Mediterr. j. math. 99 (9999), 1–11
1660-5446/99000-0, DOI 10.1007/s00009-003-0000

Mediterranean Journal
of Mathematics

Smoothness of Hessenberg and Bidiagonal Forms

Luca Dieci, M. Grazia Gasparo and Alessandra Papini

Abstract. Our purpose in this work is to explore smoothness properties of transformations of a matrix valued function A to Hessenberg and bidiagonal form. The interplay with the rank of associated Krylov functions is exploited to clarify what one should expect for smooth functions A satisfying generic properties.

Mathematics Subject Classification (2000). 65F15, 65F99.

Keywords. Smoothness, Hessenberg and bidiagonal forms, Krylov subspace, Procrustes problem.

1. Introduction

A fundamental step in obtaining an efficient implementation of the QR algorithm to find eigenvalues of a matrix $A \in \mathbb{R}^{n\times n}$ is the reduction of A to Hessenberg form: This is the typical first step for obtaining the Schur reduction of A. Likewise, to find the singular value decomposition (SVD) of A, initial reduction to bidiagonal form is mandatory for obtaining efficient techniques. These points of view are well understood, and well presented in basic textbooks (see [6]). Moreover, how to obtain Hessenberg, respectively bidiagonal, forms of a matrix A is a well studied problem, and efficient algorithms exist for this scope; again, see [6].

Here, we consider a matrix valued function of one real variable, A: $t \in \mathbb{R} \to A(t) \in \mathbb{R}^{n\times n}$, of class $\mathcal{C}^k$, $k \geq 1$, which we will write as $A \in \mathcal{C}^k(\mathbb{R}, \mathbb{R}^{n\times n})$, or more simply as $A \in \mathcal{C}^k$, since we only work with square matrices. Now, for the function A, it is not necessarily true that the Schur form or its SVD exist with the same degree of differentiability of A. Sufficient conditions to guarantee that these forms exist with some degree of differentiability (possibly less than k) are in [2]. The analytic case is treated in [1] for the SVD, and in [5] for the Schur form when A has only real eigenvalues. Now, in case the Schur or SVD forms of the function A

This work was supported in part under INDAM-GNCS and MIUR Rome-Italy.

exist, analytic or differentiable, it is natural to expect that efficient algorithms for the computation of these smooth decompositions will need to pass through smooth simplification of A to Hessenberg, respectively bidiagonal, form. Indeed, some of the algorithms put forward in [3, 4] require this to be the case.

In this work, we explore the degree of differentiability of transformations of the function A to Hessenberg, respectively bidiagonal, form. Beside the above algorithmic motivation, the issue is of independent interest and our analysis is useful to better understand theoretical performance of the standard linear algebra algorithms to find the Schur form, or the SVD, of a matrix A.

Remark 1.1. When considering smoothness of decompositions for matrix valued functions in $\mathcal{C}^k(\mathbb{R}, \mathbb{R}^{n\times n})$, one should not be interested only in exploring possible pathological behavior of a specific function A, but it is more meaningful to look at the behavior of the entire family of such $\mathcal{C}^k$ functions. For this reason, in our results in Section 2, we will make some assumptions which may at first sight appear ad hoc. However, our assumptions hold in an open and dense subset of $\mathcal{C}^k$ functions, and hence it is a *generic property* that a matrix function satisfies these assumptions. For background on genericity, we refer to [7].

In Section 2 we give results on smoothness of Hessenberg and bidiagonal forms. We first give a general sufficient condition, and then more precise conditions motivated by generic properties of smooth functions. All along, we will exploit the relation of these simplified forms to appropriate Krylov functions. Use of the "Implicit Q-Theorem" will prove essential to determine the degree of uniqueness of these simplified forms. Uniqueness makes things simpler because if there is a "unique" reduction, and we know that there is a smooth one, then that must be the one. When we do not necessarily have uniqueness, but still have guarantee of existence of a smooth decomposition, it is not obvious how to retrieve it. In Section 3 we see how –in principle– this may be done.

Remarks 1.2.

