SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 279–316, January 1993

FORWARD INSTABILITY OF TRIDIAGONAL QR[*]

BERESFORD N. PARLETT[†] AND JIAN LE[‡]

Abstract. The QR algorithm is the standard method for finding all the eigenvalues of a symmetric tridiagonal matrix. It produces a sequence of similar tridiagonals. It is well known that the QR transformation from T to $\hat{T}$ is backward stable. That means that the computed $\hat{T}$ is exactly orthogonally similar to a matrix close to T. It is also known that sometimes the computed $\hat{T}$ is not close to the exact $\hat{T}$. This is caused by the occasional extreme sensitivity of $\hat{T}$ to changes in T or the shift, and will be referred to as forward instability of the (computed) QR algorithm.

For the purpose of computing eigenvalues the property of backward stability is all that is required. However, the QR transformation has other uses and then forward stability is needed.

This paper gives examples, analyzes the forward instability, and shows that it occurs only when the shift causes "premature deflation." It is shown that forward stability is governed by the size of the last entry in normalized eigenvectors of leading principal submatrices, and the extreme values of the derivative of each entry in $\hat{T}$ as a function of the shift are found.

Key words. QR transformation, tridiagonal, sensitivity

AMS(MOS) subject classification. 65F15

1. Summary.

The QR transformation is a complicated similarity transformation that depends on a parameter σ, called the shift, and that preserves the tridiagonal form of real symmetric matrices. A QR algorithm consists of a strategy for the choice of σ and the associated sequence of QR transformations. Wilkinson's shift strategy (see Example 2.4) produces a sequence of shifts that always converges to an eigenvalue. This algorithm is implemented in several routines in the EISPACK and LAPACK libraries. See [1], [4], [5], and [8].

Each QR transform of an $n \times n$ real symmetric tridiagonal matrix T may be computed by a sequence of $n-1$ similarity transformations using plane rotations. In [8], Wilkinson showed that any algorithm employing a limited sequence of orthogonal similarity transformations is backward stable in finite precision arithmetic. This means that the output matrix is exactly orthogonally similar to a perturbation of the initial matrix that is tiny in norm.

For the purpose of computing eigenvalues, backward stability is completely satisfactory. However, the QR transform has other uses as well. It occurs in inverse eigenvalue problems (see [2]) and as a deflation procedure in contexts such as the Lanczos algorithm. In these applications forward stability is needed. That means that the computed output should be close, in norm, to the output in exact arithmetic.

Unfortunately, the computed QR transformation sometimes exhibits violent forward instability. In this paper we show examples of this phenomenon and analyze it for real symmetric tridiagonal matrices. Definition of the QR transformation $\hat{T}$ of the pair T, σ is given at the beginning of §2.1 and the sequence of plane rotations is shown in (2.18). In the rest of this section we summarize our findings for the knowledgeable reader. We follow Householder conventions in notation with a few exceptions.

[*] Received by the editors August 1, 1988; accepted for publication (in revised form) September 5, 1991. This work was supported in part by Office of Naval Research contract N00014-90-J-1372.

[†] Department of Mathematics and Computer Science Division, Electrical Engineering and Computer Science Department, University of California, Berkeley, California 94720 (parlett@math.berkeley.edu).

[‡] EEG Systems Laboratory, San Francisco, California 94107.

1. Forward instability can appear in exact arithmetic. It is merely the extreme sensitivity of some entries of the transform $\hat{T}$ to small changes in T and/or σ. Roundoff error is not needed to provoke this phenomenon. As a thought experiment we can consider an unreduced tridiagonal T with an eigenvalue λ that is irrational. If σ is a very accurate rational approximation to λ, then it is possible for the exact QR transforms of T with respect to λ and with respect to σ, to differ, in certain entries, in all figures. For this reason we need only look at the sensitivity of $\hat{T}$ as a function of σ and ignore implementation details such as whether implicit or explicit shifting is used.

2. The QR transformation is defined for all square matrices. Its instability, when it occurs, is inherited from the sensitivity of the Q factor of a matrix ($B = QR$) and this sensitivity can be bounded by traditional perturbation theory. This was done in [7] and further extension and refinement of that work is in progress.

However, for real symmetric unreduced tridiagonal matrices T, the situation is simpler and we have been able to replace perturbation bounds by exact derivatives of the transform $\hat{T}$ with respect to the shift σ.

To explain our results we need the last entries in certain eigenvectors. Let T_k denote the leading $k \times k$ submatrix of T. Let $\lambda_i^{(k)}, i = 1, \ldots, k$ denote the spectrum of T_k and let ω_{ik} or $\omega_{i,k}$ (when the comma is needed) denote the magnitude of the last entry of the normalized eigenvector for $\lambda_i^{(k)}$. We chose the last letter of the Greek alphabet to remind us that these are the last entries.

Our main result is that forward instability occurs if and only if σ is very close to a $\lambda_i^{(k)}$ with a tiny ω_{ik}. More precisely, there are entries of $\hat{T}$ whose derivatives with respect to σ are $O(1/\omega_{ik})$ when σ is close to $\lambda_i^{(k)}$. This is the only way large values can occur in the derivatives. The details are in Theorem 4.3. We recall that the derivatives are not proportional to $\|T\|$. The effect of varying sizes in the off-diagonal entries β_i of T is hidden in the values of the $\{\omega_{ik}\}$.

One simple result (see proof of Theorem 4.3(i)) is that, for the last off-diagonal entry $\hat{\beta}_n$ of $\hat{T}$,

$$\frac{d}{d\sigma}\hat{\beta}_n(\sigma)\big|_{\sigma=\lambda_i^{(n)}} = (-1)^i \tan\theta_n,$$

where θ_n is the last rotation angle in the QR factorization of $T - \sigma I$ and

$$|\cos\theta_n| = \omega_{in}.$$

Figure 1 shows that $\hat{\beta}_n'$ is only interesting for σ very close to an eigenvalue.

3. Forward instability, if it occurs, is always preceded by what we call *premature deflation of* σ. As mentioned in the second paragraph, $\hat{T}$ may be computed by a sequence of plane rotations. Let $T^{(1)} = T$, $T^{(i)} = \Theta_i T^{(i-1)} \Theta_i^t$, $i = 2, \ldots, n$, $\hat{T} = T^{(n)}$. The intermediate matrices $T^{(k)}$ that occur in the transformation of $T^{(1)}$ into $T^{(n)}$ (see (2.18)) depart from tridiagonal form only in positions $(k-1, k+1)$ and $(k+1, k-1)$. If the entries $(k, k-1)$ and $(k, k+1)$ of $T^{(k)}$ are sufficiently small and the (k, k) entry equals σ to working accuracy, then row and column k could be deleted from $T^{(k)}$ with little change to the remaining eigenvalues. We say that σ has been deflated from T at minor step k instead of at the expected place $T^{(n)} = \hat{T}$. Most implementations do not look for this phenomenon and so do not see it. In such cases the computed version of $\hat{T}$ below the kth row bears little resemblance to the output in exact arithmetic. In particular, deflation at minor step n will not occur despite the fact that σ is an eigenvalue to working precision. This is discussed in §5 and shown in Example 2.1.

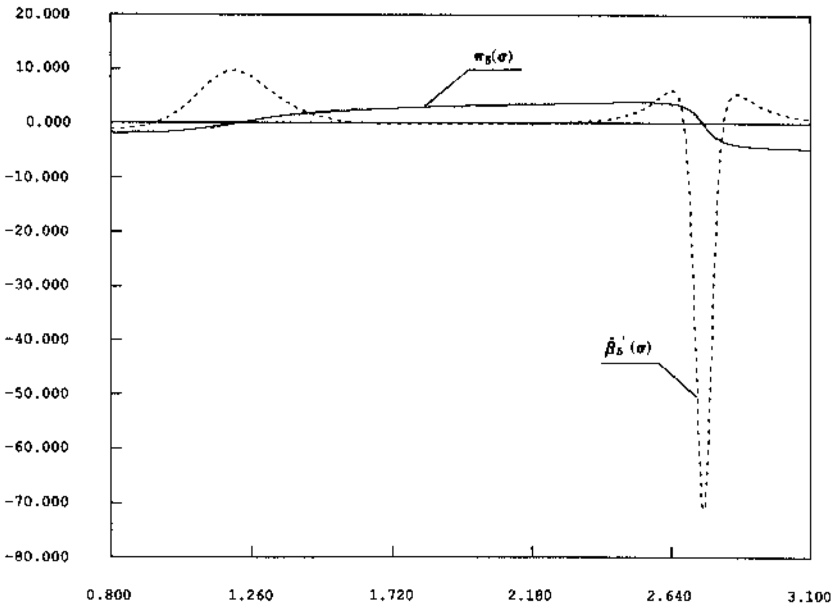

FIG. 1. *$\hat{\beta}_5'(\sigma)$ and $\pi_5(\sigma)$ on* [0.8, 3.1] *with $T = W_5^-$, defined in Example* 2.3.

4. A standard QR algorithm is somewhat protected from suffering from forward instability because of the preceding steps (transforms) in the algorithm. By the step at which σ is an eigenvalue, the last off-diagonal β_n is small and the eigenvector for σ usually has a large value for ω_{in}. This protection is not complete and Example 2.4 shows instability that occurs when Wilkinson's shift is used. However, our analysis shows that a shift strategy that picks an eigenvalue far from α_n is bound to incur forward instability when β_n is small. We are inclined to say that such a strategy has picked the "wrong" eigenvalue.

5. If $\beta_{k+1} = 0$, $k + 1 < n$, and if $\sigma = \lambda_i^{(k)}$ then the kth column of Q is not uniquely defined. Hence forward instability is inevitable close to this situation. This is when washout of the shifts occurs in the implicit implementation of QR (see [7]). This occurrence of local forward instability does not perturb the rest of the matrix when the explicit shift is used. We do not pursue this aspect further in this paper. From our point of view, when $\sigma = \lambda_i^{(k)}$ and $\beta_{k+1} = 0$, then the deflation that occurs at step k, though premature for T_n, is not premature for T_k.

6. Ultimate shifts. In [4] it was suggested that an efficient strategy for computing the spectral factorization of T would use two phases. First find the eigenvalues by any means. The QR algorithm without accumulation of the plane rotations is a leading candidate for the job. Second, run the QR algorithm with accumulation of the rotations, but using the eigenvalues as shifts. In this way the number of transformations to be accumulated is minimized. Our analysis in this paper shows that this strategy is likely to encounter *forward instability* and this might detract from the accuracy of the computed eigenvectors. However, if the monitoring algorithm described in §5 is used to terminate a QR transform at premature deflation, then this

danger may be greatly reduced. More investigation is needed.

7. Deflation in the Lanczos algorithm. A good way to compute a few eigenpairs for a large sparse symmetric matrix is to use the Lanczos algorithm; see, for example, [1] and [4]. The algorithm gradually builds up a tridiagonal matrix adding a row and a column at each step. After k steps some of the eigenvalues of the $k \times k$ tridiagonal T_k "settle down" and change very little as the algorithm proceeds. It would be convenient to deflate a converged eigenvalue from T_k. However, our analysis shows that this must be done at exactly the right step. After that, forward instability will occur. In such an application the monitoring scheme presented in §5 must be used. Indeed it was the failure to deflate eigenvalues of full accuracy that led us to this study of forward instability.

To a reader inclined to complain that we treat only instability due to a variation in the shift σ, we say two things. First, there is little loss in generality since the ω_{ik} are continuous functions of the entries of an unreduced T. A perturbed T, whatever the cause of the perturbation, will have a new set of $\lambda_i^{(k)}$ and ω_{ik}. Great sensitivity in transforming the new T will occur if σ is close to any of the new $\lambda_i^{(k)}$ with new ω_{ik} that are tiny. Second, we mention that this approach encompasses nonsymmetric Hessenberg matrices and arbitrary perturbations. Essentially, the condition for forward instability is the same. A tiny value of ω_{ik} signals that the first k columns of $T - \lambda_i^{(k)} I$ are almost linearly dependent. Note that if the first k columns of a matrix form a linearly dependent set, then the QR factorization process loses uniqueness at step k unless $k = n$.

Here is the plan of the paper. Section 2 presents notation and basic results on QR and §2.3 gives a set of examples of forward instability. Section 3 says more about the QR intermediate quantity π_k (shown in (2.1) and defined in (2.13)) than the reader will want to know. It is worth mentioning that all entries in $\hat{T}$ can be expressed in terms of the π_k and the entries in T. Not only do we give bounds on the value of π_k and its derivatives, but we give a simple model for it, and related functions, near each $\lambda_i^{(k)}$. The results are used in §4 and so §3 could be skipped by readers who have a little faith. Section 4 presents $\hat{T}'$, the derivative with respect to σ of the QR transform $\hat{T}$. Upper bounds are given first (Theorem 4.1) and show that huge values are only possible when σ is close to an eigenvalue of some principal submatrix T_k. Theorem 4.2 shows that the derivatives at the special points $\lambda_i^{(k)}$ can only be huge if σ is close to the spectrum of two successive principal submatrices. However, these results only give necessary conditions for instability. To show that small ω_{ik} are also sufficient we could see no other way than to prove Theorem 4.3. Again there is more detail than the reader may care for but the theorem is needed to establish the constants behind the O's and so to establish that values like $1/\omega_{ik}$ are attained. The functions $\hat{\beta}_k'$ can exhibit quite violent behavior close to some $\lambda_i^{(k)}$. For some of our examples the models are valid on intervals of width 10^{-14} and certain spectral points differ by 10^{-28}. A simple version of Theorem 4.3 is given in the preamble just before the detailed proof. All the results come from applying the linear model for π_k established in §3. Section 5 shows how ω_{ik} occurs in a lower bound on the premature deflation indicator. If neither β_{k+1} nor ω_{ik} is small then premature deflation cannot occur at that step.

Finally, we mention the figures whose contemplation, at the right moment, should help greatly in explaining the model and its application in Theorem 4.3. We chose a tame situation for Figs. 2 and 3 to illustrate typical situations and yet to avoid drastic rescaling of the y-axis.

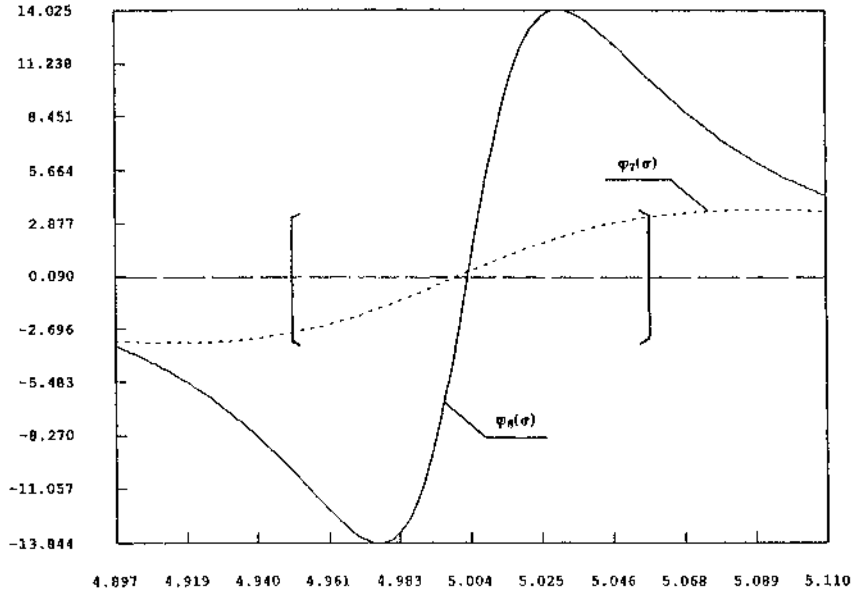

FIG. 2. φ_8 *and* φ_7 *on* $[\lambda_5^{(8)} - \omega_{5,8}m_8(5), \lambda_5^{(8)} + \omega_{5,8}m_8(5)]$ *with* $T = W_{17}^-$ *(see* §3.2).

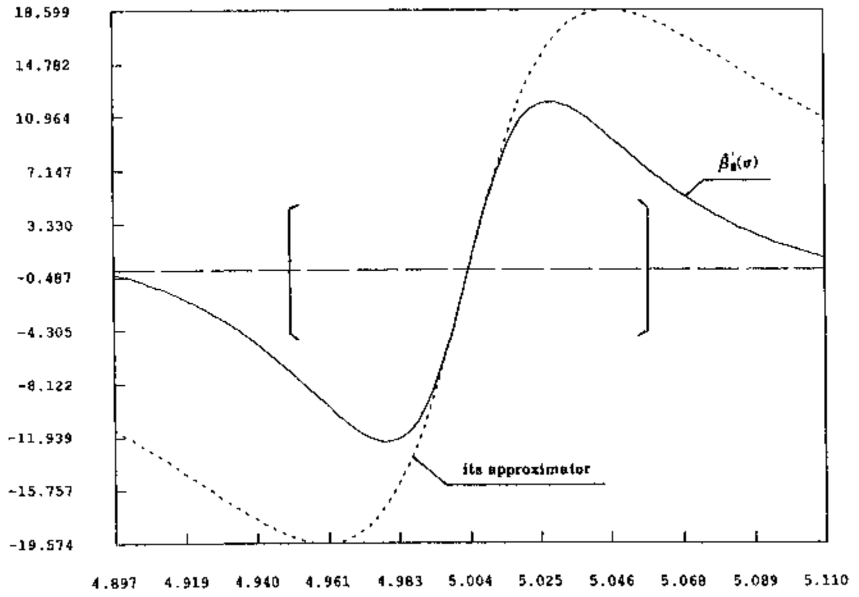

FIG. 3. β_8' *and its approximator defined in* (4.15) *on* $[\lambda_5^{(8)} - \omega_{5,8}m_8(5), \lambda_5^{(8)} + \omega_{5,8}m_8(5)]$ *with* $T = W_{17}^-$ *(see* §3.2).

