OPTIMAL SCALING OF BLOCK COMPANION PENCILS*

D. LEMONNIER AND P. VAN DOOREN†

May 14, 2004

Abstract. Computing roots of a polynomial matrix may be done by computing the generalized eigenvalues of a corresponding block companion pencil using any generalized eigenvalue algorithm. If the algorithm is backward stable, it computes exact roots of a slightly modified polynomial matrix, but it may yield very poor backward errors in the coefficients of the polynomial matrix. In this paper we investigate what can be done to improve these errors using geometric ideas. We show that preconditioning the block companion pencil using a scaling transformation may achieve this goal.

1. Introduction. The polynomial generalized eigenvalue problem consists of finding the roots of the polynomial matrix

$$P(\lambda) \;:=\; P_n\lambda^n + P_{n-1}\lambda^{n-1} + \ldots + P_1\lambda + P_0, \qquad P_i \in \mathbb{C}^{p\times p}$$

which are defined as the zeros of the scalar polynomial

$$p(\lambda) := \det P(\lambda) \;=\; 0\ .$$

These are well-defined provided that $P(\lambda)$ is square and regular, which means that $\det P(\lambda)$ is not identically zero. When $P(\lambda)$ is not regular one can still define zeros via its Smith form but these zeros are then ill-posed [8] and we will not consider this more general case here. For the regular case one often recommends to reduce it to the generalized eigenvalue problem (called a *companion pencil*)

$$\lambda B - A = \lambda \begin{bmatrix} I_p & & & \\ & \ddots & & \\ & & I_p & \\ & & & P_n \end{bmatrix} - \begin{bmatrix} 0 & & & -P_0 \\ I_p & 0 & & -P_1 \\ & \ddots & & \vdots \\ & & I_p & -P_{n-1} \end{bmatrix}. \qquad (1.1)$$

These two problems are equivalent since $\det(\lambda B - A) = \det P(\lambda)$. Finding the roots of $\det(\lambda B - A)$ can be solved via the QZ algorithm [5], which is known to be backward stable. This means that the computed eigenvalues are exactly those of a matrix pencil $\lambda(B+\Delta_B) - (A+\Delta_A)$ for some dense backward error pencil $\Delta(\lambda) = \lambda\Delta_B - \Delta_A$ with

$$\frac{\|\Delta_A\|}{\|A\|} \;=\; O(\epsilon) \quad , \quad \frac{\|\Delta_B\|}{\|B\|} \;=\; O(\epsilon) \qquad (1.2)$$

where $\|\cdot\|$ is some matrix norm and ϵ is the machine accuracy of the computer. This suggests that it is a safe way to solve for the roots of $P(\lambda)$, but it was pointed out in [7] that this is not the case. In [7] a polynomial matrix example of order $n = 2$ is given where small backward errors in the matrices $B,\ A$ of (1.1) lead to quite large backward errors in the coefficients P_i. In [3] it is shown, however, that the backward errors can significantly be reduced by an appropriate scaling of the coefficient matrices.

*This paper presents research supported by the Belgian Programme on Inter-university Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. This work was also supported by the National Science Foundation under Grant No. CCR-9912415.

†Both authors are with CESAME, Université catholique de Louvain, Belgium.

In this paper we analyze the scaling problem of pencils of the type (1.1) called *companion pencils* and we show that scaling the pencil $\lambda B - A$ using invertible left and right diagonal transformations

$$D_2^{-1}(\lambda B - A)D_1$$

can significantly improve the sensitivity of the coefficients. Moreover we describe the class of optimally scaled companion pencils. Finally we show that the optimal scalings can be obtained via the solution of a convex optimization problem. This is an extension of the results previously obtained [4, 2] for the scalar polynomial case.

2. Problem statement. In the scalar case ($p = 1$), every pencil $\lambda(B + \Delta_B) - (A+\Delta_A)$ has a uniquely defined determinant $p(\lambda)+\delta p(\lambda)$, where $\delta p(\lambda) = \delta p_n\lambda^n+\ldots+\delta p_1\lambda + \delta p_0$. The QZ algorithm thus computes generalized eigenvalues of a perturbed pencil $\lambda(B + \Delta_B) - (A + \Delta_A)$ which are also the roots of the perturbed polynomial $p(\lambda) + \delta p(\lambda)$. One would wish that $\|\delta p(\cdot)\| = O(\epsilon)\|p(\cdot)\|$ for some polynomial norm, which would mean backward stability. In [2, 4] this problem was analyzed using geometric methods and in [4] the *polynomial sensitivity*

$$\kappa_{pol} \;:=\; \lim_{\delta\to 0} \sup_{\|\Delta\| \,\le\, \delta} \frac{\|\delta p(\cdot)\|}{\|\Delta(\cdot)\|}$$

has been studied for this problem.