(1) We restrict to functions depending on one parameter for several reasons. First of all, smoothness results on decompositions of functions of two or more variables are much harder to get by, also for analytic functions (see [8]). Secondly, practical continuation algorithms for functions of several parameters proceed by holding all parameters frozen except one of them, and thus one ends up having to continue decompositions for matrices depending on one parameter.

(2) In this work, we are restricting to square matrices. Obviously, this is needed for a similarity reduction to Hessenberg form. For the case of reduction to bidiagonal form, however, our results extend easily to rectangular functions $A \in \mathcal{C}^k(\mathbb{R}, \mathbb{R}^{m\times n})$, $m > n$, as long as A is full rank. The construction now would require an initial reduction to a square problem by finding a smooth orthonormal basis for $\mathcal{N}(A^T)$ as done in [4].

2. Hessenberg and Bidiagonal forms

In this section we examine smoothness of reduction to Hessenberg form for general matrix valued functions, and then further to bidiagonal form (via reduction to tridiagonal form for symmetric functions).

2.1. Hessenberg form

The problem here is the following.

We are given a function $A \in \mathcal{C}^k$, $k \geq 1$, and a unit vector $v \in \mathbb{R}^n$. [The case of v being also a $\mathcal{C}^k$ function can be handled similarly.] We want to find Q orthogonal, such that $Q^T AQ = H$, with H upper Hessenberg and the first column of Q given by v, $Qe_1 = v$. Moreover, we want to have Q smooth, though possibly less smooth than A itself.

The fundamental theoretical backing is given by the well known interplay between the Hessenberg form and the QR factorization of the associated Krylov matrix valued function K below (for $t \in \mathbb{R}$):

$$K(t) = [v, A(t)v, \ldots, A^{n-1}(t)v]\,. \tag{2.1}$$

Using [2, Theorem 3.1], we immediately have the following existence result, essentially Theorem 3 in [3].

Theorem 2.1. *If there exists d, $0 \leq d \leq k$, such that*

$$\limsup_{h\to 0} \frac{1}{h^{2d}} \det\bigl(K^T(t+h)K(t+h)\bigr) > 0$$

for all t, then there exists $Q \in \mathcal{C}^{k-d}$, orthogonal, such that

$$Q^T(t)A(t)Q(t) \;=\; H(t)\ ,$$

where H is upper Hessenberg for all t. Moreover, H is unreduced [1] *except at most at isolated points, where K is not full rank.*

Proof. The proof is the same as in [3, Theorem 3]. □

We now want to characterize rank conditions on K in terms of choices of the first column v of Q. We state the following results purely in linear algebraic terms, that is when the objects involved are just matrices, not matrix valued functions. Of course, the rank of A itself plays a role in the value of $\mathrm{rank}(K)$. After all, the last $(n-1)$ vectors making up K all lay in the range of A. The following elementary result summarizes this consideration:

$$\text{If} \quad \mathrm{rank}(A) = n - q, \quad \text{then} \quad \mathrm{rank}(K) \leq n - q + 1\ .$$

A more precise characterization, which relates the rank of K to the choice of v and the invariant subspaces of A is expressed in the property below; this result appeared first in the Control Engineering literature, see [13, section 1.2], and later

[1] we recall that a upper Hessenberg matrix $H \in \mathbb{R}^{n\times n}$ is called *unreduced*, or *proper*, if $H_{i+1,i} \neq 0$ for $i = 1, \ldots, n-1$.

rediscovered by the numerical linear algebra community, see [10, section 12.2] and [12, section 6.2].

Property 2.1. *Let $A \in \mathbb{R}^{n\times n}$, $v \in \mathbb{R}^n$, $v^Tv = 1$, and $K = [v, Av, \dots, A^{n-1}v]$. Then,* $\mathrm{rank}(K)$ *is the dimension of the smallest invariant subspace of A containing v.*

Remark 2.2. In practical terms, since invertible matrices are dense in the set of matrices, one should expect the matrix K to be of full rank for basically all unit vectors v; see also [13, section 1.5]. [This is because a p-dimensional invariant subspace of A, with $p \le n-1$, spans a set of measure 0 in $\mathbb{R}^n$.] Now, recall that techniques which seek a Schur form of a matrix A typically perform a preprocessing step bringing A to Hessenberg form by a matrix $Q : \; Q^TAQ = H$, where H is upper Hessenberg and the first column of Q is some unit vector, v. As a consequence, since full rank of K implies that H is unreduced (see [6, Theorem 7.4.3]), one should expect the Hessenberg form of a general matrix to be unreduced.

