

The Convergence Rate of a Quasi-Newton Method for the Inverse Eigenvalue Problem

Raymond H. Chan* Shu-fang Xu† Hao-min Zhou‡

August 26, 2001

1 Introduction

Let $\{A_j\}_{j=1}^n$ be n real symmetric $n \times n$ matrices. For any vector $c = (c_1, c_2, \dots, c_n)^T$ in \mathbb{R}^n , we define

$$A(c) \equiv \sum_{j=1}^n c_j A_j, \quad (1)$$

and denote its eigenvalues by $\{\lambda_i(c)\}_{i=1}^n$ with $\lambda_1(c) \leq \dots \leq \lambda_n(c)$, and their corresponding normalized eigenvectors by $\{q_i(c)\}_{i=1}^n$. The inverse eigenvalue problem refers to the following one: Given n real numbers $\{\lambda_i^*\}_{i=1}^n$, which are ordered as $\lambda_1^* \leq \dots \leq \lambda_n^*$, find a vector $c^* \in \mathbb{R}^n$ such that $\lambda_i(c^*) = \lambda_i^*$ for $i = 1, \dots, n$. This problem can be posed as a problem of solving the nonlinear system

$$f(c) = 0, \quad (2)$$

*Department of Mathematics, Chinese University of Hong Kong, Hong Kong.

†Department of Mathematics, Peking University, Beijing, P. R. China.

‡Department of Mathematics, Chinese University of Hong Kong, Hong Kong.

where

$$f(c) = (\lambda_1(c) - \lambda_1^*, \dots, \lambda_n(c) - \lambda_n^*)^T. \quad (3)$$

We will assume that the given eigenvalues are distinct, i.e.

$$\lambda_1^* < \lambda_2^* < \dots < \lambda_n^*, \quad (4)$$

and that the Jacobian $J(c^*)$ of $f(c)$ at the true solution c^* is nonsingular. In Friedland, Nocedal and Overton [1], they have proposed to solve the nonlinear system (2) by Newton-type methods. The first one considered was the Newton method and the second one was a quasi-Newton method based on the inverse power method. It was proved that both methods converge quadratically.

In this paper, we first note that the proof of the quasi-Newton method as given in [1] is incorrect. We then give a correct proof of the convergence.

The outline of the paper is as follows. In §2, we give the algorithm of a quasi-Newton method based on the inverse power method for solving (2). The convergence rate of the method is given in §3 where we briefly explain why the proof of the convergence rate as given in Friedland, Nocedal and Overton [1] is incorrect. Some results on matrix perturbation theory and five preliminary lemmas that are useful in proving the convergence rate are given in §4 and §5 respectively. In §6, we prove the convergence rate by the mathematical induction. Finally concluding remarks are given in §7.

2 The Algorithms

Since the inverse eigenvalue problem is equivalent to the problem of solving the nonlinear system (2), we can use Newton-type methods to solve it. Before

we introduce the algorithm, we first remark that if two vectors b and c are close, then the matrices $A(b)$ and $A(c)$ they formed are also close.

Lemma 1 *For any vectors b and c in \mathbb{R}^n , we have*

$$\|A(b) - A(c)\|_F \leq \mu \|b - c\|, \quad (5)$$

where $\|\cdot\|_F$ and $\|\cdot\|$ denote the matrix Frobenius norm and vector 2-norm respectively, and

$$\mu \equiv \left(\sum_{j=1}^n \|A_j\|_F^2 \right)^{1/2}. \quad (6)$$

Proof: For any vectors $b = (b_1, \dots, b_n)^T$ and $c = (c_1, \dots, c_n)^T$, we have, by the Cauchy-Schwartz inequality,

$$\begin{aligned} \|A(b) - A(c)\|_F^2 &= \left\| \sum_{j=1}^n (b_j - c_j) A_j \right\|_F^2 \\ &= \sum_{i,l=1}^n \left(\sum_{j=1}^n (b_j - c_j) [A_j]_{i,l} \right)^2 \\ &\leq \sum_{i,l=1}^n \left(\sum_{j=1}^n (b_j - c_j)^2 \right) \left(\sum_{j=1}^n [A_j]_{i,l}^2 \right) \\ &\leq \|b - c\|^2 \sum_{i,l=1}^n \sum_{j=1}^n [A_j]_{i,l}^2 = \|b - c\|^2 \left(\sum_{j=1}^n \|A_j\|_F^2 \right). \quad \square \end{aligned}$$