In the present paper we extend these results to the matrix polynomial case ($p > 1$). The bound (1.2) seems to imply that we computed the roots of a perturbed polynomial matrix $P(\lambda) + \delta P(\lambda)$ where the perturbation $\delta P(\lambda) = \delta P_n\lambda^{n-1} + \ldots + \delta P_1\lambda + \delta P_0$ is hoped to be small. We would like to require that the computed eigenvalues are precisely the roots of the determinant of a matrix polynomial $\det(P(\lambda) + \delta P(\lambda))$ where $\|\delta P(\cdot)\| = O(\epsilon)\|P(\cdot)\|$ for some matrix polynomial norm, which would imply backward stability. For $p > 1$ every perturbation $\Delta(\lambda) = \lambda\Delta_B-\Delta_A$ of the pencil does not correspond anymore to a unique perturbation $\delta P(\lambda)$ of the polynomial matrix. There exist now several choices for $\delta P(\cdot)$, and we need to consider the most favorable ones. For this purpose, we define a transformation $\pi(\cdot)$ that maps $\lambda\Delta_B - \Delta_A$ to $\delta P(\lambda)$ via $\delta P(\cdot) = \pi(\Delta(\cdot))$. With appropriately defined norms, we then define the polynomial sensitivity under perturbations of the pair (B, A) as

$$\kappa_{pol} := \min_{\pi} \lim_{\delta\to 0} \sup_{\|\Delta(\cdot)\| \,\le\, \delta} \frac{\|\pi(\Delta(\cdot))\|}{\|\Delta(\cdot)\|} \;. \tag{2.1}$$

In this paper, we concentrate on making κ_{pol} as small as possible under a given class of transformations.

3. The monic matrix case. In this special case, we want to compute the eigenvalues λ_i of the $p\times p$ polynomial matrix $P(\lambda) \;:=\; I_p\lambda^n+P_{n-1}\lambda^{n-1}+\ldots+P_1\lambda+P_0$. For this special case the pencil $\lambda B - A$ reduces to a standard eigenvalue problem $\lambda I_{pn} - C$, where the matrix C is in block companion form

$$C = \begin{bmatrix} 0 & & & -P_0 \\ I_p & 0 & & -P_1 \\ & \ddots & & \vdots \\ & & I_p & -P_{n-1} \end{bmatrix} \tag{3.1}$$

and has as characteristic polynomial $\det(\lambda I - C) = \det P(\lambda)$. We then compute the eigenvalues of $P(\lambda)$ using any eigenvalue algorithm. If the algorithm is backward stable (think for example of the standard QR-algorithm), the computed eigenvalues are exactly those of a matrix $C + \Delta_C$ for some dense backward error matrix Δ_C. We can state that this backward error satisfies

$$\frac{\|\Delta_C\|}{\|C\|} = O(\epsilon) \tag{3.2}$$

where ϵ denotes the machine precision, and $\|\cdot\|$ some matrix norm. However we want to analyze how this perturbation is reflected in a perturbation of $P(\lambda)$. We therefore require that the computed eigenvalues are the roots of the determinant of a matrix polynomial $\det(P(\lambda) + \delta P(\lambda))$. If $\delta P(\lambda) = \delta P_{n-1}\lambda^{n-1} + \ldots + \delta P_1\lambda + \delta P_0$ satisfies $\|\delta P(\cdot)\| = O(\epsilon)\|P(\cdot)\|$ for some matrix polynomial norm, this would imply backward stability in a structured sense. Notice that for a given perturbation Δ_C, $\delta P(\cdot)$ is not unique anymore in contrast with the scalar case treated in [4]. Since there exist several choices for $\delta P(\cdot)$, we consider all transformations $\pi(\cdot)$ that reduce $\lambda I_{pn} - (C + \Delta_C)$ to a polynomial matrix $P(\lambda) + \delta P(\lambda)$ and show that up to first order the map $\delta P(\cdot) = \pi(\Delta_C)$ is a projector. We then define the polynomial sensitivity under perturbations of C as in (2.1) :

$$\kappa_{pol} := \min_{\pi} \lim_{\delta \to 0} \sup_{\|\Delta_C\| \leq \delta} \frac{\|\pi(\Delta_C)\|}{\|\Delta_C\|}. \tag{3.3}$$

In order to analyze this geometrically we consider the manifold of all matrices that are similar to C :

$$\mathbf{Orb} \;:=\; \text{orbit}(C) \;=\; \{T^{-1}CT : \det(T) \neq 0\}$$

imbedded in the euclidian matrix space $\mathbb{C}^{np \times np}$ with the Frobenius inner product $\langle A, B\rangle := \text{tr}(AB^*)$. The tangent space to **Orb** at C is given by (see [2, 4])

$$\mathbf{Tan} \;=\; \{XC - CX : X \;\in\; \mathbb{C}^{np \times np}\},$$

which generically has dimension $n^2p^2 - np$, but it can be smaller. The $p \times p$ block companion matrices (3.1) form an affine space passing through C which we will call

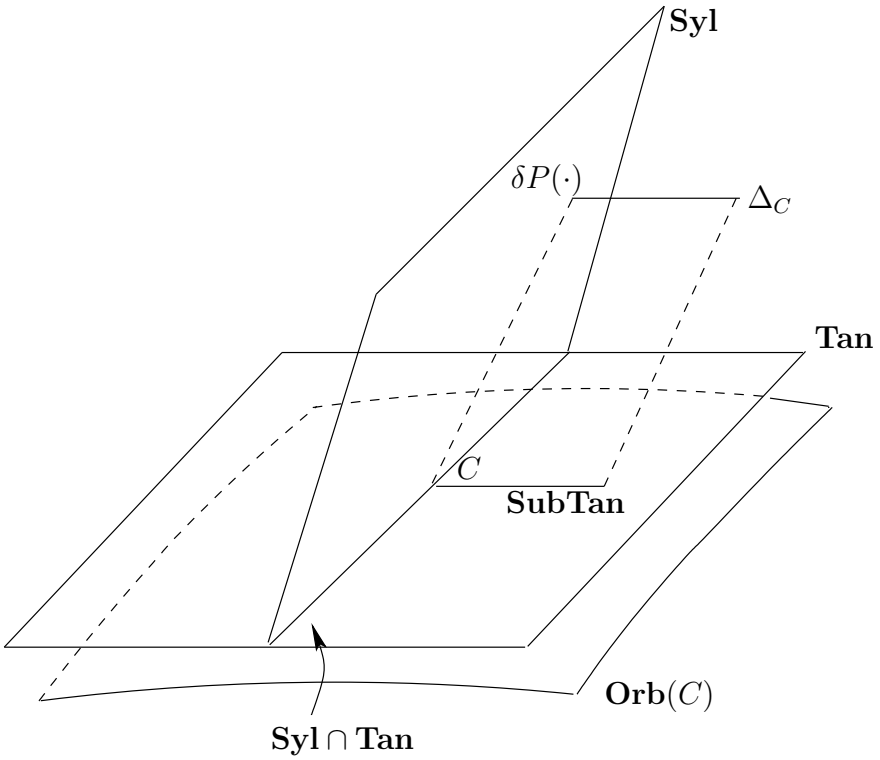
$$\mathbf{Syl} = \{X.[\,0 \;\ldots\; 0\; I_p\,],\; X \in \mathbb{C}^{np \times p}\}$$