When we consider functions, things are somewhat more complicated, but the above considerations are still useful and provide insight into what are appropriate and reasonable assumptions to make. We begin with an important observation.

Remark 2.3. As it is well known, a $\mathcal{C}^k$ function A generically has $\mathrm{rank}(A) \ge n-1$, and $\mathrm{rank}(A(t)) = n-1$ at isolated values of t. This is nothing more than the rephrasing of the fact that the only generic co-dimension one bifurcation for which A loses rank is the "quadratic fold" bifurcation; e.g., see [9]. As a consequence, relatively to the function K of (2.1), the function K will generically satisfy $\mathrm{rank}(K) \ge n-1$, and $\mathrm{rank}(K(t)) = n-1$ at isolated values of t. [This is because a one parameter p-dimensional invariant subspace of A has codimension $n-p+1$.]

Remark 2.3 has an important consequence of practical relevance, which is summarized next.

Theorem 2.4. *If* $\mathrm{rank}(K(t)) \ge (n-1)$ *for all t, where K is given in* (2.1)*, then there is a $\mathcal{C}^k$ upper Hessenberg form of A, and it is essentially unique. More precisely, for any interval $[a,b]$, for $t \in [a,b]$, and for given $\bar{Q}$ yielding a Hessenberg form for $A(a)$, there is a unique function Q in $\mathcal{C}^k$, giving Hessenberg form for A, passing through $\bar{Q}$. Furthermore, if the values of t where $K(t)$ has rank $(n-1)$ are isolated, then the Hessenberg form H is unreduced for all t, except at those points where K is of rank $n-1$, where H takes the form $H = \begin{bmatrix} H_1 & \hat{h}_n \\ 0 & h_{nn} \end{bmatrix}$, with $H_1 \in \mathbb{R}^{n-1,n-1}$ upper Hessenberg and unreduced. In other words, generically, there is an essentially unique $\mathcal{C}^k$ upper Hessenberg form as stated.*

Proof. If $\mathrm{rank}(K(t)) \ge (n-1)$, then the first $(n-1)$ columns of $K(t)$ are linearly independent for all t; from their QR-factorization we can select a $\mathcal{C}^k$ set of $(n-1)$ orthonormal vectors $[q_1(t), q_2(t), \dots, q_{n-1}(t)]$, for all t, where $q_1 = v$. These vectors are smooth as soon as we fix the signs of the diagonal of the factor R (the usual choice is to keep it positive), and they uniquely determine the last vector $q_n(t)$ up to sign. Thus, we can form a $\mathcal{C}^k$ function Q, which is indeed made unique by the requirement $Q(a) = \bar{Q}$. The statement about the smoothness of H follows at once upon recalling that $H(t) = Q^T(t)A(t)Q(t)$, for all t. The form of H at points where K has rank $(n-1)$ is a consequence of the fact that from $R = Q^T K$, we must have $R(t) = [\pm e_1, \pm H(t)e_1, \dots, \pm H^{n-1}(t)e_1]$ and the fact that the last diagonal entry of R is 0. □

2.2. Bidiagonal form

Here we seek a smooth bidiagonalization of the $\mathcal{C}^k$ function A. That is, we want to find U and V, orthogonal, such that, for all t:

$$U^T(t)A(t)V(t) = B(t) \ , \quad B = \begin{bmatrix} a_1 & b_1 & & & \\ & a_2 & b_2 & & \\ & & \ddots & \ddots & \\ & & & a_{n-1} & b_{n-1} \\ & & & & a_n \end{bmatrix} , \tag{2.2}$$

where the first column of $V(t)$ is given by a fixed unit vector v, and the first column of $U(t)$ is given by $A(t)v/\|A(t)v\|$, for all t. Of course, at each given t, we can always find orthogonal U and V as stated, giving the bidiagonal form B. But, we want to recover smooth U and V, if smooth ones exist.