2. QR basics and examples.

2.1. Relationships. For any real tridiagonal matrix T and any scalar σ (called the shift), the associated QR transformation $T \to \hat{T}$ is defined as follows:

$$T - \sigma I = QR, \qquad \hat{T} = RQ + \sigma I = Q^t T Q.$$

Let

$$T = T_n := \text{tridiag}\begin{pmatrix} & \beta_2 & & \beta_3 \cdots \beta_n & \\ \alpha_1 & & \alpha_2 & \cdots & \alpha_n \\ & \beta_2 & & \beta_3 \cdots \beta_n & \end{pmatrix}, \quad \beta_i > 0 \quad (i = 2, \ldots, n),$$

and let σ be the shift. T is said to be *unreduced* when $\beta_i \neq 0$, $i = 2, 3, \ldots, n$. Without loss of generality we may then assume that $\beta_i > 0$, $i = 2, 3, \ldots, n$. The QR factorization of $T - \sigma I$ may be represented by means of a sequence of plane rotations $\Theta_k, k = 2, \ldots, n$. Here Θ_k differs from the identity only in rows and columns $k-1$ and k where it has entries

$$\begin{pmatrix} c_k & s_k \\ -s_k & c_k \end{pmatrix},$$

and $c_k = \cos\theta_k, s_k = \sin\theta_k$. We let the dimension of Θ_k be given by its context; sometimes $k \times k$, sometimes $n \times n$. The angle θ_k is chosen to annihilate the entry in the $(k, k-1)$ position of the matrix $\Theta_{(k-1)}\Theta_{(k-2)} \cdots \Theta_2(T - \sigma I)$ and to lie in $[0, \pi]$. The result of this stage is written as

$$\Theta_k \cdots \Theta_2(T - \sigma I) = \begin{bmatrix} \xi_1 & \zeta_1 & s_2\beta_3 & & & & \\ & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \xi_{k-1} & \zeta_{k-1} & s_k\beta_{k+1} & \\ & & & & \pi_k & c_k\beta_{k+1} & \\ & & & & \beta_{k+1} & \alpha_{k+1} - \sigma & \cdot \\ & & & & & \cdot & \cdot \end{bmatrix}. \tag{2.1}$$

The first $(k-1)$ rows are in "final form" as entries of R, while the transitory but important row k will be overwritten at the next rotation. The last row of R is

$$(0, 0, \ldots, 0, \pi_n) \equiv \pi_n e_n^t,$$

where e_k denotes the kth column of the identity matrix and v^t is the transpose of v.

The Q factor is upper Hessenberg as well as orthogonal and is specified by the $(n-1)$ rotation angles $\theta_2, \ldots, \theta_n$. The detailed structure of Q is:

$$Q = \Theta_2^t \cdots \Theta_n^t = \begin{bmatrix} c_2 & (-s_2)c_3 & (-s_2)(-s_3)c_4 & \cdot & \cdot & (-s_2)\cdots(-s_n) \\ s_2 & c_2c_3 & c_2(-s_3)c_4 & \cdot & \cdot & \cdot \\ & s_3 & c_3c_4 & & \cdot & \cdot \\ & & s_4 & \cdot & \cdot & \cdot \\ & & & \cdot & c_{n-1}c_n & c_{n-1}(-s_n) \\ & & & & s_n & c_n \end{bmatrix}.$$

The sequence $\{c_2, c_3, \ldots, c_n\}$ effectively defines Q and the elements are often called its Schur parameters; see [2].

We allow the (n, n) entry of R, namely π_n, to have the same sign as $\det[T - \sigma I]$, and $\det[Q]$ equals one in our presentation. This is a trivial departure from the convention that the diagonal of R be nonnegative.

The last column of the $k \times k$ matrix $\Theta_2^t \cdots \Theta_k^t$ plays an important role and is given a name:

$$y_k = \begin{bmatrix} (-s_2)\cdots(-s_k) \\ c_2(-s_3)\cdots(-s_k) \\ \cdot \\ \cdot \\ c_{k-1}(-s_k) \\ c_k \end{bmatrix}. \tag{2.2}$$

Note that $y_n = Qe_n = q_n$. For $k < n$ we let $y_k^{(n)}$ denote a vector of the form

$$\begin{pmatrix} y_k \\ 0 \end{pmatrix} \in R^n.$$

The following relations are used frequently in the rest of the paper.

LEMMA 2.1. *With the notation developed above,*

$$Ty_k^{(n)} - y_k^{(n)}\sigma = \pi_k e_k + c_k\beta_{k+1}e_{k+1}, \qquad k < n, \tag{2.3}$$

$$\|Ty_k^{(n)} - y_k^{(n)}\sigma\|^2 = \pi_k^2 + c_k^2\beta_{k+1}^2, \qquad k < n, \tag{2.4}$$

$$\left(y_k^{(n)}\right)^t \left(Ty_k^{(n)} - y_k^{(n)}\sigma\right) = \pi_k c_k = \left(y_k^{(n)}\right)^t Ty_k^{(n)} - \sigma, \qquad k \le n, \tag{2.5}$$

$$(T - \sigma I)y_n^{(n)} = \pi_n e_n. \tag{2.6}$$

Proof. Equate the kth row on each side of (2.1) and transpose to get

$$(T - \sigma I)\Theta_2^t \cdots \Theta_k^t e_k = Ty_k^{(n)} - y_k^{(n)}\sigma = \pi_k e_k + c_k\beta_{k+1}e_{k+1}.$$

Since (2.1) holds for $k < n$, (2.3) is true for $k < n$. Equations (2.4) and (2.5) are the direct results of (2.3). Equation (2.6) is a special case of (2.3) since $\beta_{n+1} = 0$ when $k = n$. □

Equally important is the relation between $T_k \in R^{k\times k}$ and y_k.

LEMMA 2.2. *With the notation developed above,*

$$T_k y_k - y_k\sigma = \pi_k e_k, \qquad 1 \le k \le n, \tag{2.7}$$

$$y_k^t T_k y_k - \sigma = \pi_k c_k, \qquad 1 \le k \le n. \tag{2.8}$$

Proof. Equate the first k rows of (2.3) to get (2.7) for $k < n$. Equation (2.6) covers the case for (2.7) when $k = n$. Equation (2.8) is the direct result of (2.2). □

In (2.1) the relations between ξ_k, c_k, s_k, and π_k are $c_1 = 1$, $\pi_1 = \alpha_1 - \sigma$, and for $k = 2, \ldots, n$,

$$\xi_{k-1} = \sqrt{\pi_{k-1}^2 + \beta_k^2}, \tag{2.9}$$

$$c_k = \pi_{k-1}/\xi_{k-1}, \tag{2.10}$$

$$s_k = \beta_k/\xi_{k-1}, \tag{2.11}$$

$$\zeta_{k-1} = (c_k, s_k)\begin{pmatrix}\beta_k c_{k-1}\\ \alpha_k - \sigma\end{pmatrix} = c_{k-1}c_k\beta_k + s_k(\alpha_k - \sigma), \tag{2.12}$$

$$\pi_k = (-s_k, c_k)\begin{pmatrix}\beta_k c_{k-1}\\ \alpha_k - \sigma\end{pmatrix} = -s_k c_{k-1}\beta_k + c_k(\alpha_k - \sigma). \tag{2.13}$$

Finally, for $\hat{T} = RQ + \sigma I$, we have

$$\hat{\beta}_k = \xi_k s_k, \tag{2.14}$$

$$\hat{\alpha}_k = c_k\pi_k - c_{k+1}\pi_{k+1} + \alpha_{k+1}. \tag{2.15}$$

The last result comes from the invariance of the trace in $\Theta_k \cdots \Theta_2 T \Theta_2^t \cdots \Theta_k^t$ (see (2.17) below)

$$c_k\pi_k + \alpha_{k+1} - \sigma = \hat{\alpha}_k - \sigma + c_{k+1}\pi_{k+1}. \tag{2.16}$$

When $k = n$,

$$\hat{\beta}_n = \pi_n s_n, \qquad \hat{\alpha}_n = c_n\pi_n + \sigma.$$

For future reference we show the active part of the matrix $\Theta_k \cdots \Theta_2 T \Theta_k^t \cdots \Theta_k^t$ and the bulge in positions $(k+1, k-1)$ and $(k-1, k+1)$.

$$\begin{array}{ccccc}
\cdot & \hat{\beta}_{k-1} & & & \\
\hat{\beta}_{k-1} & \hat{\alpha}_{k-1} & \pi_k s_k & s_k\beta_{k+1} & \\
 & \pi_k s_k & c_k\pi_k + \sigma & c_k\beta_{k+1} & \\
 & s_k\beta_{k+1} & c_k\beta_{k+1} & \alpha_{k+1} & \beta_{k+2} \\
 & & & \beta_{k+2} & \cdot
\end{array} \tag{2.17}$$

In a study of the QR transform it is useful to exploit two different interpretations of $\hat{T} = \Theta_n \cdots \Theta_2 T \Theta_2^t \cdots \Theta_n^t$. First interpretation:

$$\begin{aligned} T \;&\rightarrow\; R = \Theta_n \cdots \Theta_2 T \\ &\rightarrow\; \hat{T} = R\Theta_2^t \cdots \Theta_n^t. \end{aligned}$$

Second interpretation:

$$\begin{aligned} T \;&\rightarrow\; T^{(2)} = \Theta_2 T \Theta_2^t \\ &\rightarrow\; T^{(3)} = \Theta_3 T^{(2)} \Theta_3^t \\ &\rightarrow\; T^{(n)} = \Theta_n T^{(n-1)} \Theta_n^t. \end{aligned} \tag{2.18}$$

Only T and $T^{(n)} = \hat{T}$ are tridiagonal matrices; the intermediate $T^{(k)}$ contain the bulge.

2.2. Deflation. The assumption that T is unreduced guarantees, in exact arithmetic, that the spectra of T_k and its submatrix T_{k-1} interlace strictly:

$$\lambda_1^{(k)} < \lambda_1^{(k-1)} < \lambda_2^{(k)} < \lambda_2^{(k-1)} < \cdots < \lambda_{k-2}^{(k-1)} < \lambda_{k-1}^{(k)} < \lambda_{k-1}^{(k-1)} < \lambda_k^{(k)}.$$

It will be shown in §3 that, as a function of σ, π_{k-1} vanishes at and only at the $\lambda_i^{(k-1)}, i = 1, \ldots, k-1$, while π_k vanishes at and only at the $\lambda_i^{(k)}, i = 1, \ldots, k$. Since

$$\xi_k = \sqrt{\pi_k^2 + \beta_{k+1}^2} \geq \beta_{k+1}, \qquad k < n,$$

it is apparent that the only entry of R that can ever vanish is $\xi_n = \pi_n$. Consequently,

$$\hat{\beta}_n = \pi_n s_n$$

vanishes at and only at $\sigma = \lambda_i^{(n)}, i = 1, \ldots, n$. To study $\hat{\beta}_n$ for σ in the neighborhood of $\lambda_i^{(n)}$, we consider its derivatives

$$\hat{\beta}'_n = \pi'_n s_n + \pi_n s'_n,$$

$$\hat{\beta}''_n = \pi''_n s_n + 2\pi'_n s'_n + \pi_n s''_n.$$

Clearly, it is necessary to study the properties of $\pi_k(\sigma)$ to understand $\hat{\beta}'_n$ and $\hat{\beta}''_n$, and this is done in §3.

2.3. Examples of forward instability. We computed the "exact" QR transform by using a very expensive method that works only for singular T as follows. The Schur parameters $\{c_i\}$ are reconstructed from the vector y_n of (2.2), which happens to be an eigenvector of T. These eigenvectors can be computed to high relative accuracy when the eigenvalue is simple, as in our examples.

Example 2.1.

$$(2.19) \qquad T = \begin{pmatrix} 6683.3333 & 14899.672 & & & & \\ 14899.672 & 33336.632 & 34.640987 & & & \\ & 34.640987 & 20.028014 & 11.832164 & & \\ & & 11.832164 & 20.001858 & 10.141851 & \\ & & & 10.141851 & 20.002287 & 7.5592896 \\ & & & & 7.5592896 & 20.002859 \end{pmatrix}.$$

Eigenvalues $\lambda_1 = 0$, $\lambda_2 = 10$, $\lambda_3 = 20$, $\lambda_4 = 30$, $\lambda_5 = 40$, $\lambda_6 = 40000$.

(i) Successful deflation with shift $\sigma = 0 = \lambda_1$.

$$\hat{T} = \begin{pmatrix} 39999.925 & 54.726511 & & & & \\ 54.726511 & 33.404823 & 8.3017268 & & & \\ & 8.3017268 & 24.730751 & 8.8065994 & & \\ & & 8.8065994 & 21.646903 & 7.2175779 & \\ & & & 7.2175779 & 20.292461 & -1.113d\text{–}14 \\ & & & & -1.113d\text{–}14 & 9.520d\text{–}13 \end{pmatrix}.$$

The computed version of $\hat{T}$ was indistinguishable from the exact to eight decimals, except that the nonzero entries in the last row are $(-7.943d\text{–}12,\ -2.344d\text{–}15)$.

(ii) Failed deflation with shift $\sigma = \lambda_6 = 40000$.

First, we present the true $\hat{T}$ rounded to eight decimals and then the computed version $\hat{T}^{(1)}$ using double precision on VAX 780.

$$\begin{aligned}
\hat{T} &= \begin{pmatrix}
19.989995 & 14.142133 & & & & \\
14.142133 & 20.003002 & 11.832160 & & & \\
& 11.832160 & 20.001858 & 10.141851 & & \\
& & 10.141851 & 20.002287 & 7.5592896 & \\
& & & 7.5592896 & 20.002859 & -1.608d{-}13 \\
& & & & -1.608d{-}13 & 40000.000
\end{pmatrix}, \\
\hat{T}^{(1)} &= \begin{pmatrix}
19.989995 & 14.142133 & & & & \\
14.142133 & 20.003002 & 11.832160 & & & \\
& 11.832160 & 20.001858 & 10.141851 & & \\
& & 10.141851 & 20.002287 & 7.5593584 & \\
& & & 7.5593584 & 20.730517 & -170.561 \\
& & & & -170.561 & 39999.272
\end{pmatrix}.
\end{aligned} \tag{2.20}$$

Insight is gained by looking at the intermediate matrices in the QR sweep. $T^{(4)}$ below shows the premature deflation mentioned in comment 3 in §1 and in (2.18).

