A Continuous Jacobi-like Approach to the Simultaneous Reduction of Real Matrices

Moody T. Chu[1]
Department of Mathematics
North Carolina State University
Raleigh, North Carolina 27695-8205

December 1989

[1] This work was supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under Contract W-31-109-Eng-38, while the author was spending his sabbatical leave at Argonne National Laboratory.

Abstract

The problem of simultaneous reduction of real matrices by either orthogonal similarity or orthogonal equivalence transformations is considered. Based on the Jacobi idea of minimizing the sum of squares of the complementary part of the desired form to which matrices are reduced, the projected gradient method is used in this paper. It is shown that the projected gradient of the objective function can be formulated explicitly. This gives rise to a system of ordinary differential equations that can be readily solved by numerical software. The advantages of this approach are that the desired form to which matrices are reduced can be almost arbitrary, and that if a desired form is not attainable, then the limit point of the corresponding differential equation gives a way of measuring the distance from the best reduced matrices to the nearest matrices that have the desired form. The general procedure for deriving these differential equations is discussed. Some applications are given.

1 Introduction

In this paper, we are interested mainly in real-valued matrices although the discussion in the sequel can be generalized to the complex-valued case. The generalization will become clear at the end of this paper.

Let $R^{n \times n}$ denote the space of $n \times n$ real-valued matrices and let $G(n)$ denote the group of all nonsingular matrices in $R^{n \times n}$. The following is a classical problem in the field of algebra:

Problem 1 *Given k arbitrary matrices $A_1, \ldots, A_k \in R^{n \times n}$, identify the similarity class (orbit)*

$$\{(B_1, \ldots, B_k) | B_i = T^{-1} A_i T, i = 1, \ldots, k; T \in G(n)\} \tag{1}$$

under the action of $G(n)$.

Let $S(n)$ denote the subspace of all symmetric matrices in $R^{n \times n}$ and let $O(n)$ denote the subgroup of all orthogonal matrices in Gn. Then an associated problem is:

Problem 2 *Given k arbitrary matrices $A_1, \ldots, A_k \in S(n)$, identify the similarity class*

$$\{(B_1, \ldots, B_k) | B_i = Q^T A_i Q, i = 1, \ldots, k; Q \in O(n)\} \tag{2}$$

under the action of $O(n)$.

It is known that the classification of similarity classes of k-tuples of matrices can be reduced to the classification of simultaneous similarity of commuting pairs of matrices [14]. Only recently have the complex-valued versions of the above two long-standing problems been theoretically solved in the paper [10]. The technique used is highly algebraic in nature. Roughly speaking, the orbit is determined by the values of certain rational functions in the entries of $A_1, \ldots, A_k$. Various problems in which the classification of orbits is needed and various results for Problem 1 can be found in [10] and the references contained therein. But no numerical procedure has ever been attempted.

Because of concerns about numerical stability, numerical analysts usually prefer orthogonal transformations to general invertible transformations. Therefore, it is of practical interest to consider the following problem:

Problem 3 *Given k arbitrary matrices $A_1, \ldots, A_k \in R^{n \times n}$, identify the similarity class*

$$\{(B_1, \ldots, B_k) | B_i = Q^T A_i Q, i = 1, \ldots, k; Q \in O(n)\} \tag{3}$$

The only difference between Problem 2 and Problem 3 is that we have replaced symmetric matrices by general matrices. We have reasons to believe that this replacement makes Problem 3 harder to analyze. We mention, for example, the well-known real Schur decomposition theorem [17, p.362] that is related to the case $k = 1$ in Problem 3:

Theorem 1.1 (RSD) *If $A \in R^{n \times n}$, then there exists an orthogonal matrix $Q \in O(n)$ such that $Q^T AQ$ is upper quasi-triangular, that is, $Q^T AQ$ is block upper-triangular where each diagonal block is either a 1×1 matrix or a 2×2 matrix having complex conjugate eigenvalues.*

Obviously the RSD Theorem has not yet fully identified the orthogonal simiiliarity orbit of a general matrix $A \in R^{n \times n}$. Only when A is symmetric, then the similarity orbit of A in $S(n)$, being diagonalizable through the transformation $Q^T AQ$, is perfectly classified.

We usually are interested in identifying a matrix by its *canonical form*. In application, quite often a canonical form is meant to be of a special *matrix structure*. In this paper, we shall further require the canonical form be such that *all matrices having that structure form a linear subspace in $R^{n \times n}$*. The structure could be, for instance, a diagonal matrix, a bidiagonal matrix, an upper triangular matrix, and so on. The Jordan canonical form, however, is out of our consideration because Jordan matrices do not form a linear subspace. A different view of Problem 3, therefore, is to consider the following problem:

Problem 4 *Given k specified (but possibly the same) canonical forms for matrices in $R^{n \times n}$, determine if the orbit of k matrices $A_1, \ldots, A_k \in R^{n \times n}$ under the action of $O(n)$ contains an element such that each $Q^T A_i Q$ has the specified structure.*

We mention a related but slightly different problem to illustrate an application of Problem 4. For decades, the problem of simultaneous diagonalization of two symmetric matrices has received much attention. See, for example, [2, 17, 20, 27, 28], and a historical survey [29]. A classical result in this direction is stated as follows:

Theorem 1.2 *If A is symmetric and B is symmetric and positive definite, then there exists a nonsingular X such that both X^TAX and X^TBX are diagonal matrices.*

We note that most of the diagonalization processes for symmetric matrices involve nonsingular (congruence) transformations which usually are not orthogonal. This is partly due to the reason that orthogonal transformations are too limited to result in the diagonal form. But then it is curious to know how much reduction orthogonal transformations can accomplish.