and it has dimension np^2. For $p > 1$ we thus have that generically dim **Syl** + dim **Tan** $> n^2p^2$. Both spaces have thus a nontrivial intersection. When we consider an arbitrary perturbed block companion matrix $C + \Delta_C$, it lies on neither of the affine spaces $C + \mathbf{Syl}$ and $C + \mathbf{Tan}$. We thus need to find the closest matrix to $C + \Delta_C$ that lies on $C + \mathbf{Syl}$ and is similar to $C + \Delta_C$. The similarity transformation will be close to the identity and can be approximated by $T = I + X$ with $\|X\|$ of the order of $\|\Delta_C\|$

$$(I + X)(C + \Delta_C)(I + X)^{-1} \in C + \mathbf{Syl}$$

or, after neglecting terms of the order of $\|\Delta_C\|^2$

$$XC - CX + \Delta_C \in \mathbf{Syl}. \tag{3.4}$$



The matrix $XC - CX$ clearly lies in **Tan** or in a subspace **SubTan** of **Tan**. The operation (3.4) is thus a projection of Δ_C on **Syl** parallel to any subspace **SubTan** $\subseteq$ **Tan**. Let us call this projector π then its norm $\|\pi\| := \max_{\Delta_C} \|\delta P(\cdot)\|/\|\Delta_C\|$ is the maximal growth factor one can expect for the perturbation $\delta P(\cdot)$ corresponding to Δ_C. We therefore need to find the projector π of minimal norm. It is well known from geometric considerations that for any *complementary* subspace **SubTan** (i.e. such that **Syl** $\oplus$ **SubTan** $= \mathbb{C}^{n^2p^2}$) the projector on **Syl** and parallel to **SubTan** has norm equal $1/cos(\theta_{min})$, where θ_{min} is the smallest canonical angle between these two subspaces. Moreover for all complimentary subspaces **SubTan** of **Tan** this is also bounded by $1/cos(\theta_{min})$, where θ_{min} is the smallest *nonzero* canonical angle between **Syl** and **Tan**. In the future, we will call this optimal subspace **SubTan**$_{opt}$. In order to describe the corresponding projector we need to find bases of the spaces we consider. For that we represent a matrix X in its vector form vec$X$ (as in the Matlab notation).

Let $\mathcal{X}$ be a basis for **Syl** and $\mathcal{Y}$ be a basis for the orthogonal complement of **SubTan** then π is described by the $n^2p^2 \times n^2p^2$ matrix

$$\Pi := \mathcal{X}(\mathcal{Y}^*\mathcal{X})^{-1}\mathcal{Y}^*.$$

Since matrices in **Syl** have their last np^2 components nonzero, an (orthogonal) basis for **Syl** is given by

$$\mathcal{X} = \begin{bmatrix} O \\ I_{np^2} \end{bmatrix}.$$

The tangent space $\{XC - CX : X \in \mathbb{C}^{np\times np}\}$ is clearly the image of the mapping

$$\mathrm{vec}X \longrightarrow (I \otimes C - C^T \otimes I)\mathrm{vec}X.$$

Its orthogonal complement is then the kernel of the adjoint mapping

$$I \otimes C^* - \overline{C} \otimes I$$

which amounts to the set of matrices commuting with C^*. The orthogonal complement $Im\mathcal{Y}$ of every complimentary subspace **SubTan** is thus spanned by np^2 matrices N_k^* commuting with C^* :

$$C^*N_k^* \;=\; N_k^*C^* \quad k = 1, \ldots, np^2,$$

$$\mathcal{Y} \;=\; \left[\ \mathrm{vec}\ N_1^* \ \middle|\ \ldots\ \middle|\ \mathrm{vec}\ N_{np^2}^*\ \right].$$

Since $\mathcal{X}$ is orthonormal, $\|\Pi\|_2 = \|(\mathcal{Y}^*\mathcal{X})^{-1}\mathcal{Y}^*\|_2$ and if we normalize the vectors N_k such that $\mathcal{Y}^*\mathcal{X} = I_{np^2}$ then we can then write :

$$\kappa_{pol} = \|\Pi\|_2 = \|\left[\ \mathrm{vec}\ N_1^* \ \middle|\ \ldots\ \middle|\ \mathrm{vec}\ N_{np^2}^*\ \right]\|_2.$$

Recall that $\|\Pi\|_2$ also equals $1/\cos\theta_{min}$ where θ_{min} is the minimal canonical angle between **SubTan** and **Syl**. If this angle is small, the sensitivity κ_{pol} can be very large.

