Least Squares Approximation by Real Normal Matrices with Specified Spectrum

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 best approximating a given real matrix in the Frobenius norm by real, normal matrices subject to a prescribed spectrum is considered. The approach is based on using the projected gradient method. The projected gradient of the objective function on the manifold of constraints can be formulated explicitly. This gives rise to a descent flow that can be followed numerically. The explicit form also facilitates the computation of the second order optimality condition from which some interesting properties of the stationary points are related to the well-known Wielandt-Hoffman Theorem.

1 Introduction

A matrix $A \in C^{n \times n}$ is normal if and only if $A^*A = AA^*$. Normality, as it includes the Hermitian, unitary and skew-Hermitian matrices, defines a rather general and important class of matrices. In [7] seventy equivalent conditions are listed to characterize a normal matrix. This again reflects that normality may arise in many different ways.

One interesting question that has received considerable attention is the determination of a closest normal matrix to a given square complex matrix. This problem has only recently been completely solved (in the Frobenius norm) in [4], and independently in [12]. It turns out that finding a nearest normal matrix is equivalent to finding a unitary similarity transformation which makes the sum of squares of moduli of the diagonal elements as large as possible [8]. The Jacobi algorithm, therefore, may be derived from this perspective to solve the nearest to normality problem.

In this paper we assume the following situation happens: Experimental data has been collected in the matrix A which, by some prior knowledge, should be a normal matrix with known spectrum. Generally, due to measurement errors, A will not satisfy these requirements. Since A still contains some useful information, we would like to retrieve its least squares approximation that satisfies these requirements.

In practice, one may well be interested in real matrices. It is well known [5, p284] that a real normal matrix is always orthogonally similar to a real quasi-diagonal matrix

$$\operatorname{diag}\left\{ \begin{bmatrix} \lambda_1 & \nu_1 \\ -\nu_1 & \lambda_1 \end{bmatrix}, \ldots, \begin{bmatrix} \lambda_q & \nu_q \\ -\nu_q & \lambda_q \end{bmatrix}, \lambda_{2q+1}, \ldots, \lambda_n \right\} \tag{1}$$

where λ_k, ν_k are real numbers and $\nu_k \neq 0$ $(k = 1, 2, \ldots, q)$. Therefore, we consider the following problem in this paper:

Problem A Given a matrix $A \in R^{n \times n}$ and a set of eigenvalues $\{\lambda_1 \pm i\nu_1, \ldots, \lambda_q \pm i\nu_q, \lambda_{2q+1}, \ldots, \lambda_n\}$ where λ_k, ν_k are real numbers and $\nu_k \neq 0 (k = 1, 2, \ldots, q)$, find an orthogonal matrix Q that minimizes the function

$$F(Q) := \frac{1}{2} ||Q^T \Lambda Q - A||^2 \tag{2}$$

where Λ is the quasi-diagonal matrix given by (1) and $||\cdot||$ means the Frobenius matrix norm.

A special case of Problem A has been considered in [3]. There it is shown that when A is symmetric and when Λ is diagonal with distinct elements arranged in descending order, the columns of the optimal Q^T should be the normalized eigenvectors of A corresponding to eigenvalues arranged in the descending order. In this paper we study the extension to more general classes of matrices.

Our idea is closely related to the setting in [1]. Our approach is parallel to that in [3]. Without using the Lagrangian function, we first formulate explicitly the projection of the gradient of the objective function F onto the the feasible set $O(n) := \{Q \in R^{n\times n} | Q^TQ = I\}$. This formula gives rise to the construction of a descent flow that can be followed numerically. We then derive the so called projected Hessian on the tangent space of $O(n)$. Wherever possible, we classify the stationary points from the second-order condition. Finally we discuss the connection between our results and the well known Wielandt-Hoffman theorem [9].

2 Preliminaries

Let $\langle A, B\rangle$ denote the Frobenius inner product of two matrices $A, B \in R^{n\times n}$:

$$\langle A, B\rangle := \mathrm{trace}(AB^T) = \sum_{i,j} a_{ij}b_{ij}. \tag{3}$$

We first consider the function F in (2) to be defined everywhere in $R^{n\times n}$. For $Z, H \in R^{n\times n}$, the Fréchet derivative of F at Z acting on H is calculated to be

$$\begin{aligned} F'(Z)H &= \langle Z^T\Lambda Z - A, H^T\Lambda Z + Z^T\Lambda H\rangle \\ &= \langle (\Lambda Z)(Z^T\Lambda Z - A)^T, H\rangle + \langle (\Lambda^T Z)(Z^T\Lambda Z - A), H\rangle. \end{aligned} \tag{4}$$

In the second equation above we have used the adjoint property

$$< A, BC > = < B^T A, C > = < AC^T, B >$$

to rearrange terms. With respect to the Frobenius inner product, the equation (4) suggests that the gradient of F at a general matrix $Z \in R^{n\times n}$ may be interpreted as the matrix

$$\nabla F(Z) := (\Lambda Z)(Z^T\Lambda Z - A)^T + (\Lambda^T Z)(Z^T\Lambda Z - A). \tag{5}$$