The type of transformation Q^TAQ with $Q \in O(n)$ will be referred to, henceforth, as the *(real) orthogonal similarity transforamtion.* In numerical analysis there is another type of transformation, Q^TAZ with $Q, Z \in O(n)$, which will be referred to as the *(real) orthogonal equivalence transformation.* The importance of the real orthogonal equivalence transformation can been seen from the singular value decomposition theorem [17, p.71]:

Theorem 1.3 (SVD) *If $A \in R^{m\times n}$, then there exist orthogonal matrices $Q \in O(m)$ and $Z \in O(n)$ such that Q^TAZ is a diagonal matrix. ($\Sigma \in R^{m\times n}$ is understood to be a diagonal matrix if and only if $\sigma_{ij} = 0$ whenever $i \neq j$.)*

Analogous to (3), the *equivalence orbit* of any given k arbitrary matrices $A_1, \ldots, A_k \in R^{m\times n}$ (under the action of $O(m)$ and $O(n)$) is defined to be the set

$$\{(B_1, \ldots, B_k) | B_i = Q^TA_iZ, i = 1, \ldots, k; Q \in O(m), Z \in O(n)\}. \qquad (4)$$

Motivated by Problem 4, we ask the following question:

Problem 5 *Given k specified (but possibly the same) canonical forms for matrices in $R^{m\times n}$, determine if the equivalence orbit of k matrices $A_1, \ldots, A_k \in R^{m\times n}$ contains an element such that each Q^TA_iZ has the specified structure.*

The SVD Theorem settles the special case $k = 1$ in Problem 5. When $k = 2$, then Problem 5 is partially answered by the so called generalized real Schur decomposition theorem [17, p.396]:

Theorem 1.4 (GRSD) *If $A, B \in R^{n\times n}$, then there exist orthogonal matrices Q and Z such that Q^TAZ is upper quasi-triangular and Q^TBZ is upper triangular.*

We should distinguish GRSD from another analogous but different application known as the generalized singular value decomposition theorem [22], [30]:

Theorem 1.5 (GSVD) *If $A \in R^{m \times n}$ and $B \in R^{p \times n}$, then there exist orthogonal $U \in O(m)$, $V \in O(p)$ and invertible $X \in R^{n \times n}$ such that $U^T AX$ and $V^T BX$ are diagonal matrices.*

We note that besides the generality of dimensions of A and B, the GSVD is fundamentally different from GRSD in that the orthogonal matrix Q is not the same for A as for B and that the orthogonalilty of the matrix Z is replaced by the nonsingularity.

All the aforementioned special cases of either Problem 4 or Problem 5 have found significant applications in numerical analysis. Enormous amount of efforts have already been devoted to the study of these special matrix decompositions (See [17] and the references cited therein). We mention just one example — Based on Theorem 1.4 a numerically stable method, called the QZ algorithm [21], has been developed to solve the important generalized eigenvalue problem $Ax = \lambda Bx$. On the other hand, for the general cases of either Problem 4 or Problem 5, we find little is known in the literature. Conceibably, when more matrices are involved, the simultaneous reduction problem becomes more difficult both theoretically and numerically.

In this paper we recast the simultaneous reduction problem as an equality-constrained optimization problem and apply the projected gradient method. We develop a differential equation approach that can be used as a numerical method for answering both Problem 4 and Problem 5. Our approach is flexible in at least two aspects:

1. The differential equations for various types of *canonical forms* can easily be derived within a uniform framework for a given k.

2. The framework can easily be modified if k is changed.

In view of these advantages, we think we have established a tool by which one may experiment with combinations of many different canonical forms with only slight modifications in the computer program. Furthermore, if the desired form is not attainable, then the limit point of the corresponding differential equation gives a way of measuring the distance from the best reduced matrices to the *nearest* matrices that have the desired form. This information sometimes is useful in applications.

The QR algorithm, the SVD algorithm and the QZ algorithm are a few of the iterative methods that play very prominent roles in matrix computations. Earlier the author has developed differential equations to model these iterative processes. Some references can be found in the review paper [5]. Most of the ideas there have been based on the fact that a finite, nonperiodic Toda lattice is a continuous analog of the QR algorithm [26].

The differential equation approach developed in this paper is based on the idea of using the projected gradient method to minimize the sum of squares of the *undesired* portions of the matrices. So in some sense our approach is a continuous analog of the so called Jacobi method for symmetric eigenvalue problems. A collection of variations and references of the Jacob method can be found in [17, p.444–459]. In the past, attempts have been made to extend the Jacobi iteration to other classes of matrices and to push through corresponding convergence results. But success has been reported only for normal matrices [16] which, then, was employed to solve the closest normal matrix problem [23]. For non-normal matrices, the situation is considerably more difficult. Simultaneous reduction of more than one general matrices is thus an even harder problem. It turns out that our differential equation approach offers a fairly easy but systematic reduction procedure. In fact, the approach is so versatile that one can examine the (similarity or equivalence) orbit of k given matrices for many different combinations of reduced forms.

In [1, p.239] the following question was raised:

> What is the simplest form to which a family of matrices depending smoothly on the parameters can be reduced by a change of coordinates depending smoothly on the parameters?

Our differential equation approach to Problem 4 and Problem 5 can be regarded as a special tool to answer this general question.

This paper is organized as follows: In Section 2, we first derive a general framework of constructing differential equations for Problem 4. The development is parallel to that in an earlier paper [6] where a framework was proposed for solving spectrally constrained least squares approximation problems. In particular, we show how the projected gradient can be calculated explicitly. In Section 3 we demonstrate a special application to the simultaneous diagonalization of two symmetric matrices. Differential equations for Problem 5 are derived in Section 4. In the last section, we combine techniques from both Section 2 and Section 4 to show how the argument can be generalized to complex-valued case. An application to the closest normal matrix problem is discussed there.

2 Orthogonal Similarity Transformation

In this section we develop an ordinary differential equation that can be used to solve Problem 4 numerically.