An important question arising here is to find a similarity transformation $T^{-1}CT$, i.e. another point on **Orb**, for which this sensitivity is improved or even optimal. Under such a transformation, the spaces $\mathcal{X}$ and $\mathcal{Y}$ are transformed according to

$$\mathcal{Y}_{\mathcal{T}}^* = \mathcal{Y}^*\mathcal{T} \quad , \quad \mathcal{X}_{\mathcal{T}} = \mathcal{T}^{-1}\mathcal{X} \quad , \quad \mathcal{T} := T^{-T} \otimes T.$$

The sensitivity $\kappa_{pol}(T)$ of the matrix $T^{-1}CT$ is therefore given by

$$\kappa_{pol}(T) = \|\mathcal{X}_{\mathcal{T}}(\mathcal{Y}_{\mathcal{T}}^*\mathcal{X}_{\mathcal{T}})^{-1}\mathcal{Y}_{\mathcal{T}}^*\|_2 = \|\mathcal{T}^{-1}\mathcal{X}\mathcal{Y}^*\mathcal{T}\|_2$$

which we need to minimize over T to find the matrix $T^{-1}CT$ with minimal sensitivity. The following theorem is given without proof here.

THEOREM 3.1. *The matrix $T^{-1}CT$ is normal if and only if T minimizes*

$$\min_T \kappa_{pol}(T) = \min_{T,\ \mathcal{Y}_{\mathcal{T}}^*\mathcal{X}_{\mathcal{T}}=I} \|\mathcal{X}_{\mathcal{T}}\mathcal{Y}_{\mathcal{T}}^*\|_2 \ . \tag{3.5}$$

Now one can ask which points are best conditioned when restricting ourselves to *diagonal scaling* transformations, which boils down to solving

$$\min_D \ \kappa_{pol}(D) \ . \tag{3.6}$$

In order to get a grip on this we need an explicit expression for the selection of matrices commuting with C. In [2] a specific subspace of **Tan** is constructed with an explicit basis for its orthogonal complement. The subspace $\mathbf{SubTan}_{subopt}$ considered there consists of all matrices of the form $CX - XC$ where X is arbitrary *except for its*

bottom p rows, which are equal to zero. This subspace is shown to be complementary to the subspace **Syl**. One first defines block matrices $M_k,\ k = 1, \ldots, n$

$$M_k = \overbrace{\begin{bmatrix} P_k & & & \\ P_{k+1} & P_k & & \\ \vdots & P_{k+1} & \ddots & \\ P_n = I_p & \vdots & \ddots & P_k \\ & P_n = I_p & \ddots & P_{k+1} \\ & & \ddots & \vdots \\ & & & I_p \end{bmatrix}}^{k} \overbrace{\begin{bmatrix} -P_0 & & \\ -P_1 & \ddots & \\ \vdots & \ddots & -P_0 \\ -P_{k-1} & \ddots & -P_1 \\ & \ddots & \vdots \\ & & -P_{k-1} \end{bmatrix}}^{n-k} . \tag{3.7}$$

While $M_0 = M_{n+1} = 0$, one easily checks that these matrices satisfy

$$M_k \ = \ CM_{k+1} + I_n \otimes P_k \qquad \text{for } k = 0, \ldots, n.$$

Let us define the *block trace* $\mathrm{tr}_p(A)$ of a $np \times np$ matrix A as the $p \times p$ matrix

$$\mathrm{tr}_p(A) := \sum_{i=1}^{n} A_{ii}.$$

This is used in the proof of the following theorem due to [2].

THEOREM 3.2. *Given the first $n-1$ block columns of a $np \times np$ matrix Z, the condition that*

$$\mathrm{tr}_p(ZM_k) \ = \ 0 \qquad k = 1, \ldots, n$$

is equivalent to the condition that

$$Z \ = \ XC - CX \quad \textit{for some } X \textit{ with } 0 \textit{ bottom block row.}$$

Either condition determines the final block column of Z uniquely, and the subspace

$$\boldsymbol{SubTan}_{subopt} \ = \ \{XC - CX : X \ \in \ \mathbb{C}^{np \times np} \text{ with 0 bottom block row}\} \ ,$$

is complementary to ***Syl****.*

Decomposing Δ_C in $\mathbf{SubTan}_{subopt} \oplus \mathbf{Syl}$, and taking the special structure of the M_k's into account, it is then shown that the first order perturbation of coefficient P_k with respect to this decomposition is given by :

$$\delta P_k = -\mathrm{tr}_p(\Delta_C \ M_{k+1}) \ + \ O(\|\Delta_C\|^2) \ . \tag{3.8}$$

We define M_{kij} to be the matrix with zero everywhere except in the i-th column of each block column where one has the j-th column of the corresponding block column of M_k. Notice that each M^*_{kij} does not belong to the normal space at C anymore since it does not commute with C^* due to the non-commutativity of the P_k's. But these matrices span a basis for all matrices commuting with C^* *except in the last block row.* The expression (3.8) can be written componentwise as follows

$$\delta(P_k)_{ij} = - \ \langle \Delta_C \ , \ M^*_{k+1ij} \ \rangle \ + \ O(\|\Delta_C\|^2)$$

One shows that $\mathbf{SubTan}_{subopt}$ is orthogonal to $\mathrm{vec}M_{kij},\ k = 1, ..., n,\ i, j = 1, ..., p$, which allows to define the coefficient sensitivity for coefficient $(P_k)_{ij}$ as

$$\kappa_{kij} \ = \ \|M_{kij}\|_F \ .$$

Notice that when $p = 1$ we retrieve the result of [4]. The scaled sensitivity now becomes

$$\kappa_{kij}(D) \ = \ \|D^{-1}M_{kij}D\|_F$$

which leads to the following optimal scaling problem

$$\min_D \ \max_{k,i,j} \ \|D^{-1}M_{kij}D\|_F \tag{3.9}$$

which can be solved via convex optimization [6].