Let $S(n)$ denote the subspace of all symmetric matrices in $R^{n\times n}$. It is easy to see that the tangent space $T_Q O(n)$ of the feasible set $O(n)$ is given by [3]

$$T_Q O(n) := QS(n)^\perp \tag{6}$$

where $S(n)^\perp$, the orthogonal complement of $S(n)$ in $R^{n\times n}$, is precisely the subspace of all skew-symmetric matrices. It is also easy to see that the orthogonal complement of $T_Q O(n)$ is the subspace

$$N_Q O(n) := QS(n). \tag{7}$$

Therefore, an orthogonal matrix Q is a stationary point of Problem A only if

$$(\Lambda Q)(Q^T\Lambda Q - A)^T + (\Lambda^T Q)(Q^T\Lambda Q - A) \in QS(n). \tag{8}$$

For convenience, we define in the sequel

$$X := Q^T \Lambda Q. \tag{9}$$

Then (8) is equivalent to

$$X(X^T - A^T) + X^T(X - A) \in S(n), \tag{10}$$

or

$$XA^T + X^T A = AX^T + A^T X. \tag{11}$$

Let $[A, B] := AB - BA$ denote the Lie bracket. It follows that

Lemma 2.1 *A necessary condition for $Q \in O(n)$ to be a stationary point for Problem A is that the matrix $[X, A^T]$ with X defined by (9) is symmetric.*

We remark that if A is symmetric and Λ is diagonal, then X is symmetric and $[X, A^T]$ is skew-symmetric. In this case, we conclude, from Lemma 2.1, that at a stationary point the matrix X must commute with A. This is one of the results discussed in [3].

The projected gradient of F on the manifold $O(n)$ can be calculated without any difficulty. Mainly this is due to the understanding that for any fixed $Q \in O(n)$,

$$R^{n\times n} = T_Q O(n) \oplus N_Q O(n) = QS(n)^{\perp} \oplus QS(n). \tag{12}$$

Any matrix $Z \in R^{n\times n}$ has a unique orthogonal splitting

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

as the sum of elements from $T_Q O(n)$ and $N_Q O(n)$. Accordingly, the projection $g(Q)$ of $\nabla F(Q)$ onto the tangent space $T_Q O(n)$ can be calculated explicitly as follows:

$$\begin{aligned} g(Q) &= \frac{Q}{2}\left\{Q^T \nabla F(Q) - \nabla F(Q)^T Q\right\} \\ &= \frac{Q}{2}\left\{Q^T\{(\Lambda Q)(Q^T\Lambda Q - A)^T + (\Lambda^T Q)(Q^T \Lambda Q - A)\}\right. \\ &\quad \left. -\{(\Lambda Q)(Q^T \Lambda Q - A)^T + (\Lambda^T Q)(Q^T\Lambda Q - A)\}^T Q\right\} \\ &= -\frac{Q}{2}\left\{[Q^T\Lambda Q, A^T] - [Q^T \Lambda Q, A^T]^T\right\}. \end{aligned} \tag{14}$$

It is clear that the vector field

$$\frac{dQ}{dt} := -g(Q) = \frac{Q}{2}\left\{[Q^T\Lambda Q, A^T] - [Q^T\Lambda Q, A^T]^T\right\} \tag{15}$$

defines a steepest descent flow $Q(t)$ on the manifold $O(n)$ for the objective function F in (2). Upon substitution, the corresponding $X(t)$ is governed by the ordinary differential equation:

$$\begin{aligned} \frac{dX}{dt} &:= \frac{dQ^T}{dt}\Lambda Q + Q^T\Lambda\frac{dQ}{dt} \\ &= \left[X, \frac{[X, A^T] - [X, A^T]^T}{2}\right]. \end{aligned} \tag{16}$$

Starting with an appropriate initial value, say $X(0) = \Lambda$, the positive orbit of (16) marches to a limit point which is a (local) least squares normal matrix approximation to A.

We remark again that if A is symmetric and Λ is diagonal, then the flow (16) is reduced to

$$\frac{dX}{dt} = [X, [X, A]] \tag{17}$$

which is analyzed in [3].

It is worth mentioning that the second term in the bracket of (16) is skew-symmetric. Therefore, the solution flow $X(t)$ of (16) naturally is isospectral [2] to the initial value $X(0)$. In particular, we have $||X(t)X(t)^T - X(t)^TX(t)|| = ||X(0)X(0)^T - X(0)^TX(0)||$ for all t. Thus, apart from numerical errors induced when solving the differential equation (16) on computers, the deviation of normality of $X(t)$ will remain the same as that of $X(0)$.