We first define some notation. Let $A_i \in R^{n\times n}, i = 1,\ldots,k$ denote k given matrices. For each i, let $V_i \subset R^{n\times n}$ denote the subspace of all matrices having the specified form to which A_i is supposed to be reduced. We shall use the Frobenius inner product

$$\langle X, Y\rangle := \text{trace}(XY^T) = \sum_{i,j=1}^{n} x_{ij}y_{ij} \tag{5}$$

and the Frobenius matrix norm $||X|| := \langle X, X\rangle^{1/2}$ in the space $R^{n\times n}$. Given any $X \in R^{n\times n}$, its projection onto the subspace V_i is denoted as $P_i(X)$. For any matrix $X \in R^{n\times n}$, we define the residual operator

$$\alpha_i(X) := X^T A_i X - P_i(X^T A_i X). \tag{6}$$

We remark here that the choice of the subspace V_i can be quite arbitrary. For example, V_1 may be taken to be the subspace of all diagonal matrices, V_2 the subspace of all upper Hessenberg matrices, and so on. Our idea in approaching Problem 4 is to consider the following optimization problem:

Problem 6

$$\begin{aligned} &\textit{Minimize} \qquad F(Q) := \frac{1}{2}\sum_{i=1}^{k} \|\alpha_i(Q)\|^2 \qquad (7)\\ &\textit{Subject to} \qquad Q^T Q = I. \end{aligned}$$

That is, while moving along the orthogonal similarity orbit of the given matrices $A_1, \ldots, A_k$, we want to minimize the total distance between the point $Q^T A_i Q$ and the subspace V_i for all i. One may regard Problem 6 as a standard equality-constrained optimization problem and thus solves the problem by many existing numerical algorithms found in, for example, [15]. In doing so, however, one has to interpret a matrix equation as a collection of n^2 nonlinear equations. The computation of derivatives in the *unpacked form* proves to be very inconvenient. In the following we discuss an interesting geometric approach that preserves matrix operations and leads to the construction of a differential equation.

We first note that the feasible set $O(n) := \{Q | Q^T Q = I\}$ is a well-defined smooth manifold in $R^{n\times n}$. It can be shown [6] that the tangent space to $O(n)$ at any orthogonal matrix Q is given by

$$T_Q(O(n)) = QS(n)^{\perp}, \tag{8}$$

and that the orthogonal complement of $T_Q(O(n))$ in $R^{n\times n}$ is given by

$$N_Q(O(n)) = QS(n). \tag{9}$$

The Fréchet derivative of the objective function F in (7) at a general $X \in R^{n\times n}$ acting on a general $Y \in R^{n\times n}$ can be calculated as follows:

$$\begin{aligned} F^{'}(X)Y &= \sum_{i=1}^{k} \langle \alpha_i(X), \alpha_i^{'}(X)Y \rangle \\ &= \sum_{i=1}^{k} \langle \alpha_i(X), X^T A_i Y + Y^T A_i X - P_i(X^T A_i Y + Y^T A_i X) \rangle \\ &= \sum_{i=1}^{k} \langle A_i^T X \alpha_i(X) + A_i X \alpha_i^T(X), Y \rangle. \end{aligned} \tag{10}$$

In the second equation above we have used the fact that the projections P_i are linear. In the third equation above we have used the fact that $\alpha_i(X)$ is perpendicular to V_i. We also have used the adjoint property

$$< X, YZ > = < Y^T X, Z > = < AZ^T, Y > \tag{11}$$

to rearrange terms. The equation (10) suggests that with respect to the Frobenius inner product, we may interpret the *gradient* of F at a general point X as the matrix

$$\nabla F(X) = \sum_{i=1}^{k} (A_i^T X \alpha_i(X) + A_i X \alpha_i^T(X)). \tag{12}$$

Since $R^{n\times n} = T_Q O(n) \oplus N_Q O(n)$, every element $X \in R^{n\times n}$ has a unique orthogonal splitting

$$X = Q\{\frac{1}{2}(Q^T X - X^T Q)\} + Q\{\frac{1}{2}(Q^T X + X^T Q)\} \tag{13}$$

as the sum of elements from $T_QO(n)$ and $N_QO(n)$. Therefore, the projection $g(Q)$ of $\nabla F(Q)$ into to the tangent space $T_QO(n)$ can be calculated as follows:

$$\begin{aligned} g(Q) &= \frac{Q}{2}\{Q^T\nabla F(X) - \nabla F(X)^T Q\} \\ &= \frac{Q}{2}\sum_{i=1}^{k}([Q^T A_i^T Q, \alpha_i(Q)] + [Q^T A_i Q, \alpha_i^T(Q)]). \end{aligned} \tag{14}$$

In (14) we have adopted the Lie bracket notion $[X,Y] := XY - YX$.

Now that $g(Q)$ is tangent to the manifold $O(n)$ (Note that the big summation in (14) ends up with a skew-symmetric matrix), the vector field

$$\frac{dQ}{dt} = -g(Q) \tag{15}$$

defines a flow on the manifold $O(n)$. By the way we construct it, this flow moves in the *steepest descent* direction for the objective function $F(Q)$.

For convenience, we define

$$X_i(t) := Q(t)^T A_i Q(t) \tag{16}$$

for $i = 1, \ldots, k$. Upon differentiation and substitution, we find that each $X_i(t)$ must satisfy the ordinary differential equation:

$$\begin{aligned} \frac{dX_i}{dt} &= \frac{dQ^T}{dt}A_iQ + Q^TA_i\frac{dQ}{dt} \\ &= \left[X_i, \frac{1}{2}\sum_{j=1}^{k}([X_j, P_j^T(X_j)] + [X_j^T, P_j(X_j)]\right]. \end{aligned} \tag{17}$$

It is worthwhile to note that the above arguments can be reversed [7]. That is, any solution $X(t)$ to (17) can be written in the form of (16) with $Q(t)$ satisfying (15). We note also that the big summation in the first bracket of (17) is always a skew-symmetric matrix. Therefore, the flow $X_i(t)$ naturally stays on the *isospectral surface* $M(A_i) := \{Q^T A_i Q | Q \in O(n)\}$ if it starts from an initial value $X_i(0) \in M(A_i)$. One obvious choice of the initial value will be $X_i(0) = A_i$. The differential system (17) may be integrated by many readily available ODE solvers. In doing so, we are following a flow that has the *potential* of solving Problem 4 for any prescribed set of canonical forms.