We will exploit the relation of the bidiagonalization procedure to the tridiagonalization of A^TA and AA^T. That is, we will consider the two problems: Find smooth V and U such that

$$\begin{aligned} &(a) \qquad V^T(A^TA)V = T_v \ , \quad Ve_1 = v \ , \\ &(b) \qquad U^T(AA^T)U = T_u \ , \quad Ue_1 = Av/\,\|Av\| \ , \end{aligned} \tag{2.3}$$

where T_v and T_u are tridiagonal. Obviously, if we have B we can get $T_v = B^TB$ and $T_u = BB^T$. The converse requires more work.

To begin with, we have the following result, which is fully general, and which highlights the importance of unreduced structure.

Lemma 2.5. *Let $A \in \mathbb{R}^{n\times n}$ and let U and V be orthogonal matrices giving tridiagonal reductions for AA^T and A^TA, respectively, with $Ve_1 = v$ and $Ue_1 = Av/\|Av\|$. Moreover, let the matrices $T_v = V^T(A^TA)V$ and $T_u = U^T(AA^T)U$ be unreduced. Then, $B = U^TAV$ is bidiagonal and unreduced.*

Proof. Let $\widehat{U}$ and $\widehat{V}$ be orthogonal matrices, with $\widehat{V}e_1 = v$ and $\widehat{U}e_1 = Av/\|Av\|$ such that $\widehat{B} = \widehat{U}^TA\widehat{V}$ is bidiagonal. Moreover, let us consider the tridiagonal reductions $\widehat{V}^TA^TA\widehat{V} = \widehat{T}_v$ and $\widehat{U}^TAA^T\widehat{U} = \widehat{T}_u$. Since $\widehat{V}e_1 = Ve_1$, $\widehat{U}e_1 = Ue_1$, and both T_v and T_u are unreduced, the Implicit-Q theorem ([6, Theorem 8.3.2]) implies that $\widehat{T}_v$ and $\widehat{T}_u$ are unreduced and $\widehat{V}$ and $\widehat{U}$ coincide with V and U, respectively, except at most for the signs of the 2nd, 3rd, ..., n-th columns. Therefore, $\widehat{B}$ is unreduced; moreover, B coincide with $\widehat{B}$ except at most for signs. In particular, B is bidiagonal and unreduced. □

Next, let A be a $\mathcal{C}^k$ function. We introduce the Krylov functions:

$$\begin{aligned} &(a) \qquad K_v = [v, (A^TA)v, \dots, (A^TA)^{n-1}v] \ , \\ &(b) \qquad K_u = [Av, (AA^T)Av, \dots, (AA^T)^{n-1}Av]\frac{1}{\|Av\|} \ . \end{aligned} \tag{2.4}$$

We observe that $K_u = AK_v/\|Av\|$; in particular, if A is full rank, then $\mathrm{rank}(K_v) = \mathrm{rank}(K_u)$.

We have the following general result about smoothness of the bidiagonal B, related to the rank of K_v and K_u (cfr. Theorem 2.1).

Theorem 2.6. *Let there exists d, $0 \le d \le k$, such that*

$$\limsup_{h\to 0} \frac{1}{h^{2d}} \det\big(K_v^T(t+h)K_v(t+h)\big) > 0 \,,$$

and

$$\limsup_{h\to 0} \frac{1}{h^{2d}} \det\big(K_u^T(t+h)K_u(t+h)\big) > 0 \,,$$

for all t. Then, there exists $U, V \in \mathcal{C}^{k-d}$, orthogonal, such that

$$U^T(t)A(t)V(t) \ = \ B(t) \ ,$$

and B is upper bidiagonal for all t. Moreover, B is unreduced except at isolated points where K_v and/or K_u loose rank.