Then

$$T^{(1)} = \text{matrix shown in (2.19).}$$

$$T^{(2)} = \begin{pmatrix}
19.989995 & 0.011185926 & 14.142128 & & & \\
0.011185926 & 39999.975 & -31.622748 & & & \\
14.142128 & -31.622748 & 20.028014 & 11.832164 & & \\
& & 11.832164 & 20.001858 & 10.141851 & \\
& & & 10.141851 & 20.002287 & 7.5592896 \\
& & & & 7.5592896 & 20.002859
\end{pmatrix},$$

$$T^{(3)} = \begin{pmatrix}
19.989995 & 14.142133 & & & & \\
14.142133 & 20.003001 & -2.76734d{-}6 & 11.832160 & & \\
& -2.76734d{-}6 & 40000.000 & 9.35882d{-}3 & & \\
& 11.832160 & 9.35882d{-}3 & 20.001858 & 10.141851 & \\
& & & 10.141851 & 20.002287 & 7.5592896 \\
& & & & 7.5592896 & 20.002859
\end{pmatrix},$$

$$T^{(4)} = \begin{pmatrix}
19.989995 & 14.142133 & & & & \\
14.142133 & 20.003001 & 11.832160 & & & \\
& 11.832160 & 20.001858 & -8.18031d{-}6 & 10.141851 & \\
& & -8.18031d{-}6 & 40000.000 & -2.37200d{-}6 & \\
& & 10.141851 & -2.37200d{-}6 & 20.002287 & 7.5592896 \\
& & & & 7.5592896 & 20.002859
\end{pmatrix},$$

$$T^{(5)} = \begin{pmatrix}
19.989995 & 14.142133 & & & & \\
14.142133 & 20.003001 & 11.832160 & & & \\
& 11.832160 & 20.001858 & 10.141851 & & \\
& & 10.141851 & 20.002287 & 0.032249815 & 7.5592896 \\
& & & 0.032249815 & 40000.000 & -6.09724d{-}6 \\
& & & 7.5592896 & -6.09724d{-}6 & 20.002859
\end{pmatrix},$$

$$T^{(6)} = \begin{pmatrix} 19.989995 & 14.142133 & & & & \\ 14.142133 & 20.003002 & 11.832160 & & & \\ & 11.832160 & 20.001858 & 10.141851 & & \\ & & 10.141851 & 20.002287 & 7.5593584 & \\ & & & 7.5593584 & 20.730517 & \mathbf{-170.56153} \\ & & & & \mathbf{-170.56153} & 39999.272 \end{pmatrix}.$$

(iii) A second QR sweep. Finally, we show $\hat{T}^{(2)}$, the QR transform of $\hat{T}^{(1)}$ in (2.20), using the same shift. This shows that, although $\hat{T}^{(2)}$ deflates nicely it is not very close to $\hat{T}$. Thus repeated application of the transform will *not* recover $\hat{T}$ if forward instability occurred.

$$\hat{T}^{(2)} = \begin{pmatrix} 19.979990 & 14.142125 & & & & \\ 14.142125 & 20.006003 & 11.832161 & & & \\ & 11.832161 & 20.003716 & 10.141851 & & \\ & & 10.141851 & 20.004574 & 7.5592897 & \\ & & & 7.5592897 & 20.005717 & 8.425d\text{–}15 \\ & & & & 8.425d\text{–}15 & 40000.000 \end{pmatrix}.$$

Example 2.2 (surprising successful deflation). This example uses a shift that is an exact eigenvalue not only of T but of the odd principal submatrices. Nevertheless, the computed QR transform is perfectly stable:

$$T = \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}.$$

Eigenvalues $\lambda_1 = -2 - \sqrt{3}$, $\lambda_2 = -3$, $\lambda_3 = -2$, $\lambda_4 = -1$, $\lambda_5 = -2 + \sqrt{3}$. With shift $\sigma = -2$,

$$\hat{T} = \begin{pmatrix} -2.0000000 & 1.4142136 & & & \\ 1.4142136 & -2.0000000 & 0.70710678 & & \\ & 0.70710678 & -2.0000000 & 1.2247449 & \\ & & 1.2247449 & -2.0000000 & 0.0000000 \\ & & & 0.0000000 & -2.0000000 \end{pmatrix}.$$

Example 2.3 (W_{21}^{-}; see [9]).

$$T = W_{21}^{-} := \text{tridiag}\begin{pmatrix} & 1 & 1 \cdots 1 & & 1 & \\ 10 & & 9 \quad \cdots & -9 & & -10 \\ & 1 & 1 \cdots 1 & & 1 & \end{pmatrix}.$$

Throughout this paper, matrices similar to W_{21}^{-} but of different sizes are used to best illustrate the properties of the underlying quantities. For simplicity, these matrices are denoted by W_n^{-} where $n = 2k - 1$, $\alpha_i = k - i$ for $i = 1, 2, \ldots, n$, and $\beta_i = 1$ for $i = 2, \ldots, n$.

This is the generic example of forward instability. W_{21}^{-} was constructed by Wilkinson for other purposes; see [9]. It has almost uniformly distributed eigenvalues. Rather than exhibit matrices of this size, we plot in Fig. 4 the exact and the computed Schur parameters of the Q matrix

$$Q = \Theta_2^t \Theta_3^t \cdots \Theta_{21}^t,$$

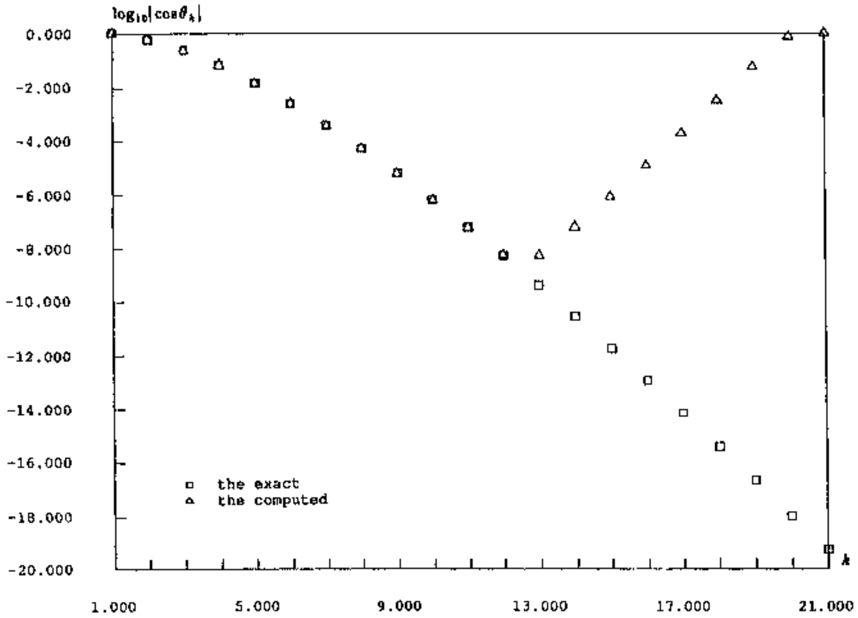

FIG. 4. *Schur parameters defining Q in* QR *transform on* W_{21}^{-} (*Example* 2.3).

obtained when the largest eigenvalue is the shift. The data in the graph is scaled from x to $\log_{10}|x|$ to show the divergence.

Example 2.4 (forward instability with Wilkinson's shift).

$$T = \text{tridiag}\begin{pmatrix} & 1 & 1 \cdots 1 & 1 & \\ 15 & 0 & \cdots & 0 & 15 \\ & 1 & 1 \cdots 1 & 1 & \end{pmatrix}.$$

Wilkinson's shift is the eigenvalue of

$$\begin{pmatrix} \alpha_{n-1} & \beta_n \\ \beta_n & \alpha_n \end{pmatrix}$$

that is closer to α_n.

This is another generic example of forward instability. The matrix has a double eigenvalue ($\lambda = 15.066666667$) in double precision. Now we run the QR algorithm on T in single precision with Wilkinson's shift strategy. At the first step Wilkinson's shift is 15.0664. After the first QR sweep, the last off-diagonal element is $-1.96302\mathbf{d}-05$. At the second step Wilkinson's shift is 15.0667. *Premature deflation* is observed at the seventh rotation since (see (2.17)) $\pi_7 * s_7 = -3.98401\mathbf{d}{-}04$ and $c_7 * \beta_8 = 4.43937\mathbf{d}{-}03$. To demonstrate forward instability, we repeat the second sweep with the same shift in double precision. The resulting $\hat{\alpha}_k$'s and $\hat{\beta}_k$'s from the second sweep in double precision and in single precision are displayed in Figs. 5 and 6, respectively. Again the data in the graphs is scaled from x to $\log_{10}|x|$ to show the divergence.

3. Properties of π_k.

The purpose of this section is to establish properties of π_k that are needed in §4, and consequently §3 may be skipped without loss of continuity.

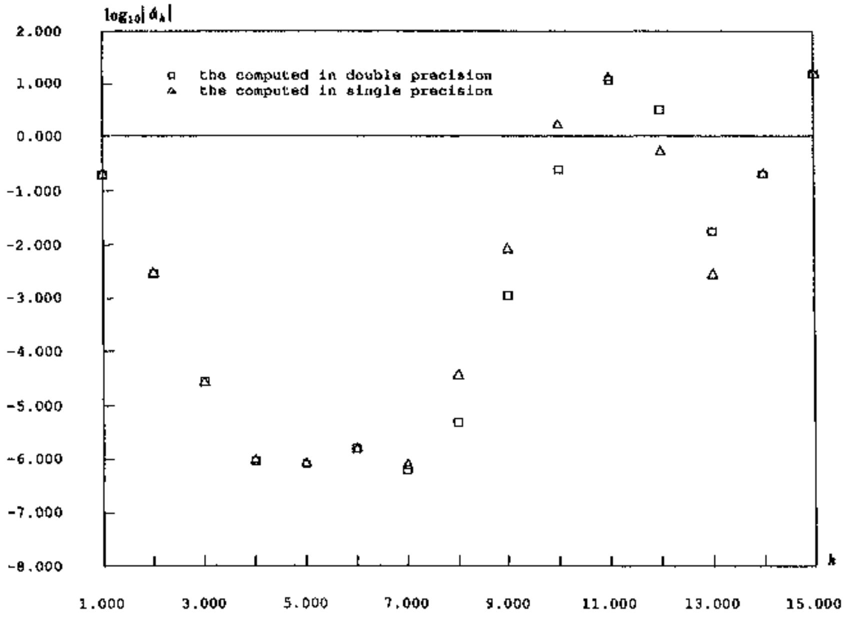

FIG. 5. *Resulting $\hat{\alpha}_k$'s from* QR *sweeps in double precision and in single precision* (*Example* 2.4).

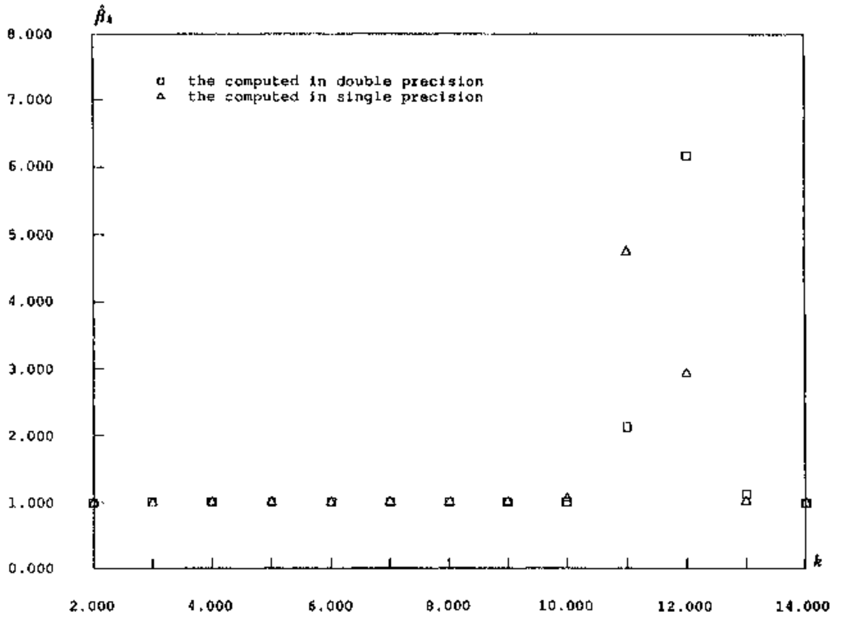

FIG. 6. *Resulting $\hat{\beta}_k$'s from* QR *sweeps in double precision and in single precision* (*Example* 2.4).

3.1. Derivatives of π_k. We discuss the smoothness of $\pi_k(\sigma)$, for $\sigma \in \mathbf{R}$.

LEMMA 3.1. *If T is unreduced (i.e., $\beta_i > 0$, $i = 2, \dots, n$), then π_k^2 is a rational function with k real double zeros and $2(k-1)$ simple poles that do not lie on the real axis. Thus π_k, c_k, and s_k are real analytic on $\mathbf{R}$. Moreover,*

$$\pi_k(\sigma) = 0 \quad \textit{if and only if } \sigma = \lambda_i^{(k)}, \qquad 1 \le i \le k, 1 \le k \le n. \tag{3.1}$$

The vector y_k of (2.2) is an eigenvector of T_k when $\sigma = \lambda_i^{(k)}$, and

$$c_k(\lambda_i^{(k)}) = (-1)^{i-1}\omega_{ik}, \quad i = 1, 2, \dots, k, \quad k = 1, 2, \dots, n. \tag{3.2}$$

Proof. Take the leading $k \times k$ submatrix on each side of (2.1) to find

$$\Theta_k \cdots \Theta_2 (T_k - \sigma I_k) = R_k$$

and

$$\det(T_k - \sigma I_k) = \xi_1 \cdots \xi_{k-1} \pi_k. \tag{3.3}$$

By (2.8), $\xi_k \ge \beta_{k+1}$ for all σ and $i < n$, and (3.1) follows directly from (3.3) and the assumption that T is unreduced.

Next consider the submatrix on each side of (2.1) in the first k rows and the first $k-1$ columns:

$$\Theta_k \cdots \Theta_2 \begin{bmatrix} T_{k-1} - \sigma I_{k-1} \\ e_{k-1}^t \beta_k \end{bmatrix} = \begin{bmatrix} \tilde{R}_{k-1} \\ o^t \end{bmatrix}.$$

Since $\Theta_k \cdots \Theta_2$ is orthogonal,

$$(T_{k-1} - \sigma I_{k-1})^2 + e_{k-1} e_{k-1}^t \beta_k^2 = \tilde{R}_{k-1}^t \tilde{R}_{k-1},$$

and, on taking determinants,

$$\det[(T_{k-1} - \sigma I_{k-1})^2 + e_{k-1} e_{k-1}^t \beta_k^2] = (\xi_1 \cdots \xi_{k-1})^2. \tag{3.4}$$

Divide $(3.3)^2$ by (3.4) to find

$$\pi_k^2 = \det{}^2[T_k - \sigma I_k] / \det[(T_{k-1} - \sigma I_{k-1})^2 + \beta_k^2 e_{k-1} e_{k-1}^t].$$

Note that with $\beta_k \ne 0$ the denominator can never vanish for real σ and consideration of the diagonal shows that it is a polynomial of degree $2(k-1)$. The numerator is the square of a polynomial of degree k and thus π_k^2 is a rational function in σ, as claimed.

Note that by (2.9), (2.10), and (2.11), ξ_{k-1}^2, c_k^2, and s_k^2 are rational since π_k^2 is rational and none have poles on the real axis.

Use (2.7) and (3.1) to find that y_k is an eigenvector of T_k when $\sigma = \lambda_i^{(k)}$. Then by definition of y_k in (2.2), find that $\omega_{ik} = |c_k(\lambda_i^{(k)})|$. Use (2.10) to see that c_k has the same sign as π_{k-1} does and $\operatorname{sign}[c_k(\lambda_i^{(k)})] = \operatorname{sign}[\pi_{k-1}(\lambda_i^{(k)})] = (-1)^{i-1}$, since from (3.3), $\pi_{k-1} > 0$ for $\sigma < \lambda_1^{(k-1)}$. That proves (3.2). □

Applying Lemma 3.1, a fundamental relation between π_k, c_k and their derivatives with respect to σ is obtained. Let $f'(\sigma) = df(\sigma)/d\sigma$.

LEMMA 3.2. *For all real σ and $1 \le k \le n$,*

$$\pi_k c_k' - c_k \pi_k' = 1. \tag{3.5}$$

Proof. Differentiate (2.7):

$$-y_k + (T_k - \sigma I_k) y_k' = \pi_k' e_k^{(k)}.$$

Multiply by $(y_k)^t$ and recall the definition of y_k in (2.2):

$$-(y_k)^t y_k + (y_k)^t (T_k - \sigma I_k) y_k' = c_k \pi_k'.$$

Equation (3.5) follows since $(y_k)^t y_k = 1$ by (2.2) and $(y_k)^t (T_k - \sigma I_k) = \pi_k (e_k^{(k)})^t$ by (2.7). □

COROLLARY 1 OF LEMMA 3.2. *For $2 \le k \le n$,*

$$\pi_k'(\lambda_i^{(k)}) = -\frac{1}{c_k(\lambda_i^{(k)})} = \frac{(-1)^i}{\omega_{ik}}, \qquad 1 \le i \le k. \tag{3.6}$$

Proof. Since $\pi_k(\lambda_i^{(k)}) = 0$ by (3.1), relation (3.5) at $\lambda_i^{(k)}$ becomes

$$c_k(\lambda_i^{(k)}) \pi_k'(\lambda_i^{(k)}) = -1.$$

The result follows from (3.2). □

COROLLARY 2 OF LEMMA 3.2.

$$\pi_k''(\lambda_i^{(k)}) = 0, \qquad 1 \le i \le k. \tag{3.7}$$

Proof. Differentiate (3.5): $\pi_k c_k'' - c_k \pi_k'' = 0$. Set $\sigma = \lambda_i^{(k)}$, then $\pi_k = 0$ by (3.1) and $c_k = \pm\omega_{ik} \ne 0$ by (3.2). □

We now turn to the behaviour of π_k for all other values of σ.