Even if the prescribed canonical form is not attainable, the solution flow $X(t)$ still provides a systematic way of simultaneously reducing the norm of the residuals. It is in this sense we think our flow is a continuous realization of the classical Jacobi approach.

We observe from (17) that the vector field for each component is, in general, a homogeneous polynomial of degree 3. Such a complicated dynamical system is difficult to analyze theoretically. The initial value problem, however, is easy to solve numerically. By varying the subspaces V_i (and, correspondingly, the projections P_i), therefore, we have established an instrument for testing numerically if a given set of matrices $A_1, \ldots, A_k$ can be simultaneously reduced to certain desired forms through orthogonal similarity transformations. We think the versatility of our approach is quite interesting.

As an application, we now consider the case $k = 1$ in the differential equation (17) and comment on the Jacobi algorithm for eigenvalue problems. The initial value problem to be solved is given by

$$\begin{aligned} \frac{dX}{dt} &= \left[X, \frac{[X, P_1^T(X)] - [X, P_1^T(X)]^T}{2}\right] \\ X(0) &= A_1 \text{ (general).} \end{aligned} \tag{18}$$

We first choose V_1 to be the subspace of all upper triangular matrices. According to our theory, the solution of (18) defines an isospectral flow that moves (for $t \geq 0$) to minimize the norm of the strictly lower triangular elements. This idea clearly generalizes that of the Jacobi method for symmetric eigenvalue problems. Indeed, we note from (18) that if X is symmetric, then so is dX/dt. If the initial value A_1 is symmetric, then so is $X(t)$ for all t. In this case, we may be better off if V_1 is chosen to be the subspace of all diagonal matrices so that the norm of *all* off-diagonal elements is being minimized. With this choice, the differential system (18) becomes

$$\begin{aligned} \frac{dX}{dt} &= [X, [X, diag(X)]] \\ X(0) &= A_1 \text{ (symmetric)} \end{aligned} \tag{19}$$

where $diag(X)$ denotes the diagonal matrix $diag\{x_{11}, x_{22}, \ldots, x_{nn}\}$. The solution flow to (19) is a continuous analog of the classical Jacobi method for symmetric eigenvalue problems.

It remains to determine to where a solution flow of (18) will converge. Our numerical experience indicates that for a general initial matrix A_1, the solution flow of (18) in minimizing the norm of the strictly lower triangular elements may converge to a limit point which does not even look like an upper triangular matrix. This can easily be demonstrated by a numerical example: Starting with the initial matrix

$$A_1 = \begin{bmatrix} 1.0000 & 3.0000 & 5.0000 & 7.0000 \\ -3.0000 & 1.0000 & 2.0000 & 4.0000 \\ 0.0000 & 0.0000 & 3.0000 & 5.0000 \\ 0.0000 & 0.0000 & 0.0000 & 4.0000 \end{bmatrix}, \tag{20}$$

we integrated the equation (18) by using the subroutine ODE in [24] with local tolerance set at 10^{-12}. We assumed that convergence had occurred if the difference between two consecutive output values (at intervals of 10) was less than 10^{-10}. We found the solution flow converged to the limit point

$$\begin{bmatrix} 2.2500 & 3.3497 & 3.1713 & 2.8209 \\ -0.3506 & 2.2500 & 8.0562 & 6.1551 \\ 0.6247 & -0.8432 & 2.2500 & 3.2105 \\ -0.0846 & 0.2727 & -0.3360 & 2.2500 \end{bmatrix}. \tag{21}$$

Although the initial matrix (20) is an upper quasi-triangular matrix (as is defined in Theorem 1.3), the limit point (21) is a full matrix. We observed also that along the solution flow, the norm of the strictly lower triangular elements had been reduced monotonically from 3 to 1.1910. This example confirms that the upper quasi-triangular matrix guaranteed by the RSD theorem is not necessarily a stationary point when minimizing *all* the strictly lower triangular elements [25].

We shall say that a matrix is of *structure* B if it is block upper-triangular and if all diagonal blocks are 2×2 matrices except possibly the last one which is 1×1. We note that structure B is more general than upper quasi-triangular. If a matrix A_1 can be reduced by orthogonal similarity transformations to be of structure B, then eigenvalues of A_1 are readily known. Toward this end, we may choose V_1 involved in equation (18) to be the subspace of all matrices of structure B. Our numerical experiments with structure B seems to indicate that the ω-limit set of any solution flow contains only a singleton which is of structure B. Thus, we conjecture that structure B is always attainable. The proof of this dynamics and the experimentation with the associated discrete Jacobi-type algorithm are currently under investigation and we shall report the result elsewhere.

Meanwhile, the classification of all critical points for (19) has been completely analyzed in [8]. It is worth noting that the diagonal matrices are proved to be the only *stable* equilibrium points for the dynamical system (19). Furthermore, any of these (isospectral) diagonal matrices corresponds to a *global* minimizer of Problem 6. Once again, this phenomenon is very analogous to that known for the classical Jacobi method.

3 A Nearest Commuting Pair Problem

In this section we discuss another application of the differential system (17). Let A_1 and A_2 be two given matrices in $R^{n\times n}$. In general, A_1 and A_2 do not commute. It is interesting to determine how far the pair (A_1, A_2) is away from being commutable. This problem can be formulated as follows:

Problem 7

$$\begin{array}{ll} \textit{Minimize} & F(E_1, E_2) := \frac{1}{2}\sum_{i=1}^{2} \|E_i - A_i\|^2 \\ \textit{Subject to} & E_1E_2 - E_2E_1 = 0. \end{array} \tag{22}$$

Again, Problem 7 is a typical equality constrained optimization problem and can be solved by many available methods. In applying the method of Lagrange multipliers, for example, we need to solve the following system of matrix equations

$$\begin{array}{rcl} E_1 - A_1 + \Lambda E_2^T - E_2^T\Lambda & = & 0 \\ E_2 - A_2 + E_1^T\Lambda - \Lambda E_1^T & = & 0 \\ E_1E_2 - E_2E_1 & = & 0 \end{array} \tag{23}$$

for the variables E_1, E_2 and the multiplier Λ. This approach suffers from some obvious difficulties.