Proof. As in the argument of [2, Theorem 3.1], the assumption on K_v implies that K_v has full rank, except at most at isolated points, and it has a $\mathcal{C}^{k-d}$ QR decomposition: $K_v(t) = V(t)R_v(t)$ for all t. Similarly, also K_u has full rank, except at most at isolated points, and it has a $\mathcal{C}^{k-d}$ QR decomposition: $K_u(t) = U(t)R_u(t)$ for all t. By virtue of Theorem 2.1, T_v and T_u are thus tridiagonal, and unreduced except at most at isolated points. Because of Lemma 2.5, we then have that B is upper bidiagonal and unreduced in these intervals. But then, because of continuity of B, B must be upper bidiagonal (though not necessarily unreduced) everywhere. □

So, we can unambiguously write the following expressions for T_v and T_u (recall the form of B in (2.2)):

$$\begin{aligned} T_v &= \begin{bmatrix} a_1^2 & a_1b_1 & & & & \\ a_1b_1 & a_2^2{+}b_1^2 & a_2b_2 & & & \\ & a_2b_2 & \ddots & \ddots & & \\ & & \ddots & \ddots & a_{n-2}b_{n-2} & \\ & & & a_{n-2}b_{n-2} & a_{n-1}^2{+}b_{n-2}^2 & a_{n-1}b_{n-1} \\ & & & & a_{n-1}b_{n-1} & a_n^2{+}b_{n-1}^2 \end{bmatrix}, \\ T_u &= \begin{bmatrix} a_1^2{+}b_1^2 & a_2b_1 & & & & \\ a_2b_1 & a_2^2{+}b_2^2 & a_3b_2 & & & \\ & a_3b_2 & \ddots & \ddots & & \\ & & \ddots & \ddots & a_{n-1}b_{n-2} & \\ & & & a_{n-1}b_{n-2} & a_{n-1}^2{+}b_{n-1}^2 & a_nb_{n-1} \\ & & & & a_nb_{n-1} & a_n^2 \end{bmatrix}. \end{aligned} \tag{2.5}$$

Now, justifying the assumptions below in a similar way to what we did in Remark 2.3, we obtain the following analog of Theorem 2.4.

Theorem 2.7. *Consider K_v and K_u in (2.4). We have the following cases.*

(i) *Suppose* $\operatorname{rank}(K_v(t)) \geq (n-1)$ *for all t, and the values of t where $K_v(t)$ has rank $(n-1)$ are isolated. Assume that A has full rank at these isolated values. Then, there is a $\mathcal{C}^k$ bidiagonal form B, which is unreduced for all t, except at those points where K_v is of rank $(n-1)$, where B takes the form $B = \left[\begin{smallmatrix} B_1 & 0 \\ 0 & a_n \end{smallmatrix}\right]$, with $B_1 \in \mathbb{R}^{n-1,n-1}$ upper bidiagonal and unreduced, that is $b_{n-1} = 0$ in* (2.2).

(ii) *Suppose now that* $\operatorname{rank}(K_v(t)) = (n-1)$ *at some isolated values of t, but that at these values of t we also have* $\operatorname{rank}(A(t)) = (n-1)$. *Then, at these values of t, we have* $\operatorname{rank}(K_u(t)) = (n-1)$ *if and only if the form of B in* (2.2) *is $B = \left[\begin{smallmatrix} B_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, with $B_1 \in \mathbb{R}^{n-1,n-1}$ upper bidiagonal and unreduced, that is $a_n = b_{n-1} = 0$ in B. In this case, again there is a $\mathcal{C}^k$ bidiagonal form B, which is unreduced for all t, except at those points where K_v is of rank $(n-1)$, and is essentially unique.*

(iii) *Finally, if* $\operatorname{rank}(K_v(t)) = (n-1)$ *at some isolated values of t, and at these t we have either (a)* $\operatorname{rank}(A(t)) = (n-1)$ *and $a_n^2 + b_{n-1}^2 \neq 0$, or (b)* $\operatorname{rank}(A(t)) = (n-2)$, *then the function U is not uniquely determined at these points. In these cases, B will look like*

$$(a)\;\; B = \begin{bmatrix} a_1 & b_1 & & & \\ & \ddots & \ddots & & \\ & & a_{n-2} & b_{n-2} & \\ & & & 0 & b_{n-1} \\ & & & & a_n \end{bmatrix}, \quad or \quad (b)\;\; B = \begin{bmatrix} a_1 & b_1 & & & \\ & \ddots & \ddots & & \\ & & a_{n-2} & b_{n-2} & \\ & & & 0 & 0 \\ & & & & 0 \end{bmatrix}.$$