The function g in (14) is defined for orthogonal matrices only. We now derive an explicit formula for the projected Hessian of the objective function F without utilizing the Lagrangian Multiplier. Readers are referred to [3] for an explanation of why this technique works. Obviously we may extend g smoothly to cover the entire space $R^{n\times n}$ simply by defining

$$G(Z) := \frac{Z}{2}\left\{[Z^T\Lambda Z, A^T]^T - [Z^T\Lambda Z, A^T]\right\}. \tag{18}$$

The Fréchet derivative of G can easily be calculated. In particular, at any stationary point Q of Problem A and for every tangent vector QK where

$K \in S(n)^{\perp}$, it holds that

$$\begin{aligned} \langle G'(Q)QK, QK\rangle &= \langle \frac{[[X,K],A^T]^T - [[X,K],A^T]}{2}, K\rangle \\ &= -\langle [[X,K],A^T], K\rangle \\ &= \langle [X,K],[A,K]\rangle. \end{aligned} \tag{19}$$

It can be proved that formula (19) is precisely the evaluation of the projected Hessian of the Lagrangian function of Problem A [6, p80]. Thus a necessary condition (and a sufficient condition if the strict inequality holds) for a stationary point Q to be a local minimum is that

$$\langle [X,K],[A,K]\rangle \geq 0 \text{ for every } K \in S(n)^{\perp}. \tag{20}$$

3 Application I — Real Eigenvalues

We now apply the first order condition (11) and the second order condition (20) to classify the stationary points for Problem A. It will prove useful if we define

$$E := QAQ^T. \tag{21}$$

We observe that the first order condition (11) and the second order condition (20) are equivalent to

$$\Lambda E^T + \Lambda^T E = E\Lambda^T + E^T\Lambda \tag{22}$$

and

$$\langle [\Lambda, K], [E, K] \rangle \geq 0 \text{ for every } K \in S(n)^{\perp}. \tag{23}$$

respectively.

In this section we first consider the case when Λ has only real eigenvalues. It follows that the matrix $X = Q^T\Lambda Q$ must be symmetric for any $Q \in O(n)$. For any general matrix $A \in R^{n\times n}$, let

$$A_S := \frac{1}{2}(A + A^T) \tag{24}$$

and

$$A_K := \frac{1}{2}(A - A^T) \tag{25}$$

denote the symmetric and skew-symmetric parts of A, respectively. We observe that

$$||X - A||^2 = ||X - A_S||^2 + ||A_K||^2. \tag{26}$$

Since the second term in (26) is fixed once A is given, a least squares approximation to A amounts to a least square approximation to A_S. Therefore, it suffices to consider the case when A is symmetric.

Suppose A is symmetric. We shall arrange eigenvalues of A in the natural ordering

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n. \tag{27}$$

We further divide our discussions according to whether or not Λ has simple eigenvalues.

Case 1 (Λ has only distinct eigenvalues)

For clarity, we shall assume the diagonal elements of Λ are arranged in the descending order

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n. \tag{28}$$

The following theorem completely classifies all the stationary points.

Theorem 3.1 *Suppose A is symmetric and has eigenvalues arranged as in (27). Suppose Λ is diagonal and has elements arranged as in (28). Then the stationary points of Problem A are classified as follows:*

1. *An orthogonal matrix Q is a stationary point of F only if columns $q_1, \ldots, q_n$ of Q^T are orthonormal eigenvectors of A.*

2. *A stationary point Q is a minimizer (or, a maximizer) of F only if columns $q_1, \ldots, q_n$ of Q^T correspond with eigenvalues $\mu_1, \ldots, \mu_n$ (or, the reverse order),respectively. All other stationary points are saddle points.*

3. *Any least squares approximation X to A is of the form*

$$X = \lambda_1 q_1 q_1^T + \cdots + \lambda_n q_n q_n^T. \tag{29}$$

The least squares approximation X is unique if A itself has distinct eigenvalues.

4. *The minimal value of F is equal to $\frac{1}{2}\sum_{i=1}^{n}(\lambda_i - \mu_i)^2$.*

5. *Local extreme points are also global extreme points.*

(pf): The proof of this theorem can be found in [3]. The main point is that the simplicity of eigenvalues of Λ and the condition (22) require that E be a diagonal matrix [11, p416].

Case 2 (Λ has multiple eigenvalues)

When multiple eigenvalues occur, the analysis becomes more complicated because the matrix E is not necessarily a diagonal matrix. For demonstration purpose, we shall only consider the special case when all eigenvalues, except the one which has multiplicity 2, of Λ are simple.

We shall assume the diagonal elements of Λ are arranged in the ordering

$$\lambda_1 > \cdots > \lambda_k = \lambda_{k+1} > \cdots > \lambda_n \tag{30}$$

with $1 \leq k \leq n-1$. Then the first order condition (22) implies that at a stationary point E must be a quasi-diagonal matrix of the form [11]

$$E = \mathrm{diag}\left\{ e_1, \ldots, e_{k-1}, \begin{bmatrix} e_k & e_* \\ e_* & e_{k+1} \end{bmatrix}, e_{k+2}, \ldots, e_n \right\}. \tag{31}$$

It follows from (21) that $e_1, \ldots, e_{k-1}, e_{k+2}, \ldots, e_n$ must be $n-2$ eigenvalues of A (Note that we are assuming A be symmetric), and that columns $q_1, \ldots, q_{k-1}, q_{k+2}, \ldots, q_n$ of the matrix Q^T must be the corresponding orthonormal eigenvectors. Obviously, the 2×2 matrix

$$R := \begin{bmatrix} e_k & e_* \\ e_* & e_{k+1} \end{bmatrix}. \tag{32}$$

determines the remaining two eigenvalues, denoted by μ_s and μ_t, of A. The columns q_k and q_{k+1} are two orthonormal vectors in the space spanned by eigenvectors of μ_s and μ_t.