Suppose both A_1 and A_2 are symmetric. A problem slightly less general than Problem 7 is to determine how far (A_1, A_2) is away from a symmetric, commuting pair [3]. Let E_1 and E_2 be any symmetric, commuting pair. We shall assume further that at least one of these two matrices has distinct eigenvalues (This is the generic case). It is not difficult to show that E_1 and E_2 can be simultaneously diagonalized by a Q^T-Q transformation [13, p.222, Corollary 1]. Let $D_i = Q^TE_iQ, i = 1, 2$ be the diagonal matrices for some orthogonal matrix Q. We observe from the relation

$$\sum_{i=1}^{2} \|E_i - A_i\|^2 = \sum_{i=1}^{2} \|D_i - Q^TA_iQ\|^2 \tag{24}$$

that the left-hand side of (24) will be minimized if one first finds an orthogonal matrix Q such that the matrices Q^TA_iQ are as close to diagonal matrices as possible, and then sets $D_i = diag(Q^T A_i Q)$. Thus the problem of finding a nearest commuting pair to a given pair of symmetric matrices is boiled down to the problem of simultaneous reduction of off-diagonal elements of the given pair by orthogonal transformations. The latter problem fits in as a special case of our general framework in the preceding section. We simply proceed as follows:

Both $V_i, i = 1, 2$ are taken to be the subspace of diagonal matrices. According to (17), the descent flow is given by the initial value problem:

$$\begin{aligned} \frac{dX_i}{dt} &= \left[X_i, \sum_{j=1}^{2}[X_j, diag(X_j)]\right] \\ X_i(0) &= A_i, \;\; i = 1, 2 \end{aligned} \tag{25}$$

since both X_i and $diag(X_i)$ are symmetric matrices.

In the event that A_1 and A_2 cannot be diagonalized simultaneously, the limit point of the flow gives a way of measuring the distance from (A_1, A_2) to the nearest commuting pair (See (24)). Comparing with the system (19), one finds immediately that (25) is a direct generalization of the Jacobi algorithm. It is known the straightforward "diagonalize one, then diagonalize the other" approach for simultaneously diagonalizing pairs of symmetric matrices is subject to numerical hazards that may prevent convergence [3]. We think our approach gives a new twist to the algorithm.

4 Orthogonal Equivalence Transformation

In this section we develop an ordinary differential equation that can be used to solvee Problem 5 numerically. Our approach is analogous to that in Section 2.

Let $A_i \in R^{m\times n}, i = 1, \ldots, k$ be given matrices. For each i, let $V_i \subset R^{m\times n}$ denote the subspace of all matrices having the specified form to which A_i is supposed to be reduced. The projection operator from $R^{m\times n}$ to V_i is denoted by P_i. For any $X \in R^{m\times m}$ and $Y \in R^{n\times n}$, we define

$$\alpha_i(X,Y) := X^T A_i Y - P_i(X^T A_i Y). \tag{26}$$

We reformulate Problem 5 as:

Problem 8

$$\begin{aligned} &\textit{Minimize} && F(Q,Z) := \frac{1}{2}\sum_{i=1}^{k} \|\alpha_i(Q,Z)\|^2 \\ &\textit{Subject to} && Q^T Q = I_n \\ & && Z^T Z = I_m. \end{aligned} \tag{27}$$

It will prove useful to consider the product topology by introducing the induced Frobenius inner product

$$\langle (X_1,Y_1),(X_2,Y_2)\rangle_P := \langle X_1, X_2\rangle + \langle Y_1, Y_2\rangle. \tag{28}$$

on the space $R^{m\times m} \times R^{n\times n}$. The feasible set of Problem 8 is considered to be the product $O(m) \times O(n)$. The tangent space to $O(m)\times O(n)$ at $(Q,Z) \in O(m)\times O(n)$ is given by

$$T_{(Q,Z)}O(m)\times O(n) = QS(m)^{\perp} \times ZS(n)^{\perp}, \tag{29}$$

and the normal space is given by

$$N_{(Q,Z)}O(m)\times O(n) = QS(m)\times ZS(n). \tag{30}$$

The Fréchet derivative of the objection function in (27) at a general $(X, Y) \in R^{m\times m} \times R^{n\times n}$ acting on a general $(H, K) \in R^{m\times m} \times R^{n\times n}$ is

$$\begin{aligned} & F^{'}(X,Y)(H,K) \\ = & \sum_{i=1}^{k} \langle \alpha_i(X,Y), H^T A_i Y + X^T A_i K - P_i(H^T A_i Y + X^T A_i K)\rangle \\ = & \sum_{i=1}^{k} \left(\langle A_i Y \alpha_i^T(X,Y), H\rangle + \langle A_i^T X \alpha_i(X,Y), K\rangle \right). \end{aligned} \tag{31}$$

Therefore, with respect to the induced Frobenius inner product, we may interpret the *gradient* of F at (X, Y) as the pair:

$$\nabla F(X,Y) = \left(\sum_{i=1}^{k} A_i Y \alpha_i^T(X,Y), \sum_{i=1}^{k} A_i^T X \alpha_i(X,Y) \right). \tag{32}$$

We note that there is a considerable similarity between (12) and (32).