We note that by using (4), (5) and results on matrix perturbation theory given in Wilkinson [5, pp.66–68], one can show that the eigenvalues and eigenvectors of $A(c)$ are differentiable functions with respect to c for c sufficiently close to the true solution c^* .

Lemma 2 (Sun [4, Theorem 2.3]) *Let $A(c) \in \mathbb{R}^{n \times n}$ be an analytic symmetric matrix-valued function defined on \mathbb{R}^n . For any given vector $c^* \in \mathbb{R}^n$, if $A(c^*)$ has n distinct eigenvalues, then there exist a scalar $\epsilon_0 > 0$, n analytic scalar functions $\{\lambda_i(c)\}_{i=1}^n$ and n analytic vector-valued functions $\{q_i(c)\}_{i=1}^n$, such that for all c with $\|c - c^*\| < \epsilon_0$, we have*

$$A(c)q_i(c) = \lambda_i(c)q_i(c), \quad i = 1, \dots, n \quad (7)$$

and

$$q_i(c)^T q_i(c) = 1, \quad i = 1, \dots, n. \quad (8)$$

According to (8), we have

$$\frac{\partial q_i(c)^T}{\partial c_j} q_i(c) = 0, \quad 1 \leq i, j \leq n. \quad (9)$$

Clearly, from (1), we have $\partial A(c)/\partial c_j = A_j$, for $j = 1, \dots, n$. Therefore, by (7) and (9), we have

$$\frac{\partial \lambda_i(c)}{\partial c_j} = q_i(c)^T \frac{\partial A(c)}{\partial c_j} q_i(c) = q_i(c)^T A_j q_i(c), \quad 1 \leq i, j \leq n.$$

Thus the Jacobian $J(c)$ of the function $f(c)$ defined in (3) is given by

$$[J(c)]_{i,j} = \left[\frac{\partial f(c)}{\partial c} \right]_{i,j} = q_i(c)^T A_j q_i(c), \quad 1 \leq i, j \leq n. \quad (10)$$

Using (1), it is easy to verify that

$$J(c) \cdot c = (\lambda_1(c), \dots, \lambda_n(c))^T. \quad (11)$$

Recall that the Newton method for $f(c) = 0$ is defined by

$$c^{k+1} = c^k - [J(c^k)]^{-1} f(c^k), \quad k = 1, 2, \dots$$

By (11) and (3), this becomes

$$J(c^k) \cdot c^{k+1} = (\lambda_1^*, \lambda_2^*, \dots, \lambda_n^*)^T, \quad k = 1, 2, \dots \quad (12)$$

Thus the algorithm of the Newton method for solving the inverse eigenvalue problem (2) is as follows:

Method I

Choose a starting vector c^1 . Then for $k = 1, 2, \dots$, do

- (i) Form $A(c^k)$ by (1).
- (ii) Compute all the eigenvalues $\lambda_i(c^k)$ and normalized eigenvectors $q_i(c^k)$ of $A(c^k)$.
- (iii) Choose the sign of all $q_i(c^k)$ to make $(q_i(c^{k-1}))^T q_i(c^k) \geq 0$ when $k > 1$.
- (iv) Stop if $\max_{i=1, \dots, n} |\lambda_i(c^k) - \lambda_i^*|$ is small enough. Otherwise, continue.
- (v) Form $J(c^k)$ by (10).
- (vi) Compute the next iterant c^{k+1} by solving (12).

