not lose more than one or two digits of accuracy, which is confirmed in the experiments.

Example 3. In this example we tested randomly generated second order polynomial matrices $P(\lambda)$ with $\|P_2\|_2 \approx 5.54 \times 10^{-5}$, $\|P_1\|_2 \approx 4.73 \times 10^3$, $\|P_0\|_2 \approx 6.01 \times 10^{-3}$, and $n = 10$, respectively. The absolute values of computed eigenvalues range between $O(10^{-7})$ and $O(10^8)$. In Table 3.3 we give the backward errors of the 5 eigenpairs of smallest modulus without and with scaling. The scaled 2-norms are reduced to $\tilde{\gamma}_0 = \tilde{\gamma}_2 \approx 2.44 \times 10^{-7}$, $\tilde{\gamma}_1 \approx 2.00$, respectively, and $\hat{\gamma} \approx 8.19 \times 10^6$. This implies that after scaling we should not lose more than six digits of accuracy.

Example 4. Here we also tested randomly generated second order polynomial matrices $P(\lambda)$ in (1.1) with $\|P_2\|_2 \approx 5.03 \times 10^5$, $\|P_1\|_2 \approx 6.53 \times 10^{-3}$, $\|P_0\|_2 \approx$

TABLE 3.2
Backward errors for Example 2.

$\|\lambda\|$	$\eta(\xi_1,\lambda)$	$\eta(\xi_2,\lambda)$	$\eta_s(\xi_1,\lambda)$	$\eta_s(\xi_2,\lambda)$
2.40e-7	5e-8	4e-7	5e-16	3e-15
4.04e-7	6e-8	3e-7	1e-15	3e-15
6.47e-7	3e-8	8e-8	4e-16	2e-15
6.70e-7	2e-8	6e-8	9e-16	3e-15
1.22e-6	5e-9	7e-9	3e-16	2e-15

TABLE 3.3
Backward errors for Example 3.

$\|\lambda\|$	$\eta(\xi_1,\lambda)$	$\eta(\xi_2,\lambda)$	$\eta_s(\xi_1,\lambda)$	$\eta_s(\xi_2,\lambda)$
2.09e-7	2e-7	1e-6	6e-11	2e-10
5.71e-7	2e-7	5e-7	2e-10	2e-10
7.44e-7	2e-7	6e-7	3e-11	3e-11
1.37e-6	2e-7	1e-7	3e-11	2e-11
1.62e-6	2e-7	1e-7	7e-12	5e-12

6.06×10^3, and $n = 10$, respectively. The absolute values of computed eigenvalues range between $O(10^{-2})$ and $O(10^{-1})$. In Table 3.4 we give the backward errors of the 5 eigenpairs of smallest modulus without and with scaling. The scaled 2-norms are now $\tilde{\gamma}_0 = \tilde{\gamma}_2 \approx 2.00$, $\tilde{\gamma}_1 \approx 2.37 \times 10^{-7}$, respectively, and the backward errors of the computed eigenpairs are reported in Table 3.4. In this case, $\hat{\gamma} \approx 1.18 \times 10^{-7}$ which means that after scaling we should not lose more than six digits of accuracy.

TABLE 3.4
Backward errors for Example 4.

$\|\lambda\|$	$\eta(\xi_1,\lambda)$	$\eta(\xi_2,\lambda)$	$\eta_s(\xi_1,\lambda)$	$\eta_s(\xi_2,\lambda)$
1.72e-2	2e-13	1e-11	7e-16	3e-15
7.21e-2	1e-12	6e-12	5e-16	6e-16
1.06e-1	1e-12	5e-12	6e-16	6e-16
1.13e-1	1e-12	2e-12	3e-16	5e-16
1.55e-1	2e-12	2e-12	5e-16	6e-16

As shown in all of these examples, the backward errors are all significantly improved by the scaling: we gain up to 10 digits of accuracy! Also the computable quantity $\hat{\gamma}$ gives an upper bound on the backward error which is often a good estimate as well, except for the last example where the accuracy is much better than predicted.

REFERENCES

[1] D. LEMONNIER AND P. VAN DOOREN, *Optimal scaling of companion pencils for the QZ algorithm*, Proceedings SIAM Applied Linear Algebra Conference, Williamsburg, July 2003, http://www.siam.org/meetings/la03/proceedings/lemonnid.pdf.

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

[3] F. TISSEUR AND K. MEERBERGEN, *The quadratic eigenvalue problem*, SIAM Rev., 43 (2001), pp. 235–286.

[4] P. VAN DOOREN AND P. DEWILDE, *The eigenstructure of an arbitrary polynomial matrix: Computational aspects*, Linear Algebra Appl., 50 (1983), pp. 545–579.