Because of the product topology, we may use the same principle as in (14) to calculate the projection $g(Q, Z)$ of $\nabla F(Q, Z)$ into the tangent space $T_{(Q,Z)}O(m) \times O(n)$. After simplification, we claim that

$$\begin{aligned} g(Q,Z) \; = \; & \left(\frac{Q}{2} \sum_{i=1}^{k} (Q^T A_i Z \alpha_i^T(Q,Z) - \alpha_i(Q,Z) Z^T A_i^T Q), \right. \\ & \left. \frac{Z}{2} \sum_{i=1}^{k} (Z^T A_i^T Q \alpha_i(Q,Z) - \alpha_i^T(Q,Z) Q^T A_i Z) \right). \end{aligned} \tag{33}$$

Therefore, the vector field

$$\frac{d(Q,Z)}{dt} = -g(Q,Z) \tag{34}$$

defines a steepest descent flow on the manifold $O(m) \times O(n)$ for the objective function $F(Q, Z)$.

For $i = 1, \ldots, k$, we define

$$X_i(t) := Q(t)^T A_i Z(t) \tag{35}$$

where $(Q(t), Z(t))$ satisfies the differential equation (34). Upon differentiation and substitution, it is not difficult to see that each $X_i(t)$ satisfies the equation

$$\frac{dX_i}{dt} = \sum_{j=1}^{k} \left\{ X_i \frac{X_j^T P_j(X_j) - P_j^T(X_j) X_j}{2} + \frac{P_j(X_j) X_j^T - X_j P_j^T(X_j)}{2} X_i \right\}. \tag{36}$$

By specifying the initial values, say $X_i(0) = A_i$, and the subspaces V_i, we now have an instrument in hand to explore various simultaneous reduction problems numerically simply by integrating the equation (36).

One special case of $k = 1$ is worth mentioning: Take V_1 to be the subspace of all diagonal matrices in $R^{m \times n}$ (In a rectangular matrix, the extra rows or columns are filled with zero). Then the equation (36) becomes

$$\frac{dX}{dt} = \frac{1}{2} \left\{ X \left(X^T (diagX) - (diagX)^T X \right) + \left((diagX) X^T - X (diagX)^T \right) X. \right\} \tag{37}$$

In spirit, the differential equation (37) is parallel to a Jacobi-like approach to the singular value decomposition [17, p.455]. The stability property of all equilibrium points of (37) can be further analyzed. In fact, it can be proved that diagonal matrices (of the singular values) are the only stable equilibrium points. Readers are referred to [6] and [9] for more details.

5 The Closest Normal Matrix Problem

All the techniques discussed in Section 2 and Section 4 can be generalized to the complex-valued case. Without given too much repetition, we demonstrate in this section how this generalization should be done by working on the closest normal matrix problem.

The determination of a closest normal matrix to a given square complex matrix has already recieved consierable attention (See [19] and the references therein). This problem has only recently been completely solved (in the Frobenius norm) in [11], and independently in [23]. We shall cast this problem into our framework from which we obtain new and clear geometric characterization of the first and the second order optimality condition.

Let $C^{n\times n}$ denote the space of $n \times n$ complex-valued matrices, $U(n)$ the group of all unitary matrices in $C^{n\times n}$ and $D(n)$ the subspace of all diagonal matrices in $C^{n\times n}$. We recall the well-known fact that [18]:

Theorem 5.1 *A matrix $Z \in C^{n\times n}$ is normal if and only if there exists a unitary $U \in U(n)$ such that $U^*ZU \in D(n)$.*

Therefore, given an arbitrary matrix $A \in C^{n\times n}$, the closest normal matrix problem can be formulated as

Problem 9

$$\begin{array}{ll} \textit{Minimize} & F(U,D) := \frac{1}{2}\|A - UDU^*\|^2 \\ \textit{Subject to} & U \in U(n) \textit{ and } D \in D(n). \end{array} \tag{38}$$

with the Frobenius norm $||Z||^2 := \sum_{i,j=1}^n |z_{ij}|^2$.

We note that in the minimization of (38), the two matrix variables U and D are considered to be independent of each other. Let $Z := UDU^*$, however, we observe the relationship

$$||A - Z||^2 = ||U^*AU - D||^2 \tag{39}$$

holds. Obviously, for any given $U \in U(n)$, the best $D \in D(n)$ that will minimize the right-hand side of (39) is $diag(U^*AU)$. Therefore, at *global* extrema, Problem 9 is equivalent to

Problem 10

$$\begin{array}{ll} \textit{Minimize} & F(U) = \frac{1}{2}\|U^*AU - diag(U^*AU)\|^2 \\ \textit{Subject to} & U^*U = I. \end{array} \tag{40}$$

Since unitary transformations do not alter the Frobenius norm of a matrix, minimizing the sum of squares of off-diagonal elements of a matrix is equivalent to maximizing the sum of squares of diagonal elements. From (40), we conclude that the closest normal matrix is characterized by the following theorem [4], [11]:

Theorem 5.2 *Let $A \in C^{n\times n}$ and let $Z = UDU^*$ where $U \in U(n)$ and $D \in D(n)$. Then Z is a closest normal matrix to A in the Frobenius norm if and only if*

1. *The unitary matrix U maximizes $||diag(V^*AV)||$ among all $V \in U(n)$.*
2. *The diagonal matrix D is such that $D = diag(U^*AU)$.*

We see from Problem 10 and Theorem 5.2 that except for complex-valued matrices, the situation is just like that discussed in Section 2 — We want to minimize the norm of the off-diagonal elements by unitary similarity transformations on A.