It is not difficult to see that

$$\begin{aligned} \langle [\Lambda, K], [E, K] \rangle = 2 &\sum_{\substack{i<j \\ i \neq k, k+1 \\ j \neq k, k+1}} (\lambda_i - \lambda_j)(e_i - e_j) k_{ij}^2 \\ +2 &\sum_{k+1<j} (\lambda_k - \lambda_j) \left\{ (e_k - e_j) k_{kj}^2 + 2 e_* k_{kj} k_{k+1,j} + (e_{k+1} - e_j) k_{k+1,j}^2 \right\} \\ +2 &\sum_{i<k} (\lambda_i - \lambda_k) \left\{ (e_i - e_k) k_{ik}^2 - 2 e_* k_{ik} k_{i,k+1} + (e_i - e_{k+1}) k_{i,k+1}^2 \right\}. \end{aligned} \tag{33}$$

We note that the three summations in (33) are mutually exclusive. Therefore, $\langle [\Lambda, K], [E, K] \rangle \geq 0$ for every $K \in S(n)^\perp$ if and only if every single term in (33) is nonnegative. Because of the specified ordering of the eigenvalues λ_i, we conclude that for a stationary point Q to be a local minimizer, it is necessary that

$$e_1 \geq e_2 \geq \cdots \geq e_{k-1} \geq e_{k+2} \geq \cdots \geq e_n, \tag{34}$$

and that the matrices

$$\begin{aligned}
\begin{bmatrix} e_i - e_k & -e_* \\ -e_* & e_i - e_{k+1} \end{bmatrix} &= e_i I - R, \text{ for every } i < k \\
\begin{bmatrix} e_k - e_j & e_* \\ e_* & e_{k+1} - e_j \end{bmatrix} &= R - e_j I, \text{ for every } k + 1 < j
\end{aligned} \qquad (35)$$

be positive semi-definite. From the above, we have proved that

Theorem 3.2 *Suppose A is symmetric and has eigenvalues arranged as in (27). Suppose Λ is diagonal and has elements arranged as in (30). Then the stationary points of Problem A are classified as follows:*

1. *An orthogonal matrix Q is a stationary point of F only if columns $q_1, \ldots, q_{k-1}, q_{k+2}, \ldots, q_n$ of the matrix Q^T are $n - 2$ orthonormal eigenvectors of A, and q_k, q_{k+1} are linear combinations of the remaining two orthonormal eigenvectors.*

2. *A stationary point Q is a local minimizer of F only if columns $q_1, \ldots, q_{k-1}$ of Q^T correspond with eigenvalues $\mu_1, \ldots, \mu_{k-1}$, and $q_{k+2}, \ldots, q_n$ correspond with eigenvalues $\mu_{k+2}, \ldots, \mu_n$, and q_k, q_{k+1} are linear combinations of eigenvectors corresponding with eigenvalues μ_k, μ_{k+1}. Similarly, a stationary point Q is a maximizer of F only if the above correspondence is in the reverse order. All other stationary points are saddle points.*

3. *Any least squares approximation X to A is of the form*

$$X = \lambda_1 q_1 q_1^T + \cdots + \lambda_k (q_k q_k^T + q_{k+1} q_{k+1}^T) + \cdots + \lambda_n q_n q_n^T. \qquad (36)$$

   The choice of q_k and q_{k+1} is immaterial. The least squares approximation is unique if the first $k - 1$ and the last $n - k - 1$ eigenvalues of A are distinct.

4. *The minimal value of F is equal to $\frac{1}{2}\sum_{i=1}^{n}(\lambda_i - \mu_i)^2$.*

5. *Local extreme points are also global extreme points.*

We remark that the proof for the above theorem can be generalized to cover other cases of multiple eigenvalues. The details are left to the readers.

4 Application II — Complex Eigenvalues

One of the difficulties associated with this case is that there is no clear way to order the eigenvalues. Even so, we have made some interesting observations .

Case 3 (A is a 2×2 matrix.)

The simple 2×2 case offers considerable insights into the understanding of higher dimensional problems. Let it be denoted

$$\Lambda = \begin{bmatrix} \lambda & \nu \\ -\nu & \lambda \end{bmatrix}. \tag{37}$$

For any $E \in R^{2\times 2}$, it is easy to see that the matrix $\Lambda E^T + \Lambda^T E$ is always symmetric. This is to say that any $Q \in O(2)$ is a stationary point. Indeed, we find that

$$X := Q^T \Lambda Q \equiv \begin{cases} \Lambda, & \text{if } \det Q = 1 \\ \Lambda^T, & \text{if } \det Q = -1. \end{cases} \tag{38}$$

So the least squares approximation problem is trivial. The objective function value is given by

$$F(Q) \equiv \frac{1}{2}\left((a_{11}-\lambda)^2 + (a_{22}-\lambda)^2 + (a_{12} \mp \nu)^2 + (a_{21} \pm \nu)^2\right) \tag{39}$$

depending upon $\det Q = \pm 1$, respectively. It is readily seen from (39) that the signs of ν and $a_{12} - a_{21}$ determine which one of Λ or Λ^T better approximates A.