LEMMA 3.3. *For T unreduced and $k \le n$,*

$$\pi_k \pi_k'' < 0 \quad \textit{for } \sigma \ne \lambda_i^{(k)}. \tag{3.8}$$

Proof. Premultiply both sides of (2.7) by $(T_k - \sigma I_k)^{-1} \pi_k^{-1}$ to find

$$\frac{y_k}{\pi_k} = (T_k - \sigma I_k)^{-1} e_k^{(k)} \quad \text{for } \sigma \ne \lambda_i^{(k)}. \tag{3.9}$$

A consequence of the spectral factorization is

$$(e_k^{(k)})^t (T_k - \sigma I_k)^{-p} e_k^{(k)} = \sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^p} \quad \text{for } \sigma \ne \lambda_i^{(k)}. \tag{3.10}$$

Since $y_k^t y_k = 1$ by (2.2), (3.9) yields

$$\frac{1}{\pi_k^2} = \left(e_k^{(k)}\right)^t (T_k - \sigma I_k)^{-2} e_k^{(k)}. \tag{3.11}$$

Combine (3.10) and (3.11) to find

$$\frac{1}{\pi_k^2} = \sum_{i=1}^{k} \left(\frac{\omega_{i,k}}{\lambda_i^{(k)} - \sigma} \right)^2 \quad \text{for } \sigma \neq \lambda_i^{(k)}. \tag{3.12}$$

Differentiate both sides of (3.12):

$$-\frac{1}{\pi_k^3}\pi_k' = \sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^3} \quad \text{for } \sigma \neq \lambda_i^{(k)}. \tag{3.13}$$

Differentiate the right-hand side of (3.13):

$$3\sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4}.$$

Differentiate the left-hand side of (3.13):

$$-\frac{1}{\pi_k^3}\pi_k'' + \frac{3}{\pi_k^4}(\pi_k')^2.$$

Therefore, for $\sigma \neq \lambda_i^{(k)}$ with $\sum$ denoting $\sum_{i=1}^{k}$,

$$\begin{aligned}
-\frac{1}{\pi_k^5}\pi_k'' &= \frac{3}{\pi_k^2}\sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4} - \frac{3}{\pi_k^6}(\pi_k')^2 \\
&= 3\sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^2} \sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4} - 3\left(\frac{1}{\pi_k^3}\pi_k'\right)^2 \quad \text{by (3.12)} \\
&= 3\left[\sum \left(\frac{\omega_{i,k}}{\lambda_i^{(k)} - \sigma} \right)^2 \sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4} - \left(\sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^3} \right)^2 \right] \quad \text{by (3.13)} \\
&\geq 0 \quad \text{by the Cauchy–Schwarz inequality.}
\end{aligned}$$

Moreover, equality holds if and only if the following two vectors,

$$z_1 = \left(\frac{\omega_{1,k}}{\lambda_1^{(k)} - \sigma}, \ldots, \frac{\omega_{k,k}}{\lambda_k^{(k)} - \sigma} \right)^t,$$

$$z_2 = \left(\frac{\omega_{1,k}}{(\lambda_1^{(k)} - \sigma)^2}, \ldots, \frac{\omega_{k,k}}{(\lambda_k^{(k)} - \sigma)^2} \right)^t,$$

are proportional. That is,

$$\lambda_1^{(k)} - \sigma = \cdots = \lambda_k^{(k)} - \sigma.$$

This is impossible, since T is unreduced. Therefore,

$$-\frac{1}{\pi_k^5}\pi_k''(\sigma) > 0 \quad \text{for } \sigma \neq \lambda_i^{(k)}.$$

Recall from (3.1) that $\pi_k(\sigma) \neq 0$ for $\sigma \neq \lambda_i^{(k)}$. Therefore,

$$-\pi_k \pi_k'' > 0 \quad \text{for } \sigma \neq \lambda_i^{(k)}.$$

□

COROLLARY OF LEMMA 3.3.

$$\pi_k'' \neq 0 \quad \textit{for } \sigma \neq \lambda_i^{(k)}. \tag{3.14}$$

Lemma 3.3 shows that the algebraic function $\pi_k(\sigma)$ is like the characteristic polynomial of T_k in that it vanishes at the eigenvalues of T_k. Moreover, it is alternatingly concave upward and downward in the intervals bounded by the eigenvalues of T_k. That shows that π_k' attains its extreme values at the $\lambda_i^{(k)}, i = 1, \ldots, k$.

LEMMA 3.4. *It holds that*

$$\lim_{\sigma\to\lambda_i^{(k)}} \left(-\frac{\pi_k''}{3\pi_k}\right) = \frac{1}{\omega_{ik}^2}\sum{}' \frac{\omega_{jk}^2}{(\lambda_i^{(k)} - \lambda_j^{(k)})^2}, \tag{3.15}$$

where $\sum'$ *indicates that the term* $j = i$ *is omitted as* $j = 1, \ldots, k$.

Proof. Rearrange $-\pi_k''/\pi_k^5$ and use (3.12) to get, with $\sum$ denoting $\sum_{i=1}^k$,

$$\begin{aligned} -\frac{1}{3}\frac{\pi_k''}{\pi_k} &= \pi_k^4\left[\sum\frac{\omega_{i,k}^2}{(\lambda_i^{(k)}-\sigma)^2}\sum\frac{\omega_{i,k}^2}{(\lambda_i^{(k)}-\sigma)^4} - \left(\sum\frac{\omega_{i,k}^2}{(\lambda_i^{(k)}-\sigma)^3}\right)^2\right] \\ &= \frac{\left[\sum\frac{\omega_{i,k}^2}{(\lambda_i^{(k)}-\sigma)^2}\sum\frac{\omega_{i,k}^2}{(\lambda_i^{(k)}-\sigma)^4} - \left(\sum\frac{\omega_{i,k}^2}{(\lambda_i^{(k)}-\sigma)^3}\right)^2\right]}{\left(\sum\frac{\omega_{i,k}^2}{(\lambda_i^{(k)}-\sigma)^2}\right)^2}. \end{aligned} \tag{3.16}$$

To analyze (3.16) in a neighbourhood of $\lambda_i^{(k)}$, it is convenient to abbreviate the terms. Let

$$\sum\frac{\omega_{i,k}^2}{(\lambda_i^{(k)}-\sigma)^l} = \frac{\omega^2}{\delta^l} + m_l,$$

where $\omega = \omega_{ik}$ and $\delta = \lambda_i^{(k)} - \sigma$. Then (3.16) becomes

$$\frac{(m_2 + \frac{\omega^2}{\delta^2})(m_4 + \frac{\omega^2}{\delta^4}) - (m_3 + \frac{\omega^2}{\delta^3})(m_3 + \frac{\omega^2}{\delta^3})}{(m_2 + \frac{\omega^2}{\delta^2})(m_2 + \frac{\omega^2}{\delta^2})}.$$

Multiply the numerator and denominator by a term δ^4 and simplify the terms to get

$$-\frac{1}{3}\frac{\pi_k''}{\pi_k} = \frac{m_2\omega^2 + O(\delta)}{\omega^4 + O(\delta)}.$$

Then the result follows as $\delta \to 0$. □

COROLLARY OF LEMMA 3.4. *If T is unreduced, then*

$$\pi_k'''(\lambda_i^{(k)}) = (-1)^{i+1}\frac{3}{\omega_{ik}^3}\sum{}'\frac{\omega_{jk}^2}{(\lambda_i^{(k)} - \lambda_j^{(k)})^2}, \tag{3.17}$$

where $\sum'$ *indicates that the term* $j = i$ *is omitted as* $j = 1, \ldots, k$.

Proof. Apply L'Hospital's rule:

$$\frac{\pi_k'''(\lambda_i^{(k)})}{\pi_k'(\lambda_i^{(k)})} = \lim_{\sigma \to \lambda_i^{(k)}} \left(\frac{\pi_k''}{\pi_k} \right) = \frac{(-3)}{\omega_{ik}^2} \sum{}' \frac{\omega_{jk}^2}{(\lambda_i^{(k)} - \lambda_j^{(k)})^2}.$$

Using (3.6), the result is obtained. □

The sum in (3.17) plays a role in the local analysis of π_k near each zero $\lambda_i^{(k)}$, $i = 1, \ldots, k$.

3.2. The linear model. Near $\lambda_i^{(k)}$, the function $\pi_k(\sigma)$ can be approximated by

$$p_k^{(i)}(\sigma) = \frac{(-1)^i}{\omega_{ik}}(\sigma - \lambda_i^{(k)}). \tag{3.18}$$

We need to know the interval on which this approximation is accurate to within, say, 10 percent.

In this analysis we need the special quantity that appeared in (3.15).

DEFINITION.

$$m_k(i) := 1 \Big/ \sqrt{\sum{}' \omega_{jk}^2 (\lambda_j^{(k)} - \lambda_i^{(k)})^{-2}}, \qquad 1 \le i \le k,$$

where $\sum{}'$ indicates omission of the term $j = i$ as $j = 1, \ldots, k$.

The quantity $m_k(i)$ is almost a mean of the singular values of $T_k - \lambda_i^{(k)} I_k$. Note that

$$\sum{}' \omega_{jk}^2 = 1 - \omega_{ik}^2,$$

and so $m_k(i)^{-2}/(1 - \omega_{ik}^2)$ is a weighted average of the $\{(\lambda_j^{(k)} - \lambda_i^{(k)})^{-2}\}_{j \ne i}$. Recall that for positive numbers τ_i, any (weighted) harmonic mean is majorized by the (weighted) arithmetic mean and both lie between the minimum and maximum values, i.e., if $\sum w_i = 1$, $w_i \ge 0$, then

$$\min\ \tau_i \le \left[\sum w_i \tau_i^{-1} \right]^{-1} \le \sum w_i \tau_i \le \max\ \tau_i.$$

With $\tau_j = (\lambda_j^{(k)} - \lambda_i^{(k)})^2$, this gives

$$\begin{aligned} \min_{j \ne i} (\lambda_j^{(k)} - \lambda_i^{(k)})^2 &\le \left(\frac{\sum' \omega_{jk}^2 (\lambda_j^{(k)} - \lambda_i^{(k)})^{-2}}{\sum' \omega_{jk}^2} \right)^{-1} \\ &\le \frac{\sum' \omega_{jk}^2 (\lambda_j^{(k)} - \lambda_i^{(k)})^2}{\sum' \omega_{jk}^2} \le \max_{j \ne i} (\lambda_j^{(k)} - \lambda_i^{(k)})^2. \end{aligned}$$

However,

$$\sum{}' \omega_{jk}^2 (\lambda_j^{(k)} - \lambda_i^{(k)})^2 = \sum_{j=1}^{k} \omega_{jk}^2 (\lambda_j^{(k)} - \lambda_i^{(k)})^2 = e_k^t (T_k - \lambda_i^{(k)} I)^2 e_k.$$

Thus

$$\min \left\{ \lambda_{i+1}^{(k)} - \lambda_i^{(k)}, \lambda_i^{(k)} - \lambda_{i-1}^{(k)} \right\}^2 \le (1 - \omega_{ik}^2) m_k(i)^2$$

and

$$(3.19)\qquad (1-\omega_{ik}^2)m_k(i)^2 \le \frac{e_k^t(T_k-\lambda_i^{(k)}I)^2e_k}{1-\omega_{ik}^2} = \frac{(\alpha_k-\lambda_i^{(k)})^2+\beta_k^2}{1-\omega_{ik}^2}.$$

Now we can compare π_k with the linear function $p_k^{(i)}$.

LEMMA 3.5. *As* $\sigma \to \lambda_i^{(k)}$,

$$(3.20)\qquad \frac{\pi_k}{p_k^{(i)}} = 1 - \frac{1}{2}\left(\frac{\sigma-\lambda_i^{(k)}}{\omega_{ik}m_k(i)}\right)^2 + O((\sigma-\lambda_i^{(k)})^3).$$

Proof. Use (3.6), (3.7), and (3.17) to find that the Taylor series of π_k around $\lambda_i^{(k)}$ is

$$\begin{aligned}\pi_k(\sigma) &= 0 + \pi_k'(\lambda_i^{(k)})(\sigma-\lambda_i^{(k)}) + 0 + \frac{1}{6}\pi_k'''(\lambda_i^{(k)})(\sigma-\lambda_i^{(k)})^3 + \cdots \\ &= \frac{(-1)^{k-i}}{\omega_{ik}}(\sigma-\lambda_i^{(k)}) + \frac{1}{6}\frac{(-1)^{k-i-1}3(\sigma-\lambda_i^{(k)})^3}{\omega_{ik}^3 m_k(i)^2} + O((\sigma-\lambda_i^{(k)})^4) \\ &= p_k^{(i)}(\sigma)\left[1-\frac{1}{2}\left(\frac{\sigma-\lambda_i^{(k)}}{\omega_{ik}m_k(i)}\right)^2 + O((\sigma-\lambda_i^{(k)})^3)\right].\end{aligned}$$ □

Remark 1. It is not valid to use the linear model for π_k if the quadratic term in (3.20) exceeds 1. Thus, in the generic case when there is no happy cancellation between the quadratic and higher terms, it is *necessary* to require

$$|\sigma-\lambda_i^{(k)}| < \tfrac{1}{2}\omega_{ik}m_k(i)$$

so that

$$\frac{\pi_k}{p_k^{(i)}} > \frac{7}{8} + O((\sigma-\lambda_i^{(k)})^3).$$

In Fig. 7, the graphs of π_k and $p_k^{(i)}$ are shown as well as the boundary points

$$\lambda_i^{(k)} \pm \tfrac{1}{2}\omega_{ik}m_k(i)$$

in a typical case.

Remark 2. Since

$$m_k(i)^{-2} \le (k-1)\max_{j\ne i}\left(\frac{\omega_{jk}}{\lambda_j^{(k)}-\lambda_i^{(k)}}\right)^2,$$

it follows that

$$\frac{|\lambda_i^{(k)}-\sigma|}{\omega_{ik}m_k(i)} \le \frac{\sqrt{k-1}|\lambda_i^{(k)}-\sigma|}{\omega_{ik}\min_{j\ne i}\left(\frac{|\lambda_j^{(k)}-\lambda_i^{(k)}|}{\omega_{jk}}\right)},$$

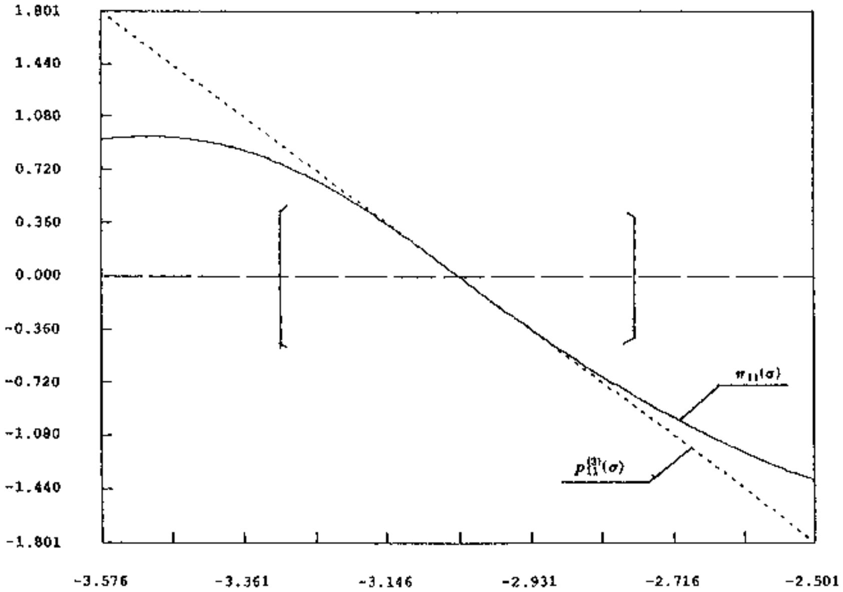

FIG. 7. π_{11} *and the linear model* $p_{11}^{(3)}$ *on* $[\lambda_3^{(11)} - \omega_{3,11} m_{11}(3), \lambda_3^{(11)} + \omega_{3,11} m_{11}(3)]$ *with* $T = W_{11}^-$ *(see* §3.2).

and when the cubic terms are negligible, the constraint

$$|\lambda_i^{(k)} - \sigma| \le \frac{\omega_{ik}}{3\sqrt{k-1}} \min_{j \ne i} \left(\frac{|\lambda_j^{(k)} - \lambda_i^{(k)}|}{\omega_{jk}} \right)$$

is a *sufficient* condition for using the linear model. Although the minimum will not be known, the expression shows how a small value of ω_{jk} tends to neutralize the effect of close $\lambda_j^{(k)}$ in restricting the domain of the linear model.

In the analysis of $\hat{\beta}_k'$ and $\hat{\alpha}_k'$ near $\lambda_i^{(k)}$, the following auxiliary function plays a role:

$$\phi_k := \pi_k \pi_k' / \xi_k^2 = \pi_k \pi_k' / (\pi_{k-1}^2 + \beta_k^2).$$

See Figs. 2 and 3 for the shape of ϕ_k near $\lambda_i^{(k)}$.