The ideas discussed in Section 2 can be applied almost without change to the complex-valued case. We briefly describe our procedure as follows: We shall regard $C^{n\times n}$ as the vector space $R^{n\times n} \times R^{n\times n}$ over the field of real numbers. That is, we shall identify the complex matrix Z as a pair of real matrices $(\Re Z, \Im Z)$, where $\Re Z$ and $\Im Z$ represent the real and the imaginary part of Z, respectively. The inner product on $C^{n\times n}$ is defined by

$$\langle X, Y\rangle_C := \langle \Re X, \Re Y\rangle + \langle \Im X, \Im Y\rangle \tag{41}$$

We note that $\langle Z, Z\rangle_C = ||Z||^2$. The topology imposed on $C^{n\times n}$ by (41) resembles that on $R^{m\times m} \times R^{n\times n}$ given by (28). We may thus also take advantage of the techniques developed in Section 4. In this context, the analog to (8) is that the tangent space to $U(n)$ at any unitary matrix U is given by

$$T_U U(n) = UH(n)^\perp \tag{42}$$

where $H(n)$ is the collection of all Hermitian matrices in $C^{n\times n}$. Furthermore, identifying $Z = (\Re Z, \Im Z)$, one can calculate the Fréchet derivative and the gradient for the objective function F in (40). It is not difficult to prove that all the calculation can be carried out formally just as in the real-valued case. In particular, one can show that the projected gradient $g(U)$ of F onto the (real) manifold $U(n)$ is given by

$$g(U) = \frac{U}{2}\left\{[diag(U^*AU), U^*A^*U] - [diag(U^*AU), U^*A^*U]^*\right\}. \qquad (43)$$

From (43), we obtain the following first order optimality condition.

Theorem 5.3 *Let $W := U^*AU$. Then for U to be a stationary point of Problem 10, it is necessary that*

$$[diag(W), W^*] = [diag(W), W^*]^*. \qquad (44)$$

Let w_{ij} denote the (i,j)-component of W. It is easy to see that condition (44) is equivalent to

$$\overline{w}_{ji}(w_{ii} - w_{jj}) = w_{ij}(\overline{w}_{jj} - \overline{w}_{ii}). \qquad (45)$$

If we define a matrix $H = (h_{ij})$ by

$$h_{ij} = \begin{cases} \frac{w_{ij}}{w_{ii}-w_{jj}} & \text{if } w_{ii} \neq w_{jj} \\ 0 & \text{if } w_{ii} = w_{jj} \end{cases}, \qquad (46)$$

condition (45) is then equivalent to assuming that H is Hermitian. This observation is in concordance with the notion of ΔH-*matrix* introduced in [11, 12, 19, 23]. We think our derivation, being different from those done in the literature, is of interest in its own right.

Furthermore, the explicit form of the projected gradient (43) can significantly facilitate the computation of the projected Hessian on the tangent space of $U(n)$. The projected Hessian, needed in describing second order optimality conditions, usually are formulated from the Lagrangian function (See, [15, p.80]). For a general nonlinear optimization problem, rarely is the closed form of the projected Hessain available. In our context, however, we can derive the explicit projected Hessian without using the Lagrangian function.

We first extend the projected gradient function g formally to the entire space $C^{n\times n}$, i.e., we assume the equation (43) is defined for general complex matrices. Since the extended g is smooth, we may formally take its Fréchet derivative. In [6] we have observed that the quadratic form of the *extended* Fréchet derivative applied to tangent vectors corresponds exactly to the projected Hessian of the Lagrangian function. We recall that the tangent space of our feasible $U(n)$ is given by $UH(n)^{\perp}$. Therefore, we are able to calculate the quadratic form

$$\langle UK, g'(U)UK\rangle = \langle [diag(W), K] - diag([W,K]), [W,K]\rangle_C \tag{47}$$

with unitary U and skew-Hermitian K. In this way, we establish a second order optimality condition for Problem 10:

Theorem 5.4 *Let $W := U^*AU$. Then necessary (sufficient) conditions for $U \in U(n)$ to be a local minimizer of Problem 10 are that:*

1. *The matrix $[diag(W), W^*]$ is Hermitian.*
2. *The quadratic form $\langle [diag(W), K] - diag([W,K]), [W,K]\rangle_C$ is nonnegative (positive) for every skew-Hermitian matrix K.*

We note that the approach in [23] utilized the Lagrangian function with a Hermitian matrix as the Lagrange multipliers. The second order condition (either formula (12) or formula (15) in [23]) also involved the Lagrangian multipliers. Our description in the above theorem does not need any information of the Lagrangian multipliers. We believe our result in Theorem 5.4 is new in its kind.

In [23], the Jacobi algorithm for normal matrices [16] was used to solve the nearest normal matrix problem — The matrix A was first transformed by rotations into a ΔH-matrix: U^*AU, then $Z := Udiag(U^*AU)U^*$ is a putative nearest normal matrix.

Based on our preceding discussion, we now propose a continuous analog which, nevertheless, does not need to compute any shift, phase or rotation angle as did in [23]. We simply need to integrate the differential system

$$\begin{aligned} \frac{dU}{dt} &= U\frac{[W, diag(W^*)] - [W, diag(W^*)]^*}{2} \\ \frac{dW}{dt} &= \left[W, \frac{[W, diag(W^*)] - [W, diag(W^*)]^*}{2}\right] \end{aligned} \tag{48}$$

for the unitary matrix $U(t)$ and the variable $W(t) := U(t)^*AU(t)$ until convergence occurs, for then the matrix $Z := \tilde{U} diag(\tilde{W})\tilde{U}^*$ (where $\tilde{}$ denotes a limit point of (48)) will be a putative nearest normal matrix. As a numerical experiment, we integrated the systems (48) with initial values $U(0) = I$ and

$$W(0) = \begin{bmatrix} 0.7616 + 1.2296i & -1.4740 - 0.4577i \\ -1.6290 - 2.6378i & 0.1885 - 0.8575i \end{bmatrix}.$$

At $t \approx 0.9$, we concluded that convergence had occurred. The approximated limit point of (48) is given by

$$\tilde{W} \approx \begin{bmatrix} 2.2671167250 + 1.9152270486i & 0.4052706333 + 0.8956586233i \\ -0.9095591045 - 0.3730293488i & -1.3170167250 - 1.5431270486i \end{bmatrix}$$

and

$$\tilde{U} \approx \begin{bmatrix} 0.8285289301 - 0.0206962995i & 0.5350877833 - 0.1636842669i \\ -0.5350877833 - 0.1636842669i & 0.8285289301 + 0.0206962995i \end{bmatrix}.$$

The matrix $\tilde{W}$ agrees with the one given in [23] only in its diagonal elements. However, after substitution, the matrix $Z = \tilde{U} diag(\tilde{W})\tilde{U}^*$ is the same as that given in [23].