Case 4 (A is a symmetric matrix.)

Again, for demonstration purpose, we shall consider only the case when Λ is of the form

$$\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \begin{bmatrix} \lambda_k & \nu_* \\ -\nu_* & \lambda_k \end{bmatrix}, \ldots, \lambda_n\} \tag{40}$$

where

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n. \tag{41}$$

and $\nu_* > 0$. Since A is symmetric, so is E. We write $\Lambda = \Lambda_S + \Lambda_K$ as the sum of its own symmetric and skew-symmetric parts. The first order condition (22) requires

$$(\Lambda^T + \Lambda)E = E(\Lambda^T + \Lambda). \tag{42}$$

Because $\Lambda^T + \Lambda = 2\Lambda_S$ is diagonal, it follows that E must be a quasi-diagonal matrix of the form (31). Furthermore, we know

$$\langle [\Lambda, K], [E, K] \rangle = \langle [\Lambda_S, K], [E, K] \rangle \tag{43}$$

since $[\Lambda_K, K]$ is skew symmetric and $[E, K]$ is symmetric. We state that

Theorem 4.1 *Suppose A is symmetric and has eigenvalues arranged as in (27). Suppose Λ is quasi-diagonal and has elements arranged as in (40) and (41). Then the stationary points of Problem A are classified as follows:*

1. *An orthogonal matrix Q is a stationary point of F only if columns $q_1, \ldots, q_{k-1}, q_{k+2}, \ldots, q_n$ of the matrix Q^T are $n-2$ orthonormal eigenvectors of A, and q_k, q_{k+1} are linear combinations of the remaining two orthonormal eigenvectors.*

2. *A stationary point Q is a local minimizer of F only if columns $q_1, \ldots, q_{k-1}$ of Q^T correspond with eigenvalues $\mu_1, \ldots, \mu_{k-1}$, and $q_{k+2}, \ldots, q_n$ correspond with eigenvalues $\mu_{k+2}, \ldots, \mu_n$, and q_k, q_{k+1} are linear combinations of eigenvectors corresponding with eigenvalues μ_k, μ_{k+1}. Similarly, a stationary point Q is a maximizer of F only if the above correspondence is in the reverse order. All other stationary points are saddle points.*

3. *Any least squares approximation X to A is of the form*

$$X = \lambda_1 q_1 q_1^T + \cdots + \lambda_k (q_k q_k^T + q_{k+1} q_{k+1}^T) + \nu_* (q_k q_{k+1}^T - q_{k+1} q_k^T) + \cdots + \lambda_n q_n q_n^T. \tag{44}$$

   The choice of q_k and q_{k+1} is immaterial. The least squares approximation is unique if the first $k-1$ and the last $n-k-1$ eigenvalues of A are distinct.

4. *The minimal value of F is equal to $\nu_*^2 + \frac{1}{2}\sum_{i=1}^n (\lambda_i - \mu_i)^2$.*

5. *Local extreme points are also global extreme points.*

(pf): The analysis of stationary points for this case is essentially identical to that of Case 2 in the preceding section.

Case 5 (A is a normal matrix.)

Obviously we should suppose A has complex eigenvalues, otherwise A would be symmetric. Now we have real difficulty in the analysis of the stationary points. In fact, we even do not have a clear way in identifying all stationary points. We can only report some partial results.

For simplicity, we shall assume that Λ is given by (40) and that (41) holds. We partition Λ into three blocks $\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$ where

$$\begin{aligned} \Lambda_1 &= \operatorname{diag}\{\lambda_1, \ldots, \lambda_{k-1}\} \\ \Lambda_2 &= \begin{bmatrix} \lambda_k & \nu_* \\ -\nu_* & \lambda_k \end{bmatrix} \qquad (45) \\ \Lambda_3 &= \operatorname{diag}\{\lambda_{k+2}, \ldots, \lambda_n\}. \qquad (46) \end{aligned}$$

It can be verified easily that any E of the form

$$E = E_1 \oplus E_2 \oplus E_3 \tag{47}$$

satisfies the first order condition (22) if $E_i + E_i^T$ is a diagonal matrix for $i = 1, 3$ and $E_2 \in R^{2\times 2}$. This, of course, is only a sufficient condition of being a stationary point.