LEMMA 3.6. *When* $|\sigma - \lambda_i^{(k)}| < \frac{1}{2}\omega_{ik} m_k(i)$ *and* $k < n$, *then*

$$\phi_k / f_k^{(i)} \in (1/2, 1]$$

where

$$f_k^{(i)}(\sigma) := \frac{\sigma - \lambda_i^{(k)}}{\beta_{k+1}^2 \omega_{ik}^2 + (\sigma - \lambda_i^{(k)})^2}, \qquad 1 \le i \le k. \tag{3.21}$$

Proof. Abbreviate $\sigma - \lambda_i^{(k)}$ by δ, ω_{ik} by ω, $m_k(i)$ by m, and use Lemma 3.5 to find

$$\begin{aligned}\frac{\phi_k}{f_k^{(i)}} &= \frac{1 - 2(\frac{\delta}{\omega m})^2 + O(\delta^3)}{1 - (\frac{\delta}{\omega m})^2 \frac{\delta^2}{\omega^2\beta^2+\delta^2} + O(\delta^5)} \\ &= 1 - \left(2 - \frac{\delta^2}{\omega^2\beta^2 + \delta^2}\right)\left(\frac{\delta}{\omega m}\right)^2 + O(\delta^3).\end{aligned}$$

Thus, when $|\delta| < \frac{1}{2}\omega m$,

$$0 \leq -\frac{\phi_k}{f_k^{(i)}} + 1 + O(\delta^3) < \frac{1}{2}.$$

Equality on the left occurs only when $\delta = 0$. □

COROLLARY.

$$|f_k^{(i)}| \leq \frac{1}{2\beta_{k+1}\omega_{ik}}$$

with equality if and only if $\sigma = \lambda_i^{(k)} \pm \omega_{ik}\beta_{k+1}$.

Remark. If $\beta_{k+1} > m_k(i)$, then $|\phi_k|$ will not attain the bound on $|f_k^{(i)}|$ since $|\phi_k| < |f_k^{(i)}|$ on the interval in Lemma 3.6, and this interval does not contain the maximizing point of $|f_k^{(i)}|$.

3.3. Pointwise bounds.

We now give a pointwise bound on $\pi_k^{'}$ that reveals the role of the distance of σ from the spectrum of T_k and from the spectrum of T_{k-1}. Some preliminary results are restated for easy reference.

$$\begin{array}{lll}
\text{(A)} & \xi_{k-1} = \sqrt{\pi_{k-1}^2 + \beta_k^2} & \text{(see definition in (2.9))}, \\
\text{(B)} & c_k = \pi_{k-1}/\xi_{k-1} & \text{(see definition in (2.10))}, \\
\text{(C)} & s_k = \beta_k/\xi_{k-1} & \text{(see definition in (2.11))}, \\
\text{(D)} & c_k^{'} = \frac{s_k^2}{\xi_{k-1}}\pi_{k-1}^{'} & \text{(derivative of (B))}, \\
\text{(E)} & \pi_k c_k^{'} - c_k \pi_k^{'} = 1 & \text{(for any real } \sigma \text{ (3.5))}, \\
\text{(F)} & (T_k - \sigma I_k)y_k = \pi_k e_k^{(k)} & \text{(for any real } \sigma \text{ (2.7))}.
\end{array}$$

DEFINITION. $\mu_k := \mu_k(\sigma) = \min_{1\leq i\leq k} |\lambda_i^{(k)} - \sigma|, k = 1, 2, \ldots, n.$

$$\text{(G)} \qquad \mu_k^2 < \pi_k^2 \quad (\text{for } \sigma \neq \lambda_i^{(k)}).$$

Proof of (G). $(F)^t(F)$ yields $(y_k)^t(T_k - \sigma I_k)^2 y_k = \pi_k^2$. Then

$$\begin{aligned}\mu_k^2 &= \min_{1\leq i\leq k}(\sigma - \lambda_i^{(k)})^2 \quad \text{(by definition)} \\ &= \lambda_{\min}((T_k - \sigma I_k)^2) \\ &< (y_k)^t(T_k - \sigma I_k)^2 y_k \quad \text{since } y_k \text{ cannot be an eigenvector since } \sigma \neq \lambda_i^{(k)} \\ &= \pi_k^2. \qquad \square\end{aligned}$$

The next result shows that $|\pi_k^{'}|$ can only be huge when $\mu_k(\sigma)$ and $\mu_{k-1}(\sigma)$ are both tiny.

LEMMA 3.7. *For any real* σ,

$$\mu_k(\sigma)|\pi_k'| = |\pi_k|, \qquad \sigma = \lambda_i^{(k)}, \tag{3.22}$$

$$\mu_k(\sigma)|\pi_k'| < |\pi_k|, \qquad \sigma \neq \lambda_i^{(k)}, \tag{3.23}$$

$$\mu_{k-1}(\sigma)|\pi_k'| < s_k^2|\pi_k| + \sqrt{\mu_{k-1}^2(\sigma) + \beta_k^2}. \tag{3.24}$$

Proof of (3.22), (3.23). When $\sigma = \lambda_i^{(k)}$, then $\pi_k = 0$ by (3.1), $\mu_k = 0$ by definition, and (3.22) holds. Now suppose that $\sigma \neq \lambda_i^{(k)}$. Premultiply (F) by $(T_k - \sigma I_k)^{-1}\pi_k{}^{-1}$:

$$\frac{y_k}{\pi_k} = (T_k - \sigma I_k)^{-1} e_k^{(k)}. \tag{3.25}$$

Differentiate (3.25):

$$\frac{y_k'\pi_k - y_k\pi_k'}{\pi_k^2} = (T_k - \sigma I_k)^{-2} e_k^{(k)}. \tag{3.26}$$

Differentiate $y_k^t y_k = 1$ to obtain $y_k^t y_k' = 0$. Premultiply (3.26) by y_k^t, getting

$$\begin{aligned} -\frac{1}{\pi_k^2}\pi_k' &= (y_k)^t(T_k - \sigma I_k)^{-2} e_k^{(k)} \quad \text{since } y_k^t y_k' = 0 \\ &= (y_k)^t (T_k - \sigma I_k)^{-1} y_k/\pi_k \quad \text{by (3.25)}. \end{aligned}$$

Thus

$$-\pi_k' = (y_k)^t (T_k - \sigma I_k)^{-1} y_k \pi_k$$

and

$$|\pi_k'| = |(y_k)^t (T_k - \sigma I_k)^{-1} y_k||\pi_k|. \tag{3.27}$$

Since $\sigma \neq \lambda_i^{(k)}$, y_k will not be an eigenvector. Therefore,

$$\begin{aligned} |(y_k)^t(T_k - \sigma I_k)^{-1} y_k| &< \max_{v^t v = 1} |v^t (T_k - \sigma I_k)^{-1} v| \\ &= \|(T_k - \sigma I_k)^{-1}\| \\ &= \frac{1}{\min_{1 \le i \le k} |\lambda_i^{(k)} - \sigma|} = \frac{1}{\mu_k}. \end{aligned}$$

Put this inequality into (3.27) and multiply by μ_k to obtain (3.23). □

Proof of (3.24). When $\sigma = \lambda_i^{(k-1)}$, then $\mu_{k-1} = 0$, and (3.24) is immediate since $\beta_k \neq 0$ and $s_k^2|\pi_k| \neq 0$ by (3.1).

Now, suppose that $\sigma \neq \lambda_i^{(k-1)}$. Then,

$$\begin{aligned} |\pi_k' c_k + 1| &= |c_k' \pi_k| \quad \text{by (E)} \\ &= \frac{s_k^2}{\xi_{k-1}} |\pi_{k-1}' \pi_k| \quad \text{by (D)} \\ &< \frac{s_k^2}{\xi_{k-1}} \frac{|\pi_{k-1}|}{\mu_k} - 1|\pi_k| \quad \text{by (3.23)} \\ & \frac{{}'c_k \pi_k|}{{}_{-1}} \quad \text{by (B)}. \end{aligned}$$

Hence,

$$|\pi_k' c_k| < \frac{s_k^2 |c_k \pi_k|}{\mu_{k-1}} + 1.$$

Since $\sigma \neq \lambda_i^{(k-1)}$, $c_k \neq 0$ by (3.1). Then the above inequality can be rearranged as

$$\begin{aligned}
\mu_{k-1}|\pi_k'| &< s_k^2|\pi_k| + \frac{\mu_{k-1}}{|c_k|} \\
&= s_k^2|\pi_k| + \mu_{k-1}\frac{\xi_{k-1}}{|\pi_{k-1}|} \quad \text{by (B)} \\
&= s_k^2|\pi_k| + \mu_{k-1}\frac{\sqrt{\pi_{k-1}^2 + \beta_k^2}}{|\pi_{k-1}|} \quad \text{by (A)} \\
&= s_k^2|\pi_k| + \sqrt{\mu_{k-1}^2 + \beta_k^2 \frac{\mu_{k-1}^2}{\pi_{k-1}^2}} \\
&< s_k^2|\pi_k| + \sqrt{\mu_{k-1}^2 + \beta_k^2} \quad \text{by (G).}
\end{aligned}$$

□

4. Sensitivity of $\hat{T}$. Recall that $\hat{T} = RQ + \sigma I = Q^t T Q$ where $T - \sigma I = QR$. In terms of the intermediate quantities generated in the QR transformation, the entries of $\hat{T}$ are given in (2.14) and (2.15), namely,

$$\begin{aligned}
&\hat{\beta}_k = \xi_k s_k = \beta_k(\xi_k/\xi_{k-1}) \quad \text{for } k < n \quad \text{and} \quad \xi_k^2 = \pi_k^2 + \beta_{k+1}^2, \\
&\hat{\alpha}_k = \alpha_{k+1} - \gamma_{k+1} + \gamma_k \quad \text{for } k < n \quad \text{and} \quad \gamma_k = \pi_k c_k, \\
&\hat{\beta}_n = \pi_n s_n, \qquad \hat{\alpha}_n = \gamma_n + \sigma.
\end{aligned} \tag{4.0}$$

Differentiating with respect to σ reveals that

$$\hat{\beta}_k' = \xi_k' s_k + \xi_k s_k' = \beta_k(\xi_k/\xi_{k-1})', \tag{4.1}$$

$$\hat{\alpha}_k' = \gamma_k' - \gamma_{k+1}', \tag{4.2}$$

$$|\hat{\beta}_n'| = |\pi_n' s_n + \pi_n s_n'|, \tag{4.3}$$

$$\hat{\alpha}_n' = \gamma_n' + 1. \tag{4.4}$$

Here is the plan of this section. Our first goal is to show that these derivatives cannot be huge unless σ is close to an eigenvalue of some T_k. Next we show that, in fact, to have "huge" derivatives, σ must be close to the spectrum of two successive T_k's. The proofs repeatedly use two results, (2.7) and (3.23), from the previous sections:

$$c_k' \pi_k - c_k \pi_k' = 1 \quad \text{for all } \sigma \quad \text{and} \quad k = 2, \ldots, n, \tag{4.5}$$

and

$$|\pi_k'| < |\pi_k|/\mu_k \quad \text{for } \sigma \notin \text{ spectrum } (T_k). \tag{4.6}$$

Here

$$\mu_k = \min_{1\le i\le k} |\sigma - \lambda_i^{(k)}| = \operatorname{dist}(\sigma, \operatorname{spectrum}(T_k)). \tag{4.7}$$

Finally, we give realistic bounds on all these derivatives.

The first goal is met by the following result.

THEOREM 4.1. *With $\mu_k(\sigma)$ defined in* (4.7) *and $k < n$,*

$$\text{(i)} \qquad |\hat{\beta}_k'| < \hat{\beta}_k \left(\frac{c_{k+1}^2}{\mu_k} + \frac{c_k^2}{\mu_{k-1}} \right), \quad \sigma \notin \operatorname{spectrum}(T_k) \cup \operatorname{spectrum}(T_{k-1}), \tag{4.8}$$

$$\text{(ii)} \qquad |\hat{\alpha}_k'| < 2 \left(\frac{|\gamma_k|}{\mu_k} + \frac{|\gamma_{k-1}|}{\mu_{k+1}} \right), \quad \sigma \notin \operatorname{spectrum}(T_{k+1}) \cup \operatorname{spectrum}(T_k), \tag{4.9}$$

$$\text{(iii)} \qquad |\hat{\beta}_n'| < \hat{\beta}_n \left(\frac{1}{\mu_n} + \frac{c_n^2}{\mu_{n-1}} \right), \quad \sigma \notin \operatorname{spectrum}(T_n) \cup \operatorname{spectrum}(T_{n-1}), \tag{4.10}$$

$$\text{(iv)} \qquad |\hat{\alpha}_n'| < 2 \left(\frac{|\gamma_n|}{\mu_n} + 1 \right), \quad \sigma \notin \operatorname{spectrum}(T_n). \tag{4.11}$$

Proof.

$$\text{(i)} \qquad \frac{\xi_k}{\xi_{k-1}} = \sqrt{\frac{\pi_k^2 + \beta_{k+1}^2}{\pi_{k-1}^2 + \beta_k^2}}.$$

Applying the quotient rule and simplifying the result yields

$$\left(\frac{\xi_k}{\xi_{k-1}} \right)' = \left(\frac{\xi_k}{\xi_{k-1}} \right) \left(\frac{\pi_k \pi_k'}{\xi_k^2} - \frac{\pi_{k-1}\pi_{k-1}'}{\xi_{k-1}^2} \right).$$

Use (4.6) to get

$$\left| \left(\frac{\xi_k}{\xi_{k-1}} \right)' \right| < \left(\frac{\xi_k}{\xi_{k-1}} \right) \left(\frac{\pi_k^2}{\xi_k^2 \mu_k} + \frac{\pi_{k-1}^2}{\xi_{k-1}^2 \mu_{k-1}} \right)$$
$$= \frac{\xi_k}{\xi_{k-1}} \left(\frac{c_{k+1}^2}{\mu_k} + \frac{c_k^2}{\mu_{k-1}} \right).$$

Multiply each side by β_k and use (4.1) to obtain (4.8).

(ii) From (4.2) and (4.0),

$$\hat{\alpha}_k' = \gamma_k' - \gamma_{k+1}' = c_k' \pi_k + c_k \pi_k' - c_{k+1}' \pi_{k+1} - c_{k+1} \pi_{k+1}'.$$

Now use (4.5) to conclude $\hat{\alpha}_k' = 1 + 2c_k\pi_k' - (1 + 2c_{k+1}\pi_{k+1}')$, and use (4.6) to find

$$|\hat{\alpha}_k'| < 2 \left(\frac{|c_k \pi_k|}{\mu_k} + \frac{|c_{k+1}\pi_{k+1}|}{\mu_{k+1}} \right) \quad \text{for } \sigma \notin \operatorname{spectrum}(T_k) \cup \operatorname{spectrum}(T_{k+1}).$$

$$\text{(iii)} \qquad s_n' = (\beta_n/\xi_{n-1})' = -\frac{\beta_n \pi_{n-1} \pi_{n-1}'}{\xi_{n-1}^3} = -s_n c_n \frac{\pi_{n-1}'}{\xi_{n-1}}. \tag{4.12}$$

$$
\begin{aligned}
|\hat{\beta}_n'| &= |\pi_n' s_n - \pi_n s_n c_n \pi_{n-1}'/\xi_{n-1}| \\
&< |\pi_n| s_n \left(\frac{1}{\mu_n} + \frac{|c_n \pi_{n-1}|}{\xi_{n-1}\mu_{n-1}} \right) \quad \text{for } \sigma \notin \text{spectrum}(T_n) \cup \text{spectrum}(T_{n-1}), \\
&= \hat{\beta}_n \left(\frac{1}{\mu_n} + \frac{c_n^2}{\mu_{n-1}} \right).
\end{aligned}
$$

$$
\text{(iv)} \qquad \hat{\alpha}_n' = \gamma_n' + 1 = \begin{cases} 2\pi_n c_n' \\ 2(\pi_n' c_n + 1) \end{cases} \qquad \text{since } c_n'\pi_n = 1 + c_n \pi_n'.
$$

Then

$$
|\hat{\alpha}_n'| < \begin{cases} 2|\gamma_n|/\mu_{n-1} \\ 2\left(\dfrac{|\pi_n c_n|}{\mu_n} + 1 \right) \end{cases} \quad \text{for } \sigma \notin \text{spectrum}(T_n) \cup \text{spectrum}(T_{n-1}). \qquad \square
$$

It may be verified that the $\gamma_i, i = 1, \ldots, n$ are intermediate quantities that appear on the diagonal when $T - \sigma I$ is being transformed into $\hat{T}$ by a sequence of plane rotations. Hence $|\gamma_i| < \|T - \sigma I\|$ for $i = 1, \ldots, n$ and all σ. Thus huge values for the derivative of $\hat{T}$ can come only from tiny values of some μ_k.

The next step is to show that the derivatives at the points excluded in the previous theorem are of modest size *unless* the excluded points, in each case, are very close to each other.