We finally analyze the choice of theorem 3.2. We have computed the minimal canonical angles between the projecting spaces for the optimal and the suboptimal decompositions and called them $\theta_{opt} := \theta(\mathbf{SubTan}_{opt}, \mathbf{Syl})$ and $\theta_{subopt} := \theta(\mathbf{SubTan}_{subopt}, \mathbf{Syl})$. In Table 3.1 we compare their cosines for different values of the coefficients P_i in the case where $n = p = 3$. Let us notice that for $P_i = 0$, i.e. in the case where $P(\lambda) = I\lambda^3$, the codimension of **Tan** is equal to np^2 implying that there is only one choice for **SubTan** : we need to take the entire tangent space for the decomposition. Therefore if $\|P_i\|$ is small, our suboptimal choice will be close to the optimal choice. From the table we conclude that the suboptimal choice is usually close to the optimal choice.

TABLE 3.1
Comparison between the optimal and the suboptimal decomposition

$n = p = 3$	max $\|P_i\| = 1e-06$	max $\|P_i\| = 1$	max $\|P_i\| = 1e+06$
$\cos\theta_{opt}$	8.16e-01	9.19e-01	9.01e-12
$\cos\theta_{subopt}$	8.16e-01	9.37e-01	9.83e-12

4. The general non-monic matrix case. Now we want to compute the eigenvalues λ_i of the $p \times p$ matrix polynomial $P(\lambda) := P_n\lambda^n + P_{n-1}\lambda^{n-1} + \ldots + P_1\lambda + P_0$ where P_n is supposed to be different from I_p. It is important to notice that we do not impose P_n to be invertible. To achieve this, we can consider the pencil

$$\lambda B - A = \lambda \begin{bmatrix} I_p & & & \\ & \ddots & & \\ & & I_p & \\ & & & P_n \end{bmatrix} - \begin{bmatrix} 0 & & & -P_0 \\ I_p & 0 & & -P_1 \\ & \ddots & & \vdots \\ & & I_p & -P_{n-1} \end{bmatrix} \tag{4.1}$$

whose determinant is $\det(\lambda B - A) = \det P(\lambda)$. We then compute the eigenvalues of $P(\lambda)$ using any eigenvalue algorithm. If the algorithm is backward stable (think for example of the QZ-algorithm), the computed eigenvalues are exactly those of a matrix pencil $\lambda(B + \Delta_B) - (A + \Delta_A)$ for some dense backward error matrix pencil $\Delta(\lambda) := \lambda\Delta_B - \Delta_A$ satisfying (1.2). We again require the polynomial sensitivity κ_{pol} (2.1) to be minimal.

We work in the $2n^2p^2$-dimensional space of $np \times np$ complex matrix pencils with Frobenius inner product $\langle \lambda B_1 - A_1, \lambda B_2 - A_2 \rangle := \mathrm{tr}(A_1A_2^* + B_1B_2^*)$. In this pencil space we consider the manifold

$$\mathbf{Orb} \ := \ \mathrm{orbit}(\lambda B - A) \ = \ \{T_2^{-1}(\lambda B - A)T_1 : \det(T_2) = \det(T_1) \neq 0\}.$$

The tangent space to **Orb** at $\lambda B - A$ is given by

$$\mathbf{Tan} \;=\; \{\lambda Z_B - Z_A = \lambda(XB - BY) - (XA - AY) \;:\; \mathrm{tr}(X) \;=\; \mathrm{tr}(Y)\, X,\; Y \;\in\; \mathbb{C}^{np\times np}\}\,.$$

The Sylvester space consists of $p \times p$ block companion pencils and has dimension $(n+1)p^2$, while the tangent space **Tan** generically has dimension $2n^2p^2 - np - 1$, but it can be smaller. For $p > 1$ the problem is the same as before : we have to choose a subspace $\mathbf{SubTan} \subseteq \mathbf{Tan}$ such that

$$\mathbf{SubTan} \oplus \mathbf{Syl} \;=\; \mathbb{C}^{np\times np} \times \mathbb{C}^{np\times np}$$

and such that $\kappa_{pol} = \|\Pi\|_2$ is as small as possible, where Π is the projector onto **Syl** along **SubTan**. To find an explicit expression for the polynomial sensitivity, we consider the mapping whose image is the tangent space

$$\begin{bmatrix} \mathrm{vec}\; X \\ \mathrm{vec}\; Y \end{bmatrix} \longrightarrow \begin{bmatrix} B^T \otimes I & -I \otimes B \\ A^T \otimes I & -I \otimes A \end{bmatrix} \begin{bmatrix} \mathrm{vec}\; X \\ \mathrm{vec}\; Y \end{bmatrix} \quad , \quad \mathrm{tr}(X) \;=\; \mathrm{tr}(Y).$$

Its orthogonal complement is the kernel of the dual mapping. This finally leads to the following theorem.