We consider a simple 3×3 example. Let

$$A = \begin{bmatrix} -0.44910244205626 & -2.69770357656912 & -0.84185971635958 \\ 0.02746606843380 & -0.23010080980457 & -2.76631903691207 \\ -2.82587649838907 & -0.61291990656488 & -1.32079674813917 \end{bmatrix}$$

and

$$\Lambda = \begin{bmatrix} 15.0 & 0.0 & 0.0 \\ 0.0 & -3.0 & 12.0 \\ 0.0 & -12.0 & -3.0 \end{bmatrix}.$$

We calculate that $||AA^T - A^TA|| \approx 4.5540 \times 10^{-14}$. So up to the 14th digit A is a normal matrix whose eigenvalues are $\{1 \pm 2i, -4\}$. Starting with $X(0) = \Lambda$, we follow the descent flow (16) by using the subroutine ODE in [13]. The local error tolerancies set at 10^{-13}. We count convergence has occurred and stop the integration whenever the difference between two consecutive output values is less than 10^{-12}. At $t \approx 0.5$, we obtain an approximate limit point

$$X = \left[\begin{array}{rrr} 5.047565112549 & -12.481140871140 & -1.983297617463 \\ 1.946294703163 & 0.447719348364 & 12.759402874230 \\ -12.486964555620 & -3.288091746472 & 3.504715539087 \end{array} \right].$$

for the flow (16). The corresponding stationary point is approximated by

$$Q = \left[\begin{array}{rrr} 0.668645609196 & -0.437652789090 & -0.601143148929 \\ 0.437652789090 & 0.885212316658 & -0.157667975945 \\ 0.601143148929 & -0.157667975945 & 0.783433292538 \end{array} \right].$$

We calculate that $||XX^T - X^TX|| \approx 2.7084 \times 10^{-10}$, $||Q^TQ - I|| \approx 1.3866 \times 10^{-13}$. So X and Q are reasonably normal and orthogonal, respectively. The corresponding matrix $E := QAQ^T$ is given by

$$E = \left[\begin{array}{rrr} 0.644444444445 & -0.801988510684 & 2.173413906502 \\ 2.314685340881 & -0.608926976624 & -1.676627286676 \\ -0.095631338793 & -2.743293953342 & -2.035517467820 \end{array} \right].$$

We calculate that $||\Lambda E^T + \Lambda^T E - E\Lambda^T + E^T\Lambda|| \approx 1.2299 \times 10^{-11}$. So we may say that up to the numerical error the matrix E satisfies the equation (22). But obviously E is not of the form (47). We think this complication is due to the fact that the spectra of A and Λ are "incompatible", i.e., the two triangles in the complex plane connecting eigenvalues of A and Λ, respectively, point to opposite directions.

In perturbation theory, one should not expect the spectrum of Λ to be distributed in a significantly different pattern from that of A. In part, this is because eigenvalues depend continuously upon components of the matrix. In part, this is because A, representing a sensible empirical data, should more or less reflect the physical reality. Now that Λ is assumed to be of the form (40), let us suppose that A also has only one pair of complex conjugate eigenvalues. Thus A can be reduced to the matrix

$$E := \operatorname{diag}\left\{ e_1, \ldots, \left[\begin{array}{cc} e_k & e_* \\ -e_* & e_k \end{array} \right], \ldots, e_n \right\}. \tag{48}$$

Now we shall see how the ordering of $\{e_1, \ldots, e_n\}$ affect the definiteness of the projected Hessian of F at such a point. By direct computation, we obtain

$$\begin{aligned}\langle [\Lambda, K], [E, K] \rangle = 2 \sum_{\substack{i<j \\ i \neq k,k+1 \\ j \neq k,k+1}} (\lambda_i - \lambda_j)(e_i - e_j)k_{ij}^2 \\ +2 \sum_{i<k} (k_{ik}^2 + k_{i,k+1}^2)((e_i - e_k)(\lambda_i - \lambda_k) + v_* e_*) \\ +2 \sum_{k+1<j} (k_{kj}^2 + k_{k+1,j}^2)((e_k - e_j)(\lambda_k - \lambda_j) + v_* e_*). \qquad (49)\end{aligned}$$

Every single term in (49) needs to be nonnegative in order that the projected Hessian of F is positive semi-definite. This, of course, will be the case if the ordering of $\{e_1, \ldots, e_n\}$ is "compatible" with (41), that is, if

$$e_1 \geq e_2 \geq \ldots \geq e_n \qquad (50)$$

and $e_* > 0$. We, therefore, has established a result of the sufficient condition:

Lemma 4.1 *Suppose A is normal. Suppose A can be reduced by orthogonal transformation Q to the canonical form E (48) whose elements are arranged as in (50). Suppose Λ is a quasi-diagonal in the form of (40) whose elements are arranged as in (41). Then*

1. *The orthogonal matrix Q is a local minimizer of F.*

2. *The local optimal value of F is given by $\frac{1}{2}||\Lambda - E||^2$.*

Remark In the 3×3 numerical example above, we have $-4 = e_1 < e_2 = 1$. Thus (49) is positive only if $e_* > (e_2 - e_1)(\lambda_1 - \lambda_3)/\nu_* = 7.5$. Since $e_* = \pm 2$ in our example, we find that our descent flow X cannot converge to an E in the form of (47). In fact, it turns out that such an E is a local maximum for F.