In order to simplify notation we shall label eigenvalues of the submatrices T_k so that the closest eigenvalue of T_{k-1} to $\lambda_i^{(k)}$ is $\lambda_{i_-}^{(k-1)}$ and of T_{k+1} is $\lambda_{i_+}^{(k+1)}$. Formally,

$$
(*) \qquad \mu_{k-1}(\lambda_i^{(k)}) = |\lambda_{i_-}^{(k-1)} - \lambda_i^{(k)}|, \qquad \mu_{k+1}(\lambda_i^{(k)}) = |\lambda_{i_+}^{(k+1)} - \lambda_i^{(k)}|.
$$

In all cases, i_- is either i or $i-1$ and i_+ is either i or $i+1$.

THEOREM 4.2. *With the notation above, for $i = 1, 2, \ldots, k$,*

$$
\begin{aligned}
&\text{(i)} && |\hat{\beta}_k'(\lambda_{i_-}^{(k-1)})| = \hat{\beta}_k |\tfrac{\pi_k \pi_k'}{\xi_k^2}|_{\lambda_{i_-}^{(k-1)}} < \hat{\beta}_k \tfrac{c_{k+1}^2}{|\lambda_{i_-}^{(k-1)} - \lambda_i^{(k)}|}, \\
&\text{(ii)} && |\hat{\beta}_k'(\lambda_i^{(k)})| = | - \hat{\beta}_k \tfrac{\pi_{k-1}\pi_{k-1}'}{\xi_{k-1}^2}|_{\lambda_i^k} < \hat{\beta}_k \tfrac{c_k^2}{|\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}|}, \\
&\text{(iii)} && |\hat{\alpha}_k'(\lambda_i^{(k)})| = |2c_k\pi_k'|_{\lambda_i^{(k)}} = 2, \\
&\text{(iv)} && |\hat{\alpha}_k'(\lambda_{i_+}^{(k+1)})| = |2(c_k\pi_k'|_{\lambda_{i_+}^{(k+1)}} + 1)| < 2\left(1 + \tfrac{|\gamma_k|}{|\lambda_{i_+}^{(k+1)} - \lambda_i^{(k)}|}\right), \\
&\text{(v)} && |\hat{\beta}_n'(\lambda_i^{(n)})| = |s_n\pi_n'|_{\lambda_i^{(n)}} < \sqrt{\tfrac{\Delta^2 + \beta_n^2}{\Delta}}; \qquad \Delta = |\lambda_i^{(n)} - \lambda_{i_-}^{(n-1)}|, \\
&\text{(vi)} && |\hat{\beta}_n'(\lambda_{i_-}^{(n-1)})| = |\pi_n' s_n|_{\lambda_{i_-}^{(n-1)}} < \tfrac{\hat{\beta}_n}{\Delta}, \\
&\text{(vii)} && \hat{\alpha}_n'(\lambda_i^{(n)}) = 0, \\
&\text{(viii)} && |\hat{\alpha}_n'(\lambda_i^{(n-1)})| = 2.
\end{aligned}
$$

Proof. (i) and (ii) Recall that

$$
\hat{\beta}_k' = \beta_k \left(\frac{\pi_k \pi_k'}{\xi_k^2} - \frac{\pi_{k-1}\pi_{k-1}'}{\xi_{k-1}^2} \right)
$$

and note that one of the two terms vanishes at the evaluation points and the other may be bounded in the same way as in the previous theorem.

(iii) Recall that $\hat{\alpha}_k' = 2(c_k\pi_k' - c_{k+1}\pi_{k+1}')$ and note that one term evaluates to -1 at the evaluation point since, for example, $c_k\pi_k' - c_k'\pi_k = -1$ and π_k vanishes at $\lambda_i^{(k)}$. Also, $c_{k+1} = 0$ at $\lambda_i^{(k)}$.

(iv) Recall that $\hat{\alpha}_k' = 2+2c_k\pi_k' - 2c_{k+1}'\pi_{k+1}$, and note that π_{k+1} vanishes at $\lambda_{i_+}^{(k+1)}$ and that the other may be bounded in the same way as in the previous theorem.

(v) and (vi) The second term $\pi_n s_n'$ vanishes at both evaluation points. To bound π_n' at $\lambda_k^{(n)}$ we invoke the inequality in (3.24) from the previous section. To bound π_n' at $\lambda_{i_-}^{(n-1)}$ we use the proof of Theorem 4.1 as before.

(vii) and (viii) $\hat{\alpha}_n' = 2\pi_n c_n' = 2(1 + \pi_n' c_n)$ and $\pi_n(\lambda_i^{(n)}) = 0$, $c_n(\lambda_{i_-}^{(n-1)}) = 0$. □

Preamble to Theorem 4.3. Theorem 4.1 shows that it is necessary for σ to be close to some $\lambda_i^{(k)}$, $i = 1, 2, \ldots, k$ in order for $\hat{\beta}_k$ and $\hat{\alpha}_k$ to have large derivatives. Theorem 4.2 shows that $\lambda_i^{(k)}$ must be very close to $\lambda_{i_-}^{(k-1)}$ in order for the derivatives to be large at these points. It remains to show that this last condition is also sufficient for extreme sensitivity of $\hat{T}$ to changes in σ. Indeed $|\hat{\beta}_k'|$, for $k < n$, may be modest at both $\lambda_i^{(k)}$ and $\lambda_{i_-}^{(k-1)}$ even when they are extremely close, and it is only at other values of σ in the neighborhood of $\lambda_i^{(k)}$ that the large values occur. Theorem 4.3 establishes this fact.

What permits this analysis is that in an interval of width $m_k(i)\omega_{ik}$ centered on $\lambda_i^{(k)}$ (see the linear model discussion in §3 for the definition of $m_k(i)$), both $\hat{\alpha}_k'$ and $\hat{\beta}_k'$ may be approximated closely by much simpler functions of σ whose maxima can be estimated. Recall that ω_{ik} is the magnitude of the last entry in $\lambda_i^{(k)}$'s normalized eigenvector. One difficulty in the analysis is to pin down $\lambda_{i_-}^{(k-1)}$, the closest eigenvalue of T_{k-1}, which must lie within the interval that we just associated with $\lambda_i^{(k)}$. Theorem 4.4, quoted at the end of this section, shows that

$$|\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}| = O(\omega_{ik}^2),$$

except in special circumstances when it is only $O(\omega_{ik})$, the width of our interval.

We digress to explain the situation. Lemma 2.1 shows that

$$\|Ty_k^{(n)} - y_k^{(n)}\lambda_i^{(k)}\| = \sqrt{\pi_k^2 + \beta_{k+1}^2 c_k^2} = \beta_{k+1}\omega_{ik},$$

since $\pi_k(\lambda_i^{(k)}) = 0$ and $|c_k(\lambda_i^{(k)})| = \omega_{ik}$. Standard results on symmetric matrices (see [4]) let us conclude that there is an eigenvalue λ of T satisfying $|\lambda - \lambda_i^{(k)}| \le \beta_{k+1}\omega_{ik}$. Moreover, the same can be said for some eigenvalue λ for each T_j with $j > k$. When $\omega_{ik} << 1$, we say that $\lambda_i^{(k)}$ has "stabilized" as an approximate eigenvalue of T. In many cases, there will only be one eigenvalue of T_{k-1} close to $\lambda_i^{(k)}$ and, in that case, their separation is $O(\omega_{ik}^2)$. Theorem 4.4 gives

$$|\Delta| := |\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}| < \omega_{ik}^2 \begin{cases} \lambda_k^{(k)} - \lambda_1^{(k)}, & i = 1, k, \\ \frac{(\lambda_k^{(k)} - \lambda_i^{(k)})(\lambda_i^{(k)} - \lambda_1^{(k)})}{|\lambda_i^{(k)} - \lambda_{i_*}^{(k-1)}|}, & 1 < i < k. \end{cases}$$

Here i_* is the second closest eigenvalue and $i_- + i_* = 2i - 1$. Thus, in all cases when $i \neq 1, k$,

$$\begin{aligned}|\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}| &\leq \sqrt{|\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}||\lambda_i^{(k)} - \lambda_{i_*}^{(k-1)}|}\\ &< \omega_{ik}((\lambda_k^{(k)} - \lambda_i^{(k)})(\lambda_i^{(k)} - \lambda_1^{(k)}))^{1/2} = O(\omega_{ik}).\end{aligned}$$

In the special cases when $|\lambda_i^{(k)} - \lambda_{i_*}^{(k-1)}|$ is comparable to Δ, our simple model for π_k is still applicable, but the maxima of the derivatives of $\hat{\alpha}_k$ and $\hat{\beta}_k$ are not so large, in general, and are complicated to express. In fact, the large derivatives will occur for a smaller value of k when the first stabilization occurs. That is why we make the assumption $\Delta = O(\omega_{ik}^2)$ in Theorem 4.3. A simple version of Theorem 4.3 is given next. The distribution of max and min was a surprise to us.

Near $\sigma = \lambda_i^{(k)}$,

$$\text{(i)} \qquad \max_\sigma |\hat{\beta}_n'| = \frac{1}{\omega_{\text{in}}}, \qquad k = n,$$

$$\text{(ii)} \qquad \max_\sigma |\hat{\beta}_k'| = O\left(\frac{1}{\min(\omega_{ik}, \omega_{i_-,k-1})}\right), \qquad k < n,$$

$$\text{(iii)} \qquad \max_\sigma |\hat{\alpha}_n'| = O\left(\frac{1}{\max(\omega_{ik}, \omega_{i_-,k-1})}\right), \qquad k = n,$$

$$\text{(iv)} \qquad \max_\sigma |\hat{\alpha}_k'| = O\left(\frac{1}{\min\left\{\max(\omega_{ik}, \omega_{i_-,k-1}), \max(\omega_{ik}, \omega_{i_+,k+1})\right\}}\right), \qquad k < n.$$

The function of Theorem 4.3 is to provide the constants hidden by the O. An interesting byproduct of the proof of Theorem 4.3 is that only for $|\hat{\beta}_n'|$ does the maximum occur in the tiny interval $[\lambda_i^{(n)}, \lambda_{i_-}^{(n-1)}]$. We remind the reader that from the preamble to the proof of Lemma 3.5

$$m_k(i) = 1 \Big/ \sqrt{\sum_{j \neq i} \omega_{jk}^2 (\lambda_j^{(k)} - \lambda_i^{(k)})^{-2}}.$$

THEOREM 4.3. *Near $\sigma = \lambda_i^{(k)}$, for $2 \leq k \leq n$ and $1 \leq i \leq k$, if*

$$\Delta_i^k := |\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}| < 100(\lambda_k^{(k)} - \lambda_1^{(k)})\omega_{ik}^2 = O(\omega_{ik}^2),$$

then the entries in $\frac{d}{d\sigma}\hat{T}$ achieve the values indicated below.

$$\text{(i)} \qquad \max_\sigma |\hat{\beta}_n'| = \frac{1}{\omega_{in}}\left[1 + \frac{1}{8}\left(\frac{\Delta_i^n}{\beta_n \omega_{i_-,n-1}}\right)^2 - \frac{3}{8}\left(\frac{\Delta_i^n}{\omega_{in} m_n(i)}\right)^2 + O(\Delta^3)\right]$$

and occurs near $(\lambda_i^{(n)} + \lambda_{i_-}^{(n-1)})/2$.

(ii) *The function $\hat{\beta}_k'$, $k < n$, is the difference of two similarly shaped functions; see Fig. 2. Under mild conditions, stated in the proof, the magnitude of the first term's model, on the appropriate interval, exceeds*

$$\frac{1}{2\omega_{ik}}.$$

Under mild conditions, stated in the proof, the magnitude of the second term's model, on the appropriate interval, exceeds

$$\left(\frac{\beta_{k+1}}{\sqrt{2}\beta_k}\right)\frac{1}{2\omega_{i_-,k-1}}.$$

When one of these terms dominates, then the larger one gives an accurate estimate of $\max|\hat{\beta}_k'|$ *near* $\lambda_i^{(k)}$*, under the stated conditions. In all other cases, the maximum value of the model is a more complicated expression and has a smaller value.*

(iii) *The function* $|\hat{\alpha}_n'|$ *attains its maximum, on the intervals associated with* $\lambda_i^{(n)}$ *and* $\lambda_{i_-}^{(n-1)}$*, at the ends of those intervals. There the model achieves*

$$\begin{cases} \dfrac{2}{\omega_{i_-,n-1}}\dfrac{m_n(i)}{\sqrt{4\beta_n^2+m_n^2(i)}} & \text{if } m_n(i)\omega_{in} < m_{n-1}(i_-)\omega_{i_-,n-1}, \\[2ex] \dfrac{2}{\omega_{in}}\dfrac{m_{n-1}(i_-)}{\sqrt{4\beta_n^2+m_{n-1}^2(i_-)}} & \text{otherwise.} \end{cases}$$

(iv) *The function* $\hat{\alpha}_k'$*,* $k < n$ *is the difference of two similarly shaped functions. On the appropriate interval about* $\lambda_i^{(k)}$*,* $i = 1, \ldots, k$*, the first term's model, in magnitude, achieves*

$$\begin{cases} \dfrac{2}{\omega_{i_-,k-1}}\dfrac{m_k(i)}{\sqrt{4\beta_k^2+m_k^2(i)}} & \text{if } m_k(i)\omega_{ik} < m_{k-1}(i_-)\omega_{i_-,k-1}, \\[2ex] \dfrac{2}{\omega_{ik}}\dfrac{m_{k-1}(i_-)}{\sqrt{4\beta_k^2+m_{k-1}^2(i_-)}} & \text{otherwise.} \end{cases}$$

On the appropriate interval about $\lambda_i^{(k)}$ *the second term's model, in magnitude, exceeds*

$$\begin{cases} \dfrac{2}{\omega_{ik}}\dfrac{m_{k+1}(i_+)}{\sqrt{4\beta_{k+1}^2+m_{k+1}^2(i_+)}} & \text{if } m_{k+1}(i_+)\omega_{i_+,k+1} < m_k(i)\omega_{ik}, \\[2ex] \dfrac{2}{\omega_{i_+,k+1}}\dfrac{m_k(i)}{\sqrt{4\beta_{k+1}^2+m_k^2(i)}} & \text{otherwise.} \end{cases}$$

Whenever one of these four expressions dominates the others, then it gives an accurate estimate of $\max|\hat{\alpha}_k'|$ *in the neighbourhood of* $\lambda_i^{(k)}$*. Otherwise, the maximum is given by a more complicated expression and it is smaller.*

Proof of (i). From §2.2 and also from (4.3),

$$\begin{aligned} \hat{\beta}_n' &= s_n\pi_n' + \pi_n s_n', \\ \hat{\beta}_n'' &= s_n\pi_n'' + s_n'\pi_n' + s_n''\pi_n. \end{aligned}$$

From Lemmas 3.1 and 3.2 and their corollaries,

$$\pi_n(\lambda_i^{(n)}) = 0 = \pi_n''(\lambda_i^{(n)}), \qquad \pi_n'(\lambda_i^{(n)}) = \pm 1/\omega_{in}, \qquad c_n(\lambda_i^{(n)}) = \pm\omega_{in},$$

and so,

$$|\hat{\beta}_n'(\lambda_i^{(n)})| = |s_n\pi_n'| = \left|\frac{s_n}{c_n}\right| = \sqrt{\omega_{in}^{-2} - 1}. \tag{4.13}$$

Using (4.12) gives

$$|\hat{\beta}_n''(\lambda_i^{(n)})| = 2|s_n'\pi_n'| = 2s_n\frac{|\pi_{n-1}'|}{\xi_{n-1}} \neq 0, \tag{4.14}$$

so the maximum does not occur at $\lambda_i^{(n)}$. For $\sigma \neq \lambda_i^{(n)}$ use (4.12) again to find

$$|\hat{\beta}_n'| = |s_n\pi_n' - \pi_n s_n c_n \pi_{n-1}'/\xi_{n-1}| = |s_n(\pi_n' - \pi_n\phi_{n-1})|,$$

where ϕ_{n-1} was defined in (3.18) and then modelled near $\lambda_{i_-}^{(n-1)}$ by $f_{n-1}^{(i_-)}$.

Although $\max|\phi_{n-1}| \approx 1/(2\beta_n\omega_{i,n-1})$ can be huge, we shall show that $|\hat{\beta}_n'|$ takes its local maximum near $(\lambda_i^{(n)} + \lambda_{i_-}^{(n-1)})/2$. We now analyze the model of the two factors in $\hat{\beta}_n'$.