Proof. To show (i), we observe that, in intervals where K_v is full rank, we must have $\operatorname{rank}(A) \geq (n-1)$, and hence $\operatorname{rank}(K_u) \geq (n-1)$ as well. So, in all cases we can select a $\mathcal{C}^k$ orthogonal V from the QR factorization of K_v. Moreover, also the first $(n-1)$ columns of K_u are linearly independent and thus the first $(n-1)$ columns of U can again be chosen smooth, and the last column of U is then determined up to sign. Thus, U can also be chosen $\mathcal{C}^k$. Now, consider the isolated values of t where $\operatorname{rank}(K_v(t)) = n-1$, but $\operatorname{rank}(A(t)) = n$, so that $\operatorname{rank}(K_u(t)) = (n-1)$ as well. Since T_v and T_u must be unreduced, and the first $(n-1)$ columns of K_v and K_u are linearly independent, from the forms (2.5), we must have

$$a_{n-1}b_{n-1} = 0\,, \quad a_n b_{n-1} = 0\,.$$

Since A is invertible, this implies $b_{n-1} = 0$, which gives the stated form of B.

Next, we show (ii). First of all, we remark that certainly $(n-2) \leq \operatorname{rank}(K_u) \leq (n-1)$ (this is simply because $\|Av\| K_u = AK_v$). Now, let V give the QR factorization of K_v : $V^T K_v = \left[\begin{smallmatrix} R & b \\ 0 & 0 \end{smallmatrix}\right]$ with $R \in \mathbb{R}^{n-1,n-1}$ invertible. Then, $V^T A^T A V = \left[\begin{smallmatrix} T_{11} & 0 \\ 0 & t_{nn} \end{smallmatrix}\right]$, with T_{11} tridiagonal, invertible and unreduced. We now show that $\operatorname{rank}(K_u) = (n-1)$ if and only if $t_{nn} = 0$. First, assume $\operatorname{rank}(K_u) = (n-1)$. But then there exist (smooth) U (from the QR factorization of K_u) and B, and thus from the structure of $T_v = B^T B$ and $T_u = BB^T$, both necessarily reduced in the last co-diagonal element, we must have

$$a_n b_{n-1} = a_{n-1} b_{n-1} = 0 \ , \quad \text{and} \quad a_{n-1} b_{n-2} \neq 0 \, .$$

It follows that $b_{n-1} = 0$, and $a_j \neq 0$, for $j = 1, \dots, n-1$. But since A has rank $(n-1)$, then we must have $a_n = 0$. Since $t_{nn} = a_n^2 + b_{n-1}^2$, then $t_{nn} = 0$. Next, suppose $t_{nn} = 0$. Observe that

$$\|Av\| A^T K_u = A^T A K_v = A^T A V \left[\begin{smallmatrix} R & b \\ 0 & 0 \end{smallmatrix}\right] \ ,$$

and also that $\mathrm{null}(K_u) \subseteq \mathrm{null}(A^T K_u)$. But since $V^T A^T K_u = \left[\begin{smallmatrix} T_{11}R & T_{11}b \\ 0 & 0 \end{smallmatrix}\right]$, then the null space of $A^T K_u$ is 1-dimensional, and therefore the same is true for the dimension of the null space of K_u.

To show (iii) is an algebraic verification. In particular, in both cases (a) and (b) we must have $a_j \neq 0,\ b_j \neq 0,\ j = 1, \dots, n-2$. In case (a), since $t_{nn} \neq 0$, we also have $a_n, b_{n-1} \neq 0$. In case (b), since $\mathrm{rank}(K_u) = (n-2)$, but T_{11} is invertible and unreduced, we must thus have $a_{n-1} = b_{n-1} = a_n = 0$. □

3. Procrustes problems

Here we collect some simple considerations on solving minimization problems of Procrustes type, which form the basis for the practical selection of an appropriate smooth path in case such path is not uniquely determined by purely algebraic considerations (e.g., see point (iii) of Theorem 2.7). A related construction was adopted by Rheinboldt in [11].