In contrast to the preceding three theorems, it is rather surprising that when A has complex eigenvalues the differential equation (16) may have multiple limit points. This phenomenon can be observed numerically by starting with different initial values on the surface $M(\Lambda) := \{Q^T \Lambda Q | Q \in$

$O(n)\}$. For instance, if we start with $X(0) = \Lambda^T \in M(\Lambda)$ for the above 3×3 example, the flow converges to another limit point

$$X = \begin{bmatrix} 13.442778205310 & -0.124823985983 & -6.168244962433 \\ -5.831716696280 & -2.460547718025 & -10.728214876180 \\ -2.013431726775 & 12.210156961630 & -1.982230487286 \end{bmatrix}$$

which is quite different from the one obtained earlier. The least squares distances from these two distinct limit points to A, nevertheless, are the same. We have experimented with many other numerical examples. It seems true that when A is normal and has complex eigenvalues, Problem A does not have a unique solution. Different least squares approximations to A may result in different optimal values of F. Problem A, therefore, has multiple local solutions.

At this point, it is worthwhile to look at Problem A from another aspect. The following general perturbation problem [15] is of significant importance in many areas:

Problem B Suppose one knows exactly the eigenvalues of the matrix A and that A is perturbed to become $A + B$. How do the eigenvalues change?

Usually one is interested in finding bounds of the perturbed eigenvalues in terms of the perturbing matrix B. In application it is not uncommon to have a situation in which both the original matrix A and the perturbing matrix B are real and symmetric. In this case, and in the more general situation in which both A and $A + B$ are normal, a comprehensive bound, known as the Wielandt-Hoffman Theorem (See, [9], [10, p368] and [15, p104]), is available on the perturbation to all the eigenvalues.

Theorem 4.2 *Let $A, B \in C^{n \times n}$. Assume that A and $A + B$ are both normal. Let $\mu_1, \ldots, \mu_n$ be the eigenvalues of A in some given order, and let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $A + B$ in some order. Then there exists a permutation $\sigma(i)$ of the integers $1, 2, \ldots, n$ such that*

$$\sum_{i=1}^{n} |\lambda_{\sigma(i)} - \mu_i|^2 \leq ||B||^2. \tag{51}$$

In Problem A we have the situation that all the eigenvalues (the original ones and the perturbed ones) are known and that we want to minimize the norm of the perturbing matrix B.

What we have shown in Theorems 3.1 and 3.2 is that, in the real and symmetric case, the minimum of $||B||$ is attained if $A + B = Q^T \Lambda Q$ where columns of Q^T are orthogonal eigenvectors of A in a certain order. In this case, the equality in (51) holds. In other words, we have shown that the bound in (51) for eigenvalues is sharp. This is a reproof of the Wielandt-Hoffman theorem. We think our proof, being different from both the original proof of [9] and the one given in [15], is of interest in its own right.

When the matrix A is real and normal, one can see immediately that the proof given in [9] for Theorem 4.2 breaks down if the perturbed matrix $A + B$ is restricted to be only real and normal. Problem A in which we try to minimize the right-hand side of the inequality (51) becomes an interesting but difficult question. In Lemma 4.1 we have proved that if eigenvalues of A and $A + B$ (both real and normal) are "compatible", then again the equality in (51) holds. Our numerical experiments seem to indicate, however, that generally the minimal $||B||$ may be far larger than any rearrangement of eigenvalues on the left-hand side of the inequality (51) if only real matrices are allowed in the perturbation. Taking the 3 × 3 example to demonstrate our point, we calculate $||X - A||^2 \approx 496.2$ in comparison with the eigenvalue variation

$$\min_{\sigma} \sum_{i=1}^{n} |\lambda_{\sigma(i)} - \mu_i|^2 = 461.$$

Case 6 (A is a general matrix.)

Given a quasi-diagonal matrix Λ as in (1), an arbitrary matrix $A \in R^{n\times n}$ and let $X := Q^T \Lambda Q$, we have established that necessary conditions for $Q \in O(n)$ to be a local minimizer for Problem A are

$$\begin{aligned} XA^T + X^T A &= AX^T + A^T X; & (52) \\ \langle [X, K], [A, K] \rangle &\geq 0 \text{ for every } K \in S(n)^{\perp}. & (53) \end{aligned}$$

If the strict inequality holds in (53), then the above conditions are sufficient for $Q \in O(n)$ to be a strong local minimizer of Problem A.

Thus far, we are able to characterize an analytical solution of Problem A from (53) and (53) for the following cases:

1. All eigenvalues of Λ are real, and $A \in R^{n\times n}$ is arbitrary.

2. Λ has complex conjugate eigenvalues, and $A \in R^{n\times n}$ is symmetric.

3. Λ has complex conjugate eigenvalues, and $A \in R^{n \times n}$ is normal but not symmetric. (Indeed, only partial results are obtained for this case.)

For a general non-normal matrix A, the analytic comprehension of solutions satisfying both (53) and (53) becomes a much harder problem.