It is convenient to abbreviate temporarily by

$$\Delta := \lambda_{i_-}^{(n-1)} - \lambda_i^{(n)}, \quad \delta := \sigma - \lambda_i^{(n)}, \quad \rho_- := \beta_n\omega_{i_-,n-1}.$$

Recall that π_{n-1} is modelled by

$$p_{n-1}^{(i_-)}(\sigma) = \frac{\pm(\sigma - \lambda_{i_-}^{(n-1)})}{\omega_{i_-,n-1}},$$

and hence $s_n = \beta_n/\sqrt{\pi_{n-1}^2 + \beta_n^2}$ is modelled by

$$\frac{\beta_n\omega_{i_-,n-1}}{\sqrt{(\delta-\Delta)^2 + \beta_n^2\omega_{i_-,n-1}^2}} = \frac{1}{\sqrt{1 + \frac{(\delta-\Delta)^2}{\rho_-^2}}}$$

on the interval of width $\omega_{i_-,n-1}m_{n-1}(i_-)$ centered at $\lambda_{i_-}^{(n-1)}$. The other factor, $\pi_n' - \pi_n\phi_{n-1}$, is modelled by $p_n^{(i)'} - p_n^{(i)}f_{n-1}^{(i_-)}$ where

$$p_n^{(i)} = \frac{(-1)^{n-i}\delta}{\omega_{\text{in}}} \quad \text{and} \quad f_{n-1}^{(i_-)} = \frac{(\delta-\Delta)}{\rho_-^2 + (\delta-\Delta)^2},$$

from (3.18) and (3.21), on the appropriate intervals centered at $\lambda_i^{(n)}$ and $\lambda_{i_-}^{(n-1)}$. Thus

$$p_n^{(i)'} - p_n^{(i)}f_{n-1}^{(i)} = \frac{(-1)^{n-i}}{\omega_{in}}\left[1 + \frac{\delta(\Delta-\delta)}{\rho_-^2 + (\Delta-\delta)^2}\right].$$

The important points are that the quantity [*] exceeds 1 only when $\delta(\Delta-\delta) > 0$, i.e., when $\sigma \in [\lambda_i^{(n)}, \lambda_i^{(n-1)}]$ and the maximum may be evaluated exactly by calculus. The details are omitted but the result is

$$\max_\delta |p_n^{(i)'} - p_n^{(i)}f_{n-1}^{(i_-)}| = \frac{1}{\omega_{in}}\frac{1}{2}\left(1 + \sqrt{1 + \left(\frac{\Delta}{\rho_-}\right)^2}\right)$$

and

$$\Delta - \delta = \frac{\Delta}{1 + \sqrt{1 + (\frac{\Delta}{\rho_-})^2}}.$$

The hypothesis on Δ gives

$$\frac{\Delta}{\rho_-} < \frac{100(\lambda_n^{(n)} - \lambda_1^{(n)})\omega_{in}^2}{\rho_-} = O(\omega_{in}),$$

and indicates that $|\delta| \approx |\Delta - \delta| \approx \Delta/2$, very close to the center of both intervals. Using this value in the model for s_n yields a maximum value for the model for $\hat{\beta}_n'$:

$$\frac{1}{2\omega_{in}} \frac{1+\sqrt{1+(\frac{\Delta}{\rho_-})^2}}{\sqrt{1+\frac{1}{4}(\frac{\Delta}{\rho_-})^2}} = \frac{1}{\omega_{in}}\left[1+\frac{1}{8}\left(\frac{\Delta}{\rho_-}\right)^2+\cdots\right].$$

However, reference to (3.18) and (3.21) shows that our model for $|\pi_n' - \pi_n\phi_{n-1}|$ carries the correction factor

$$1-\frac{3}{2}\left(\frac{\delta}{\omega_{in}m_n(i)}\right)^2 + O(\delta^3).$$

Using $|\delta| = \Delta/2$ here and combining the estimates gives

$$\max|\hat{\beta}_n'| = \frac{1}{\omega_{in}}\left[1+\frac{1}{8}\left(\frac{\Delta}{\rho_-}\right)^2 - \frac{3}{8}\left(\frac{\Delta}{\omega_{in}m_n(i)}\right)^2 + O(\Delta^3)\right]. \qquad \square$$

Proof of (ii). For all σ and $2 \le k < n$,

$$\hat{\beta}_k' = \hat{\beta}_k(\phi_k - \phi_{k-1}) = \beta_k\frac{\xi_k}{\xi_{k-1}}(\phi_k - \phi_{k-1}),$$

where

$$\xi_k = \sqrt{\pi_k^2 + \beta_{k+1}^2} \quad \text{(see (2.9))},$$

$$\phi_k := \frac{\pi_k\pi_k'}{\xi_k^2} :\approx f_k^{(i)} = \frac{\sigma - \lambda_i^{(k)}}{\beta_{k+1}^2\omega_{ik}^2 + (\sigma - \lambda_i^{(k)})^2} \quad \text{(see (3.21))}.$$

The interval associated with $f_k^{(i)}$ is

$$[\lambda_i^{(k)} - m_k(i)\omega_{ik}/2, \lambda_i^{(k)} + m_k(i)\omega_{ik}/2].$$

To analyze the model for $\hat{\beta}_k'$ it is convenient to abbreviate as follows:

$$\delta := \sigma - \lambda_i^{(k)}, \quad \Delta := \lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}, \quad \rho_k := \beta_{k+1}\omega_{ik}, \quad \rho_{k-1} := \beta_k\omega_{i_-,k-1}.$$

$\hat{\beta}_k'$ is modelled by a difference of two functions:

$$\hat{\beta}_k\left[f_k^{(i)} - f_{k-1}^{(i_-)}\right].$$

We consider $\beta_k \xi_k f_k^{(i)}/\xi_{k-1}$ first. By calculus we find that if $\beta_{k+1} < m_k(i)/2$, then $|f_k^{(i)}|$ attains its global maximum value $1/2\rho_k$ at $\delta = \pm\rho_k$ within the associated interval. At these points,

$$\frac{\xi_k}{\xi_{k-1}} \approx \left[\frac{\rho_k^2/\omega_{ik}^2 + \beta_{k+1}^2}{(\rho_k - \Delta)^2/\omega_{i_-,k-1}^2 + \beta_k^2}\right]^{1/2}$$
$$= \frac{\beta_{k+1}}{\beta_k}\left[\frac{2}{1 + (\rho_k/\rho_{k-1})^2}\right]^{1/2} [1 + O(\omega_{ik})].$$

Next consider $\beta_k \xi_k f_{k-1}^{(i_-)}/\xi_{k-1}$. By calculus we find that if $\beta_k < m_{k-1}(i_-)/2$, then $|f_{k-1}^{(i_-)}|$ attains its global maximum value $1/2\rho_{k-1}$, within its associated interval, at $\delta - \Delta = \pm\rho_{k-1}$. At these points,

$$\frac{\xi_k}{\xi_{k-1}} \approx \left[\frac{(\rho_{k-1} \pm \Delta)^2/\omega_{ik}^2 + \beta_{k+1}^2}{\rho_{k-1}^2/\omega_{i_-,k-1}^2 + \beta_k^2}\right]^{1/2}$$
$$= \frac{\beta_{k\pm1}}{\beta_k}\left[\frac{1 + (\rho_{k-1}/\rho_k)^2}{2}\right]^{1/2} [1 + O(\omega_{ik})].$$

If, however, $\beta_{k+1} > m_k(i)/2$, then the maximum of $|f_k^{(i)}|$ on its associated interval is approximately $m_k(i)/(2\beta_{k+1})$ of its global maximum $1/(2\rho_k)$ and similarly for $|f_{k-1}^{(i_-)}|$ when $\beta_k > m_{k-1}(i_-)/2$. We will not consider these two cases further because they yield more complicated expressions and smaller maxima.

Returning to the situation when $|f_k^{(i)}|$ and $|f_{k-1}^{(i_-)}|$ attain their global maxima within their appropriate intervals, as shown in Fig. 2, we distinguish two cases.

(a) $\rho_k = \beta_{k+1}\omega_{ik} < \rho_{k-1} = \beta_k \omega_{i_-,k-1}$. The first expression shows that

$$\left|\frac{\xi_k}{\xi_{k-1}}\right|_{\delta=\rho_k} > \frac{\beta_{k+1}}{\beta_k}(1 + O(\omega_{ik})).$$

Thus the first term's model $\hat\beta_k f_k^{(i)}$ satisfies

$$|\hat\beta_k f_k^{(i)}(\lambda_i^{(k)} \pm \rho_k)| > \beta_k \frac{\beta_{k+1}}{\beta_k}\frac{1}{2\rho_k}(1 + O(\omega_{ik})) = \frac{1}{2\omega_{ik}}(1 + O(\omega_{ik})).$$

(b) $\rho_{k-1} < \rho_k$. The second expression shows that

$$\left|\frac{\xi_k}{\xi_{k-1}}\right|_{|\delta-\Delta|=\rho_{k-1}} > \frac{\beta_{k+1}}{\sqrt{2}\beta_k}(1 + O(\omega_{ik})).$$

Thus the second term's model $\hat\beta_k f_{k-1}^{(i_-)}$ satisfies

$$|\hat\beta_k f_{k-1}^{(i_-)}(\lambda_i^{(k)} - \Delta \pm \rho_{k-1})| > \beta_k \frac{\beta_{k+1}}{\sqrt{2}\beta_k}\frac{1}{2\rho_{k-1}}(1 + O(\omega_{ik}))$$
$$= \frac{\beta_{k+1}}{\sqrt{2}\beta_k}\frac{1}{2\omega_{i_-,k-1}}(1 + O(\omega_{ik})).$$

When one of the two terms derived in cases (a) and (b) dominates the other (by at least a factor of 2) then the larger term gives an accurate enough estimate of the maximum magnitude of the model and of $\hat{\beta}_k'$. By the assumption $\Delta := \lambda_{i_-}^{(k-1)} - \lambda_i^{(k)} = O(\omega_{ik}^2)$, $f_k^{(i)}$ and $f_{k-1}^{(i_-)}$ have the same sign at their global maxima and so cause cancellation in the expression for $\hat{\beta}_k'$. When any of our conditions fail, either $\beta_{k+1} > m_k(i)/2$ or $\beta_k > m_{k-1}(i_-)/2$, or when $\rho_{k-1} \approx \rho_k$, then the f functions have smaller values and/or there is a significant cancellation between them. In such cases $|\hat{\beta}_k'|$ need not be

$$O\left(\frac{1}{\min\left\{\omega_{ik}, \omega_{i_-,k-1}\right\}}\right) \quad \text{near } \lambda_i^{(k)}. \qquad \square$$

Remark. Figures 2 and 3 illustrate the preceding analysis. The matrix is W_{17}^-. $k = 8$, $i = 5$. Then $k - 1 = 7$ and $i_- = 4$ by its definition in (*) preceding Theorem 4.2. Note that $\omega_{ik} = 0.032$, $\omega_{i_-,k-1} = 0.116$, $\rho_i = \beta_{k+1}\omega_{ik}/\beta_k\omega_{i_-,k-1} = 0.27$, $m_k(i) = 3.337$, $m_{k-1}(i_-) = 2.31$, $|\Delta| = 0.0037$. Since $\beta_{k+1} = \beta_k = 1$, the chosen evaluation points are within the domain of the model. The first graph shows ϕ_k and ϕ_{k-1} and the second shows $\hat{\beta}_k'$ and the model

$$\beta_k \left(\frac{\beta_{k+1}^2 + \delta^2/\omega_{ik}^2}{\beta_k^2 + (\delta - \Delta)^2/\omega_{i_-,k-1}^2}\right)^{1/2} (f_k^{(i)} - f_{k-1}^{(i_-)}). \tag{4.15}$$

Proof of (iii). From the proof of Theorem 4.1 we know that $\hat{\alpha}_n' = 2(\pi_n' c_n + 1) = 2\pi_n c_n'$. Thus $\hat{\alpha}_n'(\lambda_{i_-}^{(n-1)})$=2 since c_n vanishes on the spectrum of T_{n-1}, and $\hat{\alpha}_n'(\lambda_i^{(n)})$=0 since π_n vanishes on the spectrum of T_n. However, $|\hat{\alpha}_n'|$ can attain large values when σ is close to these points but *not* between them. The model $p_n^{(i)}$ for π_n is valid when $|\sigma - \lambda_i^{(n)}| < m_n(i)\omega_{in}/2$ and the model

$$p_{n-1}^{(i-1)}/\sqrt{\beta_n^2 + (p_{n-1}^{(i-1)})^2}$$

for c_n is valid for $|\delta| := |\sigma - \lambda_{i_-}^{(n-1)}| < m_{n-1}(i_-)\omega_{i_-,n-1}/2$. On the intersection of these intervals the model for $\hat{\alpha}_n'$ is

$$\begin{aligned} 2\left(p_n^{(i)\prime} p_{n-1}^{(i-1)} \Big/ \sqrt{\beta_n^2 + (p_{n-1}^{(i-1)})^2} + 1\right) &= 2\left[\frac{(-1)^i}{\omega_{in}} \frac{(-1)^{i_-}\delta}{\omega_{i_-,n-1}} \Big/ \frac{(\beta_n^2\omega_{i_-,n-1}^2 + \delta^2)^{1/2}}{\omega_{i_-,n-1}} + 1\right] \\ &= 2\left[\frac{(-1)^{i+i_-}\delta}{\omega_{in}(\beta_n^2\omega_{i_-,n-1}^2 + \delta^2)^{1/2}} + 1\right]. \end{aligned}$$

The function $\delta/(\tau^2 + \delta^2)^{1/2}$ is a monotone-increasing function for all δ, and so the model takes its maximum magnitude at the boundary of the associated interval, namely,

$$2\left[\frac{m_{n-1}(i_-)}{\omega_{in}(4\beta_n^2 + m_{n-1}^2(i_-))^{1/2}} + 1\right] \quad \text{if } m_{n-1}(i_-)\omega_{i_-,n-1} < m_n(i)\omega_{in},$$

$$2\left[\frac{m_n(i)}{\omega_{i_-,n-1}(4\beta_n^2 + m_n^2(i))^{1/2}} + 1\right] \quad \text{otherwise}.$$

We evaluate at the left end if $i + i_-$ is odd. The evaluation point

$$\sigma = \pm\frac{1}{2}m_n(i)\omega_{in} - \lambda_{i_-}^{(n-1)} = \pm\frac{1}{2}m_n(i)\omega_{in} - \lambda_i^{(n)} + [\lambda_i^{(n)} - \lambda_{i_-}^{(n-1)}]$$

is just outside the standard interval in some cases but the excess is $O(\omega_{in}^2)$ and permits a cleaner expression for the model. □

Proof of (iv). From (4.0) and (4.2),

$$\begin{aligned}
\hat{\alpha}_k' &= \gamma_k' - \gamma_{k+1}' \\
&= \pi_k c_k' + \pi_k' c_k - (\pi_{k+1} c_{k+1}' + \pi_{k+1}' c_{k+1}) \\
&= \begin{cases} 2\pi_k c_k' - 2\pi_{k+1} c_{k+1}' \\ 2\pi_k' c_k - 2\pi_{k+1}' c_{k+1} \end{cases} \quad \text{by (4.5)} \\
&= 2\left(\frac{\pi_k' \pi_{k-1}}{\xi_{k-1}} - \frac{\pi_{k+1}' \pi_k}{\xi_k}\right) \quad \text{by (2.10).}
\end{aligned}$$

Near $\lambda_i^{(k)}$, $\hat{\alpha}_k'$ is modelled by

$$\frac{2p_k^{(i)'} p_{k-1}^{(i_-)}}{(\beta_k^2 + (p_{k-1}^{(i_-)})^2)^{1/2}} - \frac{2p_{k+1}^{(i^+)'} p_k^{(i)}}{(\beta_{k+1}^2 + (p_k^{(i)})^2)^{1/2}}.$$

Here is the difference of two terms of the type analyzed in the proof of (iii). Each term is monotonic on the appropriate interval. The first term is valid on the intersection of the intervals

$$\begin{aligned}
&\text{center } \lambda_i^{(k)}, && \text{radius } \tau_k := \tfrac{1}{2} m_k(i) \omega_{ik}, \\
&\text{center } \lambda_{i_-}^{(k-1)}, && \text{radius } \tau_{k-1} := \tfrac{1}{2} m_{k-1}(i_-) \omega_{i_-,k-1},
\end{aligned}$$

and the second term is valid on the intervals

$$\begin{aligned}
&\text{center } \lambda_i^{(k)}, && \text{radius } \tau_k := \tfrac{1}{2} m_k(i) \omega_{ik}, \\
&\text{center } \lambda_{i_+}^{(k+1)}, && \text{radius } \tau_{k+1} := \tfrac{1}{2} m_{k+1}(i_+) \omega_{i_+,k+1}.
\end{aligned}$$

From thc analysis of $\hat{\alpha}_n'$, the magnitude of the first term achieves

$$\begin{cases} \dfrac{2}{\omega_{i_-,k-1}} \dfrac{m_k(i)}{(4\beta_k^2 + m_k^2(i))^{1/2}} & \text{if } \tau_k < \tau_{k-1}, \\[2ex] \dfrac{2}{\omega_{ik}} \dfrac{m_{k-1}(i_-)}{(4\beta_k^2 + m_{k-1}^2(i_-))^{1/2}} & \text{if } \tau_{k-1} < \tau_k. \end{cases}$$

Similarly, the second term achieves

$$\begin{cases} \dfrac{2}{\omega_{ik}} \dfrac{m_{k+1}(i_+)}{(4\beta_{k+1}^2 + m_{k+1}^2(i_+))^{1/2}} & \text{if } \tau_{k+1} < \tau_k, \\[2ex] \dfrac{2}{\omega_{i_+,k+1}} \dfrac{m_k(i)}{(4\beta_{k+1}^2 + m_k^2(i))^{1/2}} & \text{if } \tau_k < \tau_{k+1}. \end{cases}$$ □

Before we leave this section, we give Theorem 4.4, whose proof appears in [3].