We note that in step (ii), the exact eigenvalues $\{\lambda_i(c^k)\}_{i=1}^n$ and eigenvectors $\{q_i(c^k)\}_{i=1}^n$ of $A(c^k)$ are computed. That is very expensive. For a general matrix, computing a pair of eigenvalue and eigenvector requires $O(n^3)$ operations. There are now n pairs of eigenvalues and eigenvectors to be computed at each iteration. Thus the total cost per iteration of Method I is $O(n^4)$ operations.

One way to alleviate the cost is to approximate the eigenvalues and eigenvectors of $A(c^k)$ instead of computing them exactly. The following quasi-Newton method is based on using the inverse power method to find the approximate eigenvectors q_i^k to $q_i(c^k)$.

Method II

Choose a starting vector c^1 . Then form $A(c^1)$ by (1) and compute its exact eigenvalues λ_i^1 and the normalized eigenvectors q_i^1 , $1 \leq i \leq n$. Then for $k = 1, 2, \dots$, do

(i) Form $Q_k = [q_1^k, \dots, q_n^k]$, the matrix with the i th column given by q_i^k and $\Lambda^* = \text{diag}(\lambda_1^*, \dots, \lambda_n^*)$.

(ii) Stop if $\|Q_k^T A(c^k) Q_k - \Lambda^*\|_F$ is small enough. Otherwise, continue.

(iii) Form J_k (cf. (10)) where

$$[J_k]_{i,j} = (q_i^k)^T A_j q_i^k, \quad 1 \leq i, j \leq n. \quad (13)$$

(iv) Compute the next iterant c^{k+1} by solving (cf. (12))

$$J_k \cdot c^{k+1} = (\lambda_1^*, \lambda_2^*, \dots, \lambda_n^*)^T. \quad (14)$$

(v) Form $A(c^{k+1})$ by (1).

(vi) For each $i = 1, \dots, n$, solve v_i^k in

$$(A(c^{k+1}) - \lambda_i^* I) v_i^k = q_i^k. \quad (15)$$

Here I is the identity matrix.

(vii) Normalize v_i^k , $i = 1, \dots, n$,

$$u_i^{k+1} = \frac{v_i^k}{\|v_i^k\|}. \quad (16)$$

(viii) Compute the next approximate eigenvector by

$$q_i^{k+1} = \text{sign}((u_i^{k+1})^T q_i^k) u_i^{k+1}. \quad (17)$$

We note that the main cost per iteration of Method II is at step (vi) where n linear systems are solved. Usually, it saves much computational cost comparing with step (ii) of Method I where the eigenvalues and eigenvectors of $A(c^{k+1})$ are computed exactly and directly. Since the eigenvalues and eigenvectors of $A(c^1)$ are computed exactly, we see that the iterants c^2 generated by Methods I and II are the same.

Comparing our Methods I and II with the methods given in [1], our methods fix the signs of q_i^k and $q_i(c^k)$ per iteration, see Steps (iii) and (viii) in Methods I and II respectively. We remark that the new iterants in both methods actually do not depend on the signs of the eigenvectors and the approximate eigenvectors, because the Jacobians don't change as the signs of the eigenvectors and approximate eigenvectors are changed. Here we fix the signs just for convenience in the convergence proof that we will give in the subsequent sections.

3 The Convergence Rate

Both Methods I and II have been studied in many literatures. The quadratic convergence rate of Method I has been proven in [1].

Theorem 1 (Friedland, Nocedal and Overton [1]) *Suppose that the inverse eigenvalue problem (2) has a solution c^* and that the Jacobian matrix $J(c^*)$ is nonsingular. Then there exist scalars $\epsilon_1, \rho_1 > 0$ such that if $\|c^1 - c^*\| < \epsilon_1$, then the iterants c^k of Method I converge quadratically to c^* , i.e.*

$$\|c^{k+1} - c^*\| \leq \rho_1 \|c^k - c^*\|^2, \quad k = 1, 2, \dots$$

In the same paper, they have claimed that the convergence rate of Method II is also quadratic.