The typical problem is the following. Suppose we have a smooth orthonormal function $Q : \mathbb{R} \to \mathbb{R}^{n \times p}$. For a given t, let Q_c be an orthonormal matrix such that its columns span the same subspace as that spanned by the columns of $Q(t)$. The goal is to bring Q_c into $Q(t)$ by an orthogonal transformation. There is a trivial way to do this if $Q(t)$ is known, since two orthonormal basis matrices of a given subspace are right equivalent by a unique orthogonal transformation Z: in fact, in our case $Z = Q_c^T Q(t)$.

Of course, in a practical situation, we do not know the exact $Q(t)$, so the above argument is of no immediate application. Still, it is possible to get around lack of knowledge of the exact $Q(t)$ in many situations of practical interest.

The simplest case occurs when one knows that Q_c equals $Q(t)$ except possibly for the signs of the columns. For example, this will happen when computing a smooth bidiagonalization and t is a point where the bidiagonal form is unreduced, or more generally $\mathrm{rank}(K_v)$ and $\mathrm{rank}(K_u)$ are $\geq (n-1)$. So, say we have $\widehat{Q} \approx Q(t)$; in particular $\widehat{Q}$ can be assumed to be sufficiently accurate so that the signs of its columns are the same as those of $Q(t)$. By same signs of the columns, we mean that $(\widehat{Q} e_i)^T (Q e_i) > 0$, where e_i are the standard unit vectors, for all $i = 1, \dots, n$. Therefore, it will be sufficient to adjust the signs of the columns of Q_c conformally to $\widehat{Q}$, which is easy to do. This way, we will be bringing Q_c exactly into $Q(t)$.

It is a more challenging case when Q_c differs from $Q(t)$ by more than just signs of the columns. E.g., this situation arises when we have $\mathrm{rank}(K_u) = n-2$ (as when $\mathrm{rank}(A) = n-2$). In this case, we do not have an exact way to bring Q_c into the unknown $Q(t)$, though we still have an approximation $\widehat{Q}$ to $Q(t)$, in principle with $\|\widehat{Q} - Q(t)\|$ arbitrarily small (see Example 3.1 below). In our experiments, we have proceeded as follows. Notice that we can (and do) use the process below without forcing $\widehat{Q}$ to be orthogonal.

- First, let $Z = Q_c^T\widehat{Q}$. Then, replace Z by its closest orthogonal matrix, that is by its orthogonal polar factor. This means that if $U_rS_rV_r^T$ is the SVD of Z, then we use $Z = U_rV_r^T$ to correct Q_c and approximate $Q(t)$ by Q_cZ.

We observe that the matrix Z obtained this way is the solution of the constrained minimization problem (orthogonal Procrustes problem):

$$\text{Find } Z \in \mathbb{R}^{p\times p},\ Z^TZ = I\ :\ \|Q_cZ - \widehat{Q}\|_F \text{ is minimized}\,.$$

Of course, the accuracy of the above minimization process will depend on how accurately we can provide $\widehat{Q}$. We report on an experiment below.

Example 3.1. We take $A(t) = U(t)B(t)V^T(t)$, $t \in [0,1]$, where:

$$U = e^S\ ,\quad S = -S^T\ ,\quad S_{ij} = \frac{(-1)^{i+j}}{j+1}(t-2)(t+2)^{j-1}\ ,\ j = 1:5, i = 1:j-1,$$

$$V = \left[\begin{smallmatrix} 1 & 0 \\ 0 & e^C \end{smallmatrix}\right]\ ,\quad C = -C^T\ ,\quad C_{ij} = \frac{(-1)^{i+j}}{j+1}(t-2)(t+1)^{j-1},$$
$$j = 1:4, i = 1:j-1,$$

$$\text{and}\qquad B = \left[\begin{smallmatrix} 1 & e^t & 0 & 0 & 0 \\ 0 & 2{+}\cos t & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & \sin(t{-}\widehat{t}) & (t{-}\widehat{t})^3 \\ 0 & 0 & 0 & 0 & t{-}\widehat{t} \end{smallmatrix}\right], \widehat{t} = 0.5.$$