THEOREM 4.1. *Let* ***Syl*** *be spanned by* $\mathrm{Im}\mathcal{X}$ *and let* $\mathrm{Im}\mathcal{Y}_{opt}$ *be the orthogonal compliment of a complimentary subspace* ***SubTan*** *of* ***Tan*** *with the largest minimal angle with* ***Syl*** *then*

$$\kappa_{pol} = \|\Pi_{opt}\|_2 = \|\mathcal{X}(\mathcal{Y}_{opt}^*\mathcal{X})^{-1}\mathcal{Y}_{opt}^*\|_2.$$

The polynomial sensitivity $\kappa_{pol}(T_1, T_2)$ of the pencil $T_2^{-1}(\lambda B - A)T_1$ is then defined to be

$$\kappa_{pol}(T_1, T_2) \;=\; \|\mathcal{T}^{-1}\mathcal{X}(\mathcal{Y}_{opt}^*\mathcal{X})^{-1}\mathcal{Y}_{opt}^*\mathcal{T}\|_2 \quad , \quad \mathcal{T} := T_1^{-T} \otimes T_2.$$

Following the same reasoning as before, one can also prove the following theorem that shows which pencils of **Orb** have a good polynomial sensitivity.

THEOREM 4.2. *The pencil* $T_2^{-1}(\lambda B - A)T_1$ *is normal if* T_1 *and* T_2 *minimize*

$$\min_{\det(T_1)=\det(T_2)} \kappa_{pol}(T_1, T_2).$$

Now restricting ourselves to scaling transformations, one would like to solve

$$\min_{\det(D_1)=\det(D_2)} \kappa_{pol}(D_1, D_2).$$

Here we need again a specific choice of complimentary subspace **SubTan** with an explicit basis for the orthogonal complement $\mathbf{SubTan}^{\perp}$. We derive such a basis and show how it relates the perturbation pencil $\lambda\Delta_B - \Delta_A$ of the block companion pencil $\lambda B - A$ with first order perturbations δP_k in the coefficients P_k of $P(\lambda)$. The subspace $\mathbf{SubTan}_{subopt}$ of the tangent space consists of all pencils of the form $\lambda(XB - BY) - (XA - AY)$ where Y is arbitrary *with last bottom block row equal to zero* and X is arbitrary *satisfying* $\mathrm{tr}_p(X) \;=\; \mathrm{tr}_p(Y)$. We show that it is complementary to **Syl**. We

first introduce the matrices

$$\overline{M}_k = \overbrace{\begin{bmatrix} P_k & & & \\ P_{k+1} & P_k & & \\ \vdots & P_{k+1} & \ddots & \\ P_n & \vdots & \ddots & P_k \\ & P_n & \ddots & P_{k+1} \\ & & \ddots & \vdots \\ & & & I_p \end{bmatrix}}^{k} \overbrace{\begin{bmatrix} -P_0 & & \\ -P_1 & \ddots & \\ \vdots & \ddots & -P_0 \\ -P_{k-1} & \ddots & -P_1 \\ & \ddots & \vdots \\ & & -P_{k-1} \end{bmatrix}}^{n-k} . \tag{4.2}$$

With $\overline{M}_0 = \overline{M}_{n+1} = 0$, one can check that these matrices satisfy

$$B\overline{M}_k \;=\; A\overline{M}_{k+1} + I_n \otimes P_k \qquad \text{for } k = 0, \ldots, n. \tag{4.3}$$

If one perturbs the block companion pencil $\lambda B - A$ with a dense perturbation pencil $\Delta(\lambda) = \lambda\Delta_B - \Delta_A$, then first order perturbations of the coefficient P_k are given by

$$\delta P_k = \mathrm{tr}_p(\Delta_B\ \overline{M}_k - \Delta_A\ \overline{M}_{k+1}) + O(\|\Delta(\cdot)\|^2) \tag{4.4}$$

The proof is based on the following theorem.

THEOREM 4.3. *Given the first $n-1$ block columns of a $np \times np$ matrix Z_A and given a $np \times np$ matrix Z_B except for its bottom right $p \times p$ block, the condition that*

$$\mathrm{tr}_p(Z_B\overline{M}_k - Z_A\overline{M}_{k+1}) = 0 \qquad k = 0, \ldots, n \tag{4.5}$$

is equivalent to the condition that

$$\lambda Z_B - Z_A = \lambda(XB - BY) - (XA - AY) \tag{4.6}$$
$$\text{for some } X,\ Y \text{ such that } \begin{cases} \mathrm{tr}_p(X) \;=\; \mathrm{tr}_p(Y) \\ Y \text{ has zero bottom block row.} \end{cases}$$

Either condition determines the final block column of Z_A and the bottom right block of Z_B uniquely.

Proof. Let us first show that (4.5) implies (4.6). For $k < n$, the relation $\mathrm{tr}_p(Z_B\overline{M}_k - Z_A\overline{M}_{k+1}) \;=\; 0$ determines the $(k+1, n)$ block of Z_A uniquely. For $k = n$, this becomes $\mathrm{tr}_p(Z_B\overline{M}_n) \;=\; 0$ and it determines the (n, n) block of Z_B uniquely. One verifies that $\mathrm{tr}_p(X) = \mathrm{tr}_p(Y)$ and $[\, 0 \cdots 0\ I_p\,].Y = 0$ makes the following mapping injective

$$(X, Y) \longrightarrow (\ XB - BY \;=\; \left[\begin{array}{cc|c} & & \\ & & \\ \hline & & ? \end{array}\right] ,\ XA - AY \;=\; \left[\begin{array}{cc|c} & & ? \\ & & ? \\ & & ? \\ & & ? \end{array}\right])$$

where the ? blocks are not taken into account. This amounts to saying that $\mathbf{SubTan}_{subopt}$ and **Syl** do not intersect, and implies that $\lambda Z_B - Z_A$ is uniquely determined by (4.6).