Theorem 2 *Suppose that the inverse eigenvalue problem (2) has a solution c^* and that the Jacobian matrix $J(c^*)$ is nonsingular. Then there exist scalars $\epsilon, \rho > 0$ such that if $\|c^1 - c^*\| < \epsilon$, then the iterants c^k of Method II converge quadratically to c^* , i.e.*

$$\|c^{k+1} - c^*\| \leq \rho \|c^k - c^*\|^2, \quad k = 1, 2, \dots$$

In [1], the theorem was proved like this: Let $Q = [q_1^k, \dots, q_n^k]$ and $P = [q_1(c^*), \dots, q_n(c^*)]$. Define X by

$$e^X = Q^T P. \quad (18)$$

Then they claimed that X is a skew-symmetric matrix. Hence by Corollary 3.1 [1], they got,

$$\|X\| \leq \sigma \|Q - P\|, \quad (19)$$

where σ is a constant independent of k . Since P is a matrix of eigenvectors of $A(c^*)$, then

$$e^X \Lambda^* e^{-X} = Q^T A(c^*) Q. \quad (20)$$

Take expansion of (20), then

$$\Lambda^* + X\Lambda^* - \Lambda^*X = Q^T A(c^*)Q + O(\|X\|^2). \quad (21)$$

Therefore, the diagonal equation of (21) is

$$\lambda_i^* = (q_i^k)^T A(c^*)q_i^k + O(\|X\|^2).$$

Comparing it with the iteration formula (14), it has

$$J_k(c^{k+1} - c^*) = O(\|X\|^2). \quad (22)$$

Then by the nonsingularity assumption on $J(c^*)$ and (19), the quadratic convergence follows.

We note that in this proof, X is assumed to be a skew-symmetric matrix and this is not true. Since the matrix Q in Method II is computed by one step inverse power method, it is not guaranteed to be orthogonal. Therefore $Q^T P$ in general is not an orthogonal matrix. Hence X in general is not a skew-symmetric matrix. That implies that the inverse of e^X may not be the transpose of e^X and (20) may be incorrect. Thus we cannot obtain the expansion (21) and (22). Moreover, Corollary 3.1 of [1] cannot be used to derive (19). In particular, we cannot use (22) and (19) to get the required quadratic convergence.

In the remaining of the paper, we will give a correct proof of this quadratic convergence. The idea of the proof is to use the mathematical induction to prove that if c^1 is sufficiently close to c^* , then the following two inequalities hold for $k = 1, 2, \dots$:

$$\|q_i^k - q_i(c^*)\| \leq \gamma \|c^k - c^*\|, \quad i = 1, 2, \dots, n, \quad (23)$$

and

$$\|c^{k+1} - c^*\| \leq \rho \|c^k - c^*\|^2. \quad (24)$$

Here γ and ρ are constants independent of k .

4 Perturbation Theory

Before we go on, we need some standard results in matrix perturbation theory. The following lemmas show that if a vector c is close to the true solution c^* , then the corresponding eigenvalues $\{\lambda_i(c)\}_{i=1}^n$ and eigenvectors $\{q_i(c)\}_{i=1}^n$ of $A(c)$ are also close to that of $A(c^*)$.

Lemma 3 *For any c in \mathbb{R}^n , we have*

$$|\lambda_i(c) - \lambda_i^*| \leq \mu \|c - c^*\|, \quad i = 1, \dots, n \quad (25)$$

where μ is given by (6).