It is interesting to note that there is an obvious similarity between (18) and (48).

6 Conclusion

Two types of simultaneous reductions of real matrices by orthogonal transformations are formulated as constrained optimization problems. The objective functions are formed following the spirit in the well-known Jacobi method.

The projected gradients of the objective functions onto the feasible set can be calculated explicitly. Thus we are able to develop systems of ordinary differential equations ((17) and (36)). The framework in deriving these equations is quite general in that the number of the given matrices and the desired forms to which the given matrices are supposed to be reduced can be almost arbitrary.

By integrating the corresponding differential equation, we have thus established a general numerical tool that one can use to answer Problem 4 and Problem 5 for various reduced forms. In the event that a specified form is not attainable, the limit point of the corresponding differential equation still gives a way of measuring how far the matrices can be reduced.

The framework can also be generalized to the complex-valued case. The nearest normal matrix problem can be treated as a special application of our theory.

References

[1] V. I. Arnold, Geometrical Methods in the Theory of Ordinary Differntial Equations, 2nd ed., Springer-Verlag, New York, 1988.

[2] A. Berman and A. Ben-Israel, A note on pencils of Hermitian or symmetric matrices, SIAM J. Appl. Math., 21(1971), 51-54.

[3] A. Bunse-Gerstner, R. Byers and V. Mehrmann, Numerical methods for simultaneous diaglonalization, University of Kansas, Lawerence, preprint, 1988.

[4] R. L. Causey, On Closest Normal Matrices, Ph.D. Thesis, Department of Computer Science, Stanford University, 1964.

[5] M. T. Chu, On the continuous realization of iterative processes, SIAM Review, 30(1988), 375-387.

[6] M. T. Chu and K. R. Driessel, The projected gradient method for least squares matrix approximations with spectral constraints, SIAM J. Numer. Anal., to appear.

[7] M. T. Chu and L. K. Norris, Isospectral flows and abstract matrix factorizations, SIAM J. Numer. Anal., 25(1988), 1383-1391.

[8] K. R. Driessel, On isospectral gradient flows — solving matrix eigenproblems using differential equations, in Inverse Problems, ed. J. R. Cannon and U. Hornung, ISNM 77, Birkhäuser, 1986, 69-91.

[9] K. R. Driessel, On finding the singular values and singular vectors of a matrix by means of an isospectral gradient flow, Technical Report #87-01, Department of Mathematics, Idaho State University, 1987.

[10] S. Friedland, Simultaneous similarity of matrices, Advances in Math., 50(1983), 189-265.

[11] R. Gabriel, Matrizen mit maximaler Diagonale bei unitärer Similarität, J. Reine Angew. Math., 307/308(1979), 31-52.

[12] R. Gabriel, The normal ΔH-matrices with connection to some Jacobi-like methods, Linear Alg. Appl., 91(1987), 181-194.

[13] F. R. Gantmacher, The Theory of Matrices, Vols. 1-2, Chelsea, New York, 1959.

[14] I. M. Gelfand and V. A. Ponomarev, Remarks on the classification of a pair of commuting linear transformations in a finite dimensional space, Functional Anal. Appl., 3(1969), 325-326.

[15] P. E. Gill, W. Murray and M. H. Wright, Practical Optimization, Academic Press, London, 1981.

[16] H. H. Goldstine and L. P. Horowitz, A procedure for the diagonalization of normal matrices, J. Assoc. Comp. Mach., 6(1959), 176-195.

[17] G. H. Golub and C. F. Van Loan, Matrix Computations, 2ed., Johns Hopkins, Baltimore, 1989.

[18] R. Grone, C. R. Johnson, E. M. Sa and H. Wolkowicz, Normal matrices, Linear Alg. Appl., 87(1987), 213-225.

[19] N. J. Higham, Matrix nearest problems and applications, Proceeding of the IMA Conference on Application of Matrix Theory, S. Barnett and M. J. C. Govers, eds., Oxford University Press, 1989, to appear.

[20] K. N. Majindar, Linear combinations of Hermitian and real symmetric matrices, Lin. Alg. Appli., 25(1979), 95-105.

[21] C. B. Moler and G. W. Stewart, An algorithm for generalized matrix eigenvalue problems, SIAM J. Numer. Anal., 10(1973), 241-256.

[22] C. C. Paige and M. Saunders, Towards a generalized singular value decomposition, SIAM J. Numer. Anal., 18(1981), 398-405.

[23] A. Ruhe, Closest normal matrix finally found!, BIT, 27(1987), 585-598.

[24] L. F. Shampine and M. K. Gordon, Computer Solution of Ordinary Differential Equations, The Initial Value Problem, Freeman and Company, San Francisco, 1975.

[25] G. W. Stewart, A Jacobi-like algorithm for computing the Schur decomposition of a non-hermitian matrix, SIAM J. Sci. Stat. Comp., 1985(6), 853-862.

[26] W. W. Symes, The QR algorithm and scattering for the finite nonperiodic Toda lattice, Phys. 4D(1982), 275-280.

[27] F. Uhlig, Simultaneous block diagonalization of two real symmetric matrices, Lin. Alg. Appli., 7(1973), 281-289.

[28] F. Uhlig, A canonical form for a pair of real symmetric matrices that generate a nonsingular pencil, Lin. Alg. Appli., 14(1976), 189-210.

[29] F. Uhlig, A recurring theorem about pairs of quadratic forms and extensions: A survey, Lin. Alg. Appli, 25(1979), 219-237.

[30] C. F. Van Loan, Generalizing the singular value decomposition, SIAM J. Num. Anal., 13(1976), 76-83.