Let us now prove that (4.6) implies (4.5). We have thus to prove that for X, Y such that $\mathrm{tr}_p(X) \;=\; \mathrm{tr}_p(Y)$ where Y has zero bottom block row it implies that

$$\mathrm{tr}_p((XB - BY)\ \overline{M}_k - (XA - AY)\ \overline{M}_{k+1}) = 0 \qquad k = 0, \ldots, n\ . \tag{4.7}$$

We will first prove (4.7) when $k \notin \{0, n\}$. We can rewrite (4.7) as follows

$$\mathrm{tr}_p(X(B\overline{M}_k - A\overline{M}_{k+1}) - BY\overline{M}_k + AY\overline{M}_{k+1}) \;=\; 0\ .$$

Using (4.3), we can write

$$\begin{aligned}\mathrm{tr}_p(X(B\overline{M}_k - A\overline{M}_{k+1})) &= \mathrm{tr}_p(X(I_n \otimes P_k)) = \mathrm{tr}_p(X)P_k \\ = \mathrm{tr}_p(Y)P_k &= \mathrm{tr}_p(Y(I_n \otimes P_k)) = \mathrm{tr}_p(Y(B\overline{M}_k - A\overline{M}_{k+1}))\ .\end{aligned}$$

Therefore we need to prove that

$$\mathrm{tr}_p(Y(B\overline{M}_k - A\overline{M}_{k+1}) - BY\overline{M}_k + AY\overline{M}_{k+1}) \;=\; 0$$

or equivalently that

$$\mathrm{tr}_p((YB - BY)\ \overline{M}_k) = \mathrm{tr}_p((YA - AY)\ \overline{M}_{k+1})\ . \tag{4.8}$$

Because $(YB - BY)$ is zero everywhere except in the last column, we have that

$$\mathrm{tr}_p((YB - BY)\ \overline{M}_k) \;=\; Y_{kn}(P_n - I_p)\ .$$

Now using (3.7) and (4.2), one can write $\overline{M}_{k+1}$ as

$$\overline{M}_{k+1} = M_{k+1} + \begin{bmatrix} \overbrace{\begin{matrix} 0 & & \\ \vdots & \ddots & \\ P_n - I_p & \ddots & 0 \\ & & \vdots \\ & & P_n - I_p \\ & & 0 \end{matrix}}^{kp} & \overbrace{\begin{matrix} \\ \\ \\ 0 \\ \\ \\ \end{matrix}}^{(n-k)p} \end{bmatrix}\ .$$

It follows from theorem (3.2) that $\mathrm{tr}_p((YA - AY)\ M_{k+1}) \;=\; 0$ because Y has zero bottom block row. After some computation one checks that

$$\mathrm{tr}_p((YA - AY), \begin{bmatrix} \overbrace{\begin{matrix} 0 & & \\ \vdots & \ddots & \\ P_n - I_p & \ddots & 0 \\ & & \vdots \\ & & P_n - I_p \\ & & 0 \end{matrix}}^{kp} & \overbrace{\begin{matrix} \\ \\ \\ 0 \\ \\ \\ \end{matrix}}^{(n-k)p} \end{bmatrix}) = Y_{kn}(P_n - I_p),$$

which proves (4.8). If k is 0, then $\overline{M}_{k+1} = \overline{M}_1 = M_1$, and thus (4.7) becomes

$$\mathrm{tr}_p((XA - AY)\ \overline{M}_1) \;=\; 0$$

which is true because of theorem (3.2). If k is n, from (4.7) we have to prove that

$$\mathrm{tr}_p((XB - BY)\ \overline{M}_n) \;=\; 0\ .$$

The left side of this equation can be rewritten as follows :

$$\mathrm{tr}_p((XB - BY)\ \overline{M}_n) = \mathrm{tr}_p(XB\overline{M}_n) - \mathrm{tr}_p(BY\overline{M}_n) = \mathrm{tr}_p(X)P_n - \mathrm{tr}_p(Y)P_n = 0$$

because $\mathrm{tr}_p(X) = \mathrm{tr}_p(Y)$. This ends the proof. ∎

REMARK 4.1. *Defining $\overline{M}_{kij}$ to be the matrix with zero everywhere except in the i-th column of each block column where one has the j-th column of the corresponding block column of $\overline{M}_k$, we have that (4.5) is equivalent to*

$$\langle \lambda Z_B - Z_A\ ,\ \lambda \overline{M}^{\ *}_{kij} + \overline{M}^{\ *}_{k+1ij} \rangle \qquad k = 0, \ldots, n, \quad i, j = 1, \ldots, p\ .$$

This theorem actually shows that defining $\mathbf{SubTan}_{subopt}$ as

$$\{\lambda(XB - BY) - (XA - AY)\ : \begin{cases} \mathrm{tr}_p(X) = \mathrm{tr}_p(Y) \\ Y \text{ has zero bottom block row} \end{cases} \qquad X, Y \in \mathbb{C}^{np \times np}\},$$

implies that $\mathbf{SubTan}^{\perp}_{subopt}$ is spanned by the pencils $\lambda\overline{M}_{kij} + \overline{M}^{\ *}_{k+1ij}$ and that $\mathbf{Syl} \oplus \mathbf{SubTan}_{subopt} = \mathbb{C}^{np \times np} \times \mathbb{C}^{np \times np}$. Notice here again that even though the $\lambda\overline{M}^{\ *}_{k} + \overline{M}^{\ *}_{k+1}$'s are orthogonal to $\mathbf{SubTan}_{subopt}$, they don't belong to the normal space at A anymore.