We compute the smooth factors $U(t), V(t)$ and $B(t)$ for $t \in [0,1]$ starting from the exact matrices $U(0), V(0), B(0)$. The bidiagonal form is constructed at the points $t_k = k \times h$,$k = 1, 2, \dots$ for a given h. At each t_k we compute by standard algorithms a bidiagonalization $U_c^TA(t_k)V_c = B_c$ such that $V_ce_1 = V(0)e_1$ and $U_ce_1 = A(t_k)V_ce_1/\|A(t_k)V_ce_1\|$. For $t_k \neq \widehat{t}$, $B(t_k)$ (and then B_c) is unreduced and we only need to adjust the signs of the columns of U_c and V_c to recover smoothness. This is attained by using $\widehat{U} = U(t_{k-1})$ and $\widehat{V} = V(t_{k-1})$, whose columns have the correct signs for sufficiently small h. At the point $t_k = \widehat{t}$, the bidiagonal form $B(\widehat{t})$ is reduced, and we are in the case (iii)-(b) of Theorem 2.7, so that U_c is not uniquely determined and we use the procedure described above to rebuild $U(\widehat{t})$. In this case, we use $\widehat{U} = P(t_k)$, where $P(t)$ is a matrix valued polynomial interpolating past values of $U(t)$. In the table below we report on our experiments. We vary the discretization stepsize h and the degree of the interpolant used to construct $\widehat{U}$ at $\widehat{t}$: line, parabola, or cubic, which interpolates $U(t)$ at the (two, three, or four) grid-points before $\widehat{t}$. These are just plain polynomial interpolants, so that $\widehat{U}$ is not exactly orthogonal (though it is obviously close to being orthogonal with defect from orthogonality of the same order as the interpolation error). Under the headings `line, parabola, cubic`, we report the 2-norm of the errors in U at the point $\widehat{t}$. Under the heading `Err` we report the worse errors at all other grid points.

h	`line`	`parabola`	`cubic`	`Err`
0.01	9.8E − 6	1.1E − 6	4.7E − 8	6.9E − 14
0.001	1.0E − 7	1.1E − 9	4.9E − 12	3.5E − 13
0.0001	1.0E − 9	2.9E − 12	4.9E − 12	5.0E − 12

References

[1] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols. Numerical computation of an analytic singular value decomposition by a matrix valued function. *Numer. Math.*, 60:1–40, 1991.

[2] L. Dieci and T. Eirola. On smooth decomposition of matrices. *SIAM J. Matrix Anal. Appl.*, 20:800–819, 1999.

[3] L. Dieci and A. Papini. Continuation of eigendecompositions. *Future Generation Computer Systems*, 19:1125-1137, 2003.

[4] L. Dieci, M.G. Gasparo and A. Papini. Continuation of Singular Value Decompositions. *Mediterranean Journal of Mathematics*, 2:179-203, 2005.

[5] H. Gingold and P.F. Hsieh. Globally analytic triangularization of a matrix function. *Linear Algebra Appl.*, 169:75–101, 1992.

[6] G. H. Golub and C. F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 2nd edition, 1989.

[7] M. Hirsch. *Differential Topology.* Springer–Verlag, New York, 1976.

[8] Tosio Kato. *A short introduction to perturbation theory for linear operators.* Springer-Verlag, New York, 1982.

[9] Y.A. Kuznetsov. *Elements of Applied Bifurcation Theory.* Springer-Verlag, New York, 1995.

[10] B.M. Parlett. *The Symmetric Eigenvalue Problem.* Classics in Applied Mathematics, 20, SIAM, Philadelphia, 1998.

[11] W. Rheinboldt. On the computation of multi-dimensional solution manifolds of parametrized equations. *Numer. Math.*, 53:165–181, 1988.

[12] Y. Saad. *Iterative Methods for Large Linear Systems.* SIAM, 2nd edition, 2003.

[13] W. M. Wonham. *Linear Multivariable Control: a Geometric Approach.* Springer-Verlag, New York, 2nd Edition, 1979.

Luca Dieci
School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 U.S.A.
e-mail: `dieci@math.gatech.edu`

M. Grazia Gasparo
Dip. Energetica S. Stecco, Univ. of Florence, via C. Lombroso 6/17, 50134 Florence, Italy
e-mail: `mariagrazia.gasparo@unifi.it`

Alessandra Papini
Dip. Energetica S. Stecco, Univ. of Florence, via C. Lombroso 6/17, 50134 Florence, Italy
e-mail: alessandra.papini@unifi.it