We have pointed out (Case 3) that when $n = 2$, all orthogonal matrices $Q \in O(2)$ are stationary points and the corresponding X can only be either Λ or Λ^T. From here, we might be able to characterize some stationary points for higher dimensional cases. For example, suppose Λ is given by (40). Suppose A can be reduced by orthogonal similarity to the matrix

$$E := \operatorname{diag}\left\{ e_1, \ldots, \begin{bmatrix} e_k^{(11)} & e_k^{(12)} \\ e_k^{(21)} & e_k^{(22)} \end{bmatrix}, \ldots, e_n \right\} \tag{54}$$

which is conformal with Λ except that $e_k^{(ij)}, 1 \leq i, j \leq 2$ are arbitrary real numbers. Then one can show that the equation (53) is satisfied. This, of course, is just one special type of stationary points.

Recently, the Wielandt-Hoffman Theorem has been generalized to nondefective matrices [14, 16]:

Theorem 4.3 *Let $A, B \in C^{n \times n}$. Suppose both A and $A+B$ are nondefective, i.e., suppose there exist nonsingular matrices S and T such that*

$$\begin{aligned} S^{-1}AS &= diag\{\mu_1, \ldots, \mu_n\} \\ T^{-1}(A+B)T &= diag\{\lambda_1, \ldots, \lambda_n\}. \end{aligned}$$

Then there exists a permutation $\sigma(i)$ of intergers $1, 2, \ldots, n$ such that

$$\sum_{i=1}^{n} |\lambda_{\sigma(i)} - \mu_i|^2 \leq (\kappa_2(S)\kappa_2(T))^2 ||B||^2 \tag{55}$$

where $\kappa_2(S) := ||S||_2||S^{-1}||_2$ is the condition number of S and $||\dot{}||_2$ means $2-$norm.

In the context of our discussion, the matrix $A+B$ is required to be a real and normal matrix. In this case, clearly $\kappa_2(T) = 1$. Suppose the given matrix A is nondefective, then the inequality (55) becomes

$$\sum_{n=1}^{n} |\lambda_{\sigma(i)} - \mu_i|^2 \leq \kappa_2(S)||B||^2. \tag{56}$$

The inequality (56 suggests that when A is a general non-normal matrix, the minimum value of $||X - A||$ may be smaller than the so called eigenvalue variation. That it indeed is the case can be seen from the 2×2 matrix considered in Case 3 — Supposee $a_{21} = 0, a_{12} > 0, \nu > 0$. Then it holds that

$$\min_{\sigma} \sum_{i=1}^{2} |\lambda_{\sigma(i)} - \mu_i|^2 = (a_{11} - \lambda)^2 + (a_{22} - \lambda)^2 + 2\nu^2 \tag{57}$$

while

$$\min_{Q \in O(2)} ||Q^T \Lambda Q - A||^2 = (a_{11} - \lambda)^2 + (a_{22} - \lambda)^2 + (a_{12} - \nu)^2 + \nu^2. \tag{58}$$

Obviously, the value in (58) is less than that in (57) if $a_{12} < 2\nu$. This observation is interesting when compared with the Wielandt-Hoffman Theorem for normal matrices. In the latter case, the minimum value of $||X - A||$ is always bounded *below* by the eigenvalue variation.

Although closed forms of solutions of (53) and (53) are difficult to obtain in general, our approach offers an alternative way to solve Problem A. We note that the differential equation (16), derived from the projected gradient of the objective function F, is numerically traceable for arbitrary matrix A. Thus, by following trajectories of (16), we may locate stationary solutions of the least squares problem numerically. Different starting points may lead to different stationary points. The asymptotic rate of convergence is expected to be similar to that of the usual steepest descent method. But the flow, by its definition, is guaranteed to converge regardless of the location of the starting point. Our numerical experience is that the flow usually reaches a stable equilibrium point within a reasonable interval of integration.

5 Acknowledgement

The author wishes to thank an anonymous referee for bringing the references [14] and [16] to his attention. The provision by the same referee of additional statistics that enlightens the numerical examples is also gratefully acknowledged.

References

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

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

[3] 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.

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

[5] F. R. Gantmacher, Matrix Theory, Vol. 1, Chelsea, New York, 1959.

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

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

[8] N. J. Higham, Matrix nearest problems and applications, Proceedings of the IMA Conference on Applications of Matrix Theory, S. Barnett and M. J. C. Gover, eds., Oxford University Press, 1989, to appear.

[9] A. J. Hoffman and H. Wielandt, The variation of the spectrum of a normal matrix, Duke Math. J., 20(1953), 37-39.

[10] R. A. Horn and C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1987.

[11] P. Lancaster and M. Tismenetsky, The Theory of Matrices, 2nd Ed., Academic, London, 1985.

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

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

[14] J. G. Sun, On the perturbation of the eigenvalues of a normal matrix, Math. Numer. Sinica, 6(1984), #3, 334-336.

[15] J. H. Wilkinson, The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.

[16] Z. Y. Zhang, On the perturbation of the eigenvalues of a nondefective matrix, Math. Numer. Sinica, 8(1986), #1, 106-108.