THEOREM 4.4. *Suppose that T is unreduced. Then*

$$\omega_{ik}^2 < \begin{cases} \dfrac{\lambda_1^{(k-1)} - \lambda_1^{(k)}}{\lambda_2^{(k)} - \lambda_1^{(k)}}, & i = 1, \\[2ex] \dfrac{\lambda_i^{(k)} - \lambda_{i-1}^{(k-1)}}{\lambda_i^{(k)} - \lambda_{i-1}^{(k)}} \cdot \dfrac{\lambda_i^{(k-1)} - \lambda_i^{(k)}}{\lambda_{i+1}^{(k)} - \lambda_i^{(k)}}, & i \neq 1, k, \\[2ex] \dfrac{\lambda_k^{(k)} - \lambda_{k-1}^{(k-1)}}{\lambda_k^{(k)} - \lambda_{k-1}^{(k)}}, & i = k, \end{cases}$$

and

$$\omega_{ik}^2 > \begin{cases} \dfrac{\lambda_1^{(k-1)} - \lambda_1^{(k)}}{\lambda_k^{(k)} - \lambda_1^{(k)}}, & i = 1, \\[2ex] \dfrac{\lambda_i^{(k)} - \lambda_{i-1}^{(k-1)}}{\lambda_i^{(k)} - \lambda_1^{(k)}} \cdot \dfrac{\lambda_i^{(k-1)} - \lambda_i^{(k)}}{\lambda_k^{(k)} - \lambda_i^{(k)}}, & i \neq 1, k, \\[2ex] \dfrac{\lambda_k^{(k)} - \lambda_{k-1}^{(k-1)}}{\lambda_k^{(k)} - \lambda_1^{(k)}}, & i = k. \end{cases}$$

Proof. See [3]. □

5. Premature deflation and the monitoring algorithm.

The reader is referred to (2.17) for a picture of the active submatrix of the intermediate matrix $T^{(k)}$ that occurs in the transformation of $T = T^{(1)}$ to $\hat{T} = T^{(n)}$.

Observe that if $s_k \pi_k$ and $\beta_{k+1} c_k$ are replaced by zero to give $\dot{T}^{(k)}$, then $\dot{T}^{(k)}$ must exhibit an eigenvalue in the (k, k) position, and if row and column k are deleted, then the new $(n-1) \times (n-1)$ matrix is tridiagonal.

When $s_k \pi_k$ and $\beta_{k+1} c_k$ are small enough and the (k, k) entry equals the shift σ then we say that *premature deflation* occurred at step k in the implicit shift version of the QR transform.

The next result implies that a negligible value of some ω_{ik}^2 (compared to 1) is a necessary condition for premature deflation.

LEMMA 5.1. *On the interval* $\left[\lambda_i^{(k)}, \lambda_{i_-}^{(k-1)}\right]$

$$s_k^2 \pi_k^2 + \beta_{k+1}^2 c_k^2 \geq 2 \left(\frac{\beta_{k+1}^2 \omega_{ik}^2}{\beta_k^2 \omega_{i_-,k-1}^2 + \beta_{k+1}^2 \omega_{i,k}^2} \right) \left(\Psi_i^{(k)} \omega_{ik} \right)^2 \left(1 + O\left[\left(\frac{\Delta_i}{\beta_k \omega_{i_-,k-1}} \right)^2 \right] \right),$$

where

$$\Psi_i^{(k)} = \begin{cases} \lambda_2^{(k)} - \lambda_1^{(k)}, & i = 1, \\[1ex] \dfrac{\left(\lambda_{i+1}^{(k)} - \lambda_i^{(k)}\right)\left(\lambda_i^{(k)} - \lambda_{i-1}^{(k)}\right)}{|\lambda_i^{(k)} - \lambda_i^{(k-1)}|}, & i \neq 1, k, \\[1ex] \lambda_k^{(k)} - \lambda_{k-1}^{(k)}, & i = k; \end{cases}$$

$$\Delta_i^{(k)} = |\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}|.$$

Proof. Since $\Delta_i^{(k)} = O(\omega_{ik}^2)$ the linear models for π_k and π_{k-1} are valid. Write $\delta = \sigma - \lambda_i^{(k)}, \omega = \omega_{ik}, \tilde{\omega} = \omega_{i_-,k-1}$, and observe that

$$\begin{aligned} s_k^2\pi_k^2 + \beta_{k+1}^2 c_k^2 &= \frac{\frac{\beta_k^2\delta^2}{\omega^2} + \frac{\beta_{k+1}^2(\tilde{\delta}-\Delta)^2}{\omega}^{2}}{\beta_k^2 + (\delta-\Delta)^2/\tilde{\omega}^2} \\ &= \frac{\tilde{\omega}^2\beta_k^2\delta^2 + \omega^2\beta_{k+1}^2(\Delta-\delta)^2}{\beta_k^2\omega^2\tilde{\omega}^2}\left(1 + O\left(\frac{\Delta}{\beta_k\tilde{\omega}}\right)^2\right) \\ &> \frac{2\tilde{\omega}^2\beta_k^2\omega^2\beta_{k+1}^2\Delta^2}{(\tilde{\omega}^2\beta_k^2 + \omega^2\beta_{k+1}^2)\beta_k^2\omega^2\tilde{\omega}^2}\left(1 + O\left(\frac{\Delta}{\beta_k\tilde{\omega}}\right)^2\right) \\ &\quad \text{(the minimum for } \delta \in [0,\Delta]) \\ &= 2\frac{\beta_{k+1}^2}{(\tilde{\omega}^2\beta_k^2 + \omega^2\beta_{k+1}^2)}\Delta^2\left(1 + O\left(\frac{\Delta}{\beta_k\tilde{\omega}}\right)^2\right). \end{aligned}$$

By Theorem 4.4, $\Delta_i^{(k)} > \Psi_i^{(k)}\omega_{ik}^2$, the result follows. Note that the term ω_{ik}^4 has been rearranged in the lemma's statement. □

Note that at $\sigma = \lambda_i^{(k)}$,

$$s_k^2\pi_k^2 + \beta_{k+1}^2 c_k^2 = \beta_{k+1}^2\omega_{ik}^2$$

since $\pi_k = 0$, and outside $[\lambda_i^{(k)}, \lambda_{i_-}^{(k-1)}]$ the sum rises rapidly to $O(\text{spread}_k^2)$.

COROLLARY. *Premature deflation occurs if and only if σ is very close to $\lambda_i^{(k)}$ and ω_{ik}^2 is negligible. By neglecting the entries $s_k\pi_k$ and $\beta_{k+1}c_k$ and deleting the kth row and column the changes made in the eigenvalues are bounded by $(s_k^2\pi_k^2 + \beta_{k+1}^2c_k^2)/\min(|\sigma-\lambda_{i-1}^{(k)}|, |\sigma-\lambda_{i+1}^{(k)}|)$ and the error angles in the eigenvectors are bounded by $(s_k^2\pi_k^2 + \beta_{k+1}^2c_k^2)^{1/2}/\min(|\sigma - \lambda_{i-1}^{(k)}|, |\sigma - \lambda_{i+1}^{(k)}|)$.*

Thus premature deflation at step k accompanies great sensitivity of $\hat{\beta}_k, \hat{\beta}_{k+1}$, and $\hat{\alpha}_{k+1}$ to small changes in σ. In many, but not all, cases a tiny value of ω_{ik} will be associated with a tiny value of $\omega_{i_+,k+1}$.

The pattern is as follows. If ω_{ik} is tiny then $y_k^{(n)}$, T_k's eigenvector for $\lambda_i^{(k)}$ with zero entries appended, will be an excellent approximate eigenvector for all extensions to T_k. The only way that subsequent values ω_{in} can increase is by the presence of two or more eigenvalues of T_n close to $\lambda_i^{(k)}$. In such a case, $y_k^{(n)}$ is a linear combination of the eigenvectors belonging to all $\lambda_j^{(n)}$ close to $\lambda_i^{(k)}$ but not close to any of others. In fact, $y_k^{(n)}$ is usually close to a bisector of a pair of eigenvectors, and their last entries will be nearly equal and not necessarily small.

Here is an illustration of this phenomenon.

Example 5.1.

$$T = \text{tridiag}\begin{pmatrix} & 1 & 1\cdots 1 & 1 & \\ 13 & 0 & \cdots & 0 & 13 \\ & 1 & 1\cdots 1 & 1 & \end{pmatrix},$$

$$\alpha_i = 0, (i = 2, \ldots, 24), \quad \alpha_1 = \alpha_{25} = 13, \quad \beta_i = 1, (i = 2, \ldots, 25).$$

This matrix has two close eigenvalues. The instability occurs midway through the QR transform with shift at one of these two eigenvalues. Rather than exhibit matrices of

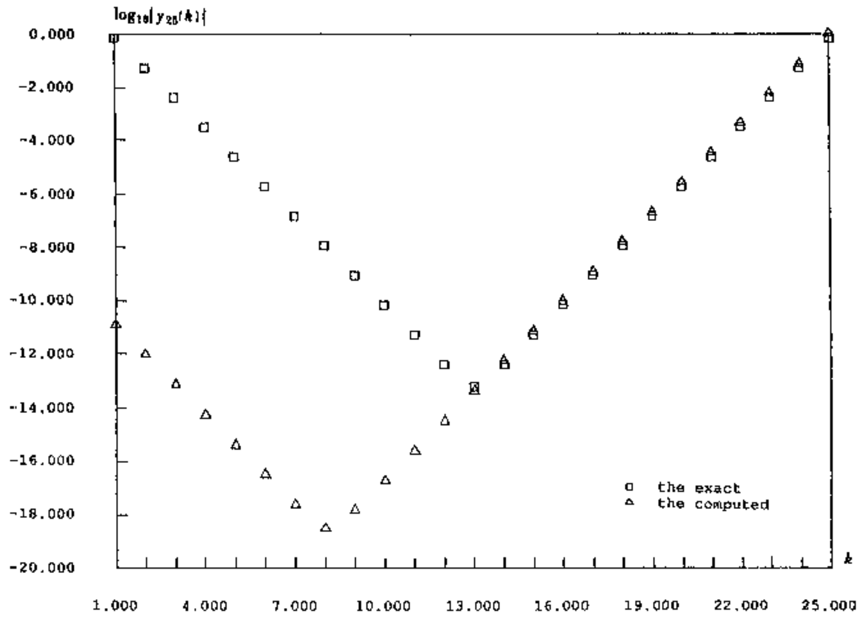

FIG. 8. *Vector components of* y_{25} *defined in* (2.2), *in Example* 5.1.

this size, we plot in Fig. 8 the exact and the computed vectors y_{25} on a logarithmic scale. At first sight it is surprising that the top half of y is quite wrong, while the lower half seems reasonable. However, reference to (2.2) reveals that late bad values of s all propagate to the top of y. Further inspection of the exact and computed c and s values shows that the first eight are good. Then instability sets in, but the last two rows are only wrong by a factor of 2. So the lower half of y only looks reasonable on a logarithmic scale; it has barely one bit of accuracy.

5.1. Forward stability for suitable shift strategies. Our analysis has focused on the role of the numbers ω_{ik} in one QR transform. In the QR algorithm a sequence of QR transformations is applied to the original matrix with carefully chosen shifts. At each transformation the eigenvectors change and the ω_{ik}'s change along with them. In the last two steps of the algorithm the last off-diagonal entry diminishes rapidly. To see the effect on the ω_{in} it suffices to consider an extreme case. Suppose that $\beta_n = 0$ but the algorithm does not notice this fact. Any reasonable shift strategy will force $\sigma = \alpha_n$ in this case, and the associated eigenvector is e_n, with its $\omega_{in} = 1$. All the other values of ω_{in} vanish since $\sum_i \omega_{in}^2 = 1$. The analysis of the previous section shows that the QR transform with shift α_n is stable, but for any shift at an eigenvalue of T_{n-1} (not T_n), the last plane rotation is arbitrary and $\hat{\beta}_n$ is undefined. In this case, the QR algorithm with any sensible shift strategy is forward stable, although the QR transform with "wrong" shifts is completely unstable.

However, forward instability can still occur in the standard QR algorithm, but only when the shift is very close to a cluster of eigenvalues equal to working precision. This was shown in Example 2.4 of §2.3.

5.2. The ultimate shift strategy. The QR transformation of an $n \times n$ real symmetric tridiagonal matrix requires about $10n$ flops. The most expensive feature of a QR algorithm that produces eigenvectors is the accumulation of all the plane rotations. This is an $O(n^2)$ process for each transform, and the cost is directly proportional to the number of QR transformations.

These facts suggested the use of a two-phase process. First, compute the eigenvalues by any means, keeping a copy of the original tridiagonal. Second, apply the QR transformation using the eigenvalues as shifts (see [4, p. 164]).

Our analysis shows that this strategy invites forward instability. In fact, it will occur for each eigenvalue whose normalized eigenvector has a tiny bottom element.

This is the first reason we have seen for being cautious about the use of the ultimate shift strategy. However, the simple modification of the QR transform given below can preserve forward stability.

5.3. QR with monitoring. The idea of the algorithm is to check for premature deflation and stop the QR transformation as soon as it occurs. Then remove the row and column with the isolated eigenvalue. This is not a pure QR algorithm, but it does use plane rotations and does restore tridiagonal form.

Since the monitoring test usually fails, it is preferable to break it into two parts so that the part that is always made involves no arithmetic operations.

```
If $|\pi_k| < \sqrt{\epsilon}\|T\|/\sqrt{n}$ then
    If $|\pi_k| + \beta_{k+1}|c_k| < \sqrt{\epsilon}\,(|\alpha_{k-1}| + |\alpha_k| + |\alpha_{k+1}| + \beta_{k-1} + \beta_k)$ then
        pull up the remaining entries of $T$ to overwrite row and column $k$.
```

Such an algorithm has been used regularly for deflation purposes in the context of the Lanczos algorithm. Since a copy of the undeflated T matrix is preserved for the eigenvector computations at the end of the Lanczos run there is no harm in suppressing off-diagonal entries as large as $\sqrt{\epsilon}\|T\|$ in the deflated T. By the Corollary to Lemma 5.1, such modifications could indeed induce a $\sqrt{\epsilon}$ twist in T's eigenvectors. Whether such alterations damage the Ritz vectors (the approximate eigenvectors of the operator driving the Lanczos algorithm) is not clear.

6. Conclusion. The occasional forward instability of the QR transformation is not a well-known phenomenon and so we have tried to give a full account of it. In particular, we have shown its intimate connection with premature deflation of the shift and with the quantities ω_{ik}. Since the $\{\omega_{ik}\}$ change at each step in the QR algorithm, because the eigenvectors change, forward instability is unlikely to occur, although, by Example 2.4, it can happen.

On the other hand, forward instability is seen quite frequently when deflating stabilized eigenvalues in the Lanczos algorithm and sometimes when using the ultimate shift strategy with the QR algorithm.

Our analysis of the function $\pi_k(\sigma)$ and of the derivative of the tridiagonal QR transform are full of details that will burden the reader. To put this aspect in perspective, we would like to say that the entries in $\hat{T}$ are complicated functions and it is far from obvious where their derivatives peak. We could see no other way than direct elucidation of $\max_\sigma |\hat{\beta}_k'|$ and $\max_\sigma |\hat{\alpha}_k'|$ to establish that small values of some ω_{ik} are generically sufficient as well as necessary for the occurrence of forward instability.

Acknowledgement. The authors are grateful for the extremely careful and helpful reviewing by the two referees and the editor Linda Kaufman throughout several revisions of this work.

REFERENCES

[1] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[2] W. B. GRAGG AND W. J. HARROD, *The numerically stable reconstruction of Jacobi matrices from spectral data*, Numer. Math., 44 (1984), pp. 317–336.

[3] R. O. HILL AND B. N. PARLETT, *Refined interlacing properties*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 239–247.

[4] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[5] B. T. SMITH, J. M. BOYLE, B. S. GARBOW, Y. IKEBE, V. C. KLEMA, AND C. B. MOLER, *Matrix Eigensystem Routines*—EISPACK *Guide*, in Lecture Notes in Computer Science, 2nd Ed., Vol. 6, Springer-Verlag, New York, 1976.

[6] G. W. STEWART, *Incorporating origin shifts into the* QR *algorithm for symmetric tridiagonal matrices*, Comm. Assoc. Comput. Mach., 13 (1970), pp. 365–367.

[7] ———, *Perturbation bounds for the* QR *factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.

[8] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

[9] ———, *The calculation of the eigenvectors of codiagonal matrices*, Comput. J., 1 (1958), pp. 90–96.