We have now everything we need to prove (4.4). It suffices to decompose $\lambda\Delta_B - \Delta_A$ in $\mathbf{Syl} \oplus \mathbf{SubTan}_{subopt}$, and to take the block trace on both sides. Then taking the special structure of the $\overline{M}_k$'s into account, we obtain the desired equation. The expression (4.4) can be written componentwise like

$$\delta(P_k)_{ij} = \langle \lambda\ \Delta_B - \Delta_A\ ,\ \lambda\overline{M}^{\ *}_{kij} + \overline{M}^{\ *}_{k+1ij} \rangle\ +\ O(\|\Delta\|^2)\ .$$

This allows us to define a coefficient sensitivity for coefficient $(P_k)_{ij}$ as

$$\kappa_{kij} = \|\lambda\overline{M}_{kij} + \overline{M}_{k+1ij}\|_F\ .$$

Notice that when $p = 1$ we retrieve the result of [4]. And it follows that

$$\kappa_{kij}(D1, D_2) = \|D_2^{-1}(\lambda\overline{M}_{kij} + \overline{M}_{k+1ij})D_1\|_F$$

which leads to the following convex optimal scaling problem

$$\min_{\det(D_1)=\det(D_2)} \max_{k,i,j} \|D_2^{-1}(\lambda\overline{M}_{kij} + \overline{M}_{k+1ij})D_1\|_F\ . \tag{4.9}$$

Now let us show that the choice of theorem 4.2 is good. For both the optimal decomposition and the suboptimal one we have computed the minimal canonical angle between the projecting spaces. We denote $\theta_{opt} := \theta(\mathbf{SubTan}_{opt}, \mathbf{Syl}_{opt})$ and $\theta_{subopt} := \theta(\mathbf{SubTan}_{subopt}, \mathbf{Syl})$. In the table below we compare their values for different values of the coefficients P_i in the case where $n = p = 3$. From the table we conclude that this suboptimal choice is relatively speaking close to the optimal choice.

TABLE 4.1
Comparison between the optimal and the suboptimal decomposition

$n = p = 3$	$\max \|P_i\| = 1e-06$	$\max\ \|P_i\| = 1$	$\max\ \|P_i\| = 1e+06$
$\cos\theta_{opt}$	9.22e-01	9.48e-01	9.44e-12
$\cos\theta_{subopt}$	9.22e-01	9.80e-01	9.87e-12

5. Numerical example. We consider the nuclear power plant problem in [7]. The characteristic frequencies of the system are the eigenvalues of a 8×8 matrix polynomial $P(\lambda) = P_2\lambda^2 + P_1\lambda + P_0$. A good way to measure the efficiency of the scaling strategy we propose here is to look at the structured backward error (i.e. in the coefficients P_i) corresponding to the computed eigenvalues. For this quadratic problem an expression for this error is given in [7]. After applying the QZ algorithm directly to the pencil

$$\lambda \begin{bmatrix} I & 0 \\ 0 & P_2 \end{bmatrix} - \begin{bmatrix} 0 & -P_0 \\ I & -P_1 \end{bmatrix}$$

we have for some eigenvalues a backward error in the coefficients that is of the order of $1e-06$. After applying our scaling strategy, the QZ algorithm gives for all eigenvalues a backward error of the order of $1e-16$!

6. Conclusions. Computing roots of a polynomial may be done by using any backward stable algorithm on a corresponding companion pencil. But this may yield very poor backward errors in the coefficients of the polynomial matrix. In this paper we analyzed what can be done to improve these errors, using a geometric approach. Starting from the definition of *polynomial sensitivity* in each point of the orbit of the companion pencil, we described the optimal points with minimal sensitivity. Restricting ourselves to diagonal scaling transformations, we have then defined a convex optimal scaling procedure. Finally, we have shown a numerical example to illustrate our theoretic results.

REFERENCES

[1] J.-P. Charlier and P. Van Dooren. A Jacobi-like algorithm for computing the generalized Schur form of a regular pencil. *J. Comput. Appl. Math.*, 27(1-2):17–36, 1989.

[2] A. Edelman, H. Murakami, Polynomial roots from companion matrix eigenvalues, *Math. Comput.* 64(210):763–776, 1995.

[3] H.Y. Fan, W.W. Lin, P. Van Dooren, *Normwise scaling of second order polynomial matrices*, accepted for *SIAM J. Matrix Anal. Appl.*, 2003.

[4] D. Lemonnier, P. Van Dooren, *Optimal scaling of companion pencils for the QZ algorithm*, Paper CP7-4, CD-Rom *Proceedings SIAM Applied Linear Algebra Conference*, Williamsburg, July, 2003.

[5] C. Moler and G. Stewart, An algorithm for generalized matrix eigenvalue problems., *SIAM J. Numer. Anal.*, 10:241–256, 1973.

[6] Yu. Nesterov, Private communication.

[7] F. Tisseur and K. Meerbergen, *The Quadratic Eigenvalue Problem*, SIAM Review, 43:235–286, 2001.

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