Proof: By (5) and the Hoffman-Widlandt Theorem, see for instance Wilkinson [5, p.104], we have for all $i = 1, \dots, n$,

$$|\lambda_i(c) - \lambda_i^*|^2 \leq \sum_{j=1}^n |\lambda_j(c) - \lambda_j^*|^2 \leq \|A(c) - A(c^*)\|_F^2 \leq \mu^2 \|c - c^*\|^2. \quad \square$$

As a corollary, we can prove that the eigenvalues $\{\lambda_i(c)\}_{i=1}^n$ of $A(c)$ are distinct if c is close to c^* .

Corollary 1 *Let*

$$\nu = \min_{1 \leq i \neq j \leq n} \frac{|\lambda_i^* - \lambda_j^*|}{2} > 0. \quad (26)$$

Then if $\|c - c^*\| < \nu/\mu$, we have

$$|\lambda_i(c) - \lambda_j^*| \geq \nu > 0, \quad 1 \leq i \neq j \leq n, \quad (27)$$

and

$$|\lambda_i(c) - \lambda_j(c)| > 0, \quad 1 \leq i \neq j \leq n, \quad (28)$$

i.e., the eigenvalues of $A(c)$ are distinct. In particular, its normalized eigenvectors $\{q_i(c)\}_{i=1}^n$ form an orthonormal basis of \mathbb{R}^n .

Proof: For $1 \leq i \neq j \leq n$, we have by (25) and (26),

$$|\lambda_i(c) - \lambda_j^*| \geq |\lambda_i^* - \lambda_j^*| - |\lambda_i(c) - \lambda_i^*| \geq 2\nu - \mu\|c - c^*\| \geq \nu.$$

Also, by (25) and (26) again,

$$|\lambda_i(c) - \lambda_j(c)| \geq |\lambda_i^* - \lambda_j^*| - |\lambda_i(c) - \lambda_i^*| - |\lambda_j(c) - \lambda_j^*| \geq 2\nu - 2\mu\|c - c^*\| > 0. \quad \square$$

This corollary also shows that as c is close to c^* , the eigenvalue $\lambda_i(c)$ of $A(c)$ is well separated by a gap ν from $A(c^*)$'s eigenvalues other than λ_i^* . In this case, the following lemma proves that the eigenvector $q_i(c)$ of $A(c)$ is close to the eigenvector $q_i(c^*)$ of $A(c^*)$, (see B. N. Parlett, [3], page 14 and page 222).

Lemma 4 For any c in set $\{c : \|c - c^*\| < \nu/\mu\}$, there exist eigenvectors $\{q_i(c)\}_{i=1}^n$ of $A(c)$ satisfying:

$$\|q_i(c) - q_i(c^*)\| \leq \frac{\sqrt{2}\mu}{\nu}\|c - c^*\|, \quad i = 1, \dots, n, \quad (29)$$

where μ and ν are given by (6) and (26).

Proof: Let $q_i(c)$ be the unit eigenvector of $A(c)$ that satisfies $(q_i(c))^T q_i(c^*) \geq 0$. Let ϕ denote the angle between $q_i(c)$ and $q_i(c^*)$. Then we have $\cos \phi \geq 0$. Decompose $q_i(c)$ in the form $q_i(c) = q_i(c^*) \cos \phi + p \sin \phi$, where p is the unit vector orthogonal to $q_i(c^*)$ in the plane spanned by $q_i(c)$ and $q_i(c^*)$. Hence,

$$\begin{aligned} (A(c) - A(c^*))q_i(c) &= (\lambda_i(c)I - A(c^*))q_i(c) \\ &= (\lambda_i(c)I - A(c^*))q_i(c^*) \cos \phi + (\lambda_i(c)I - A(c^*))p \sin \phi \\ &= (\lambda_i(c) - \lambda_i^*)q_i(c^*) \cos \phi + (\lambda_i(c)I - A(c^*))p \sin \phi. \end{aligned}$$

Using the facts $\|q_i(c^*)\| = \|p\| = 1$ and $(q_i(c^*))^T (\lambda_i(c)I - A(c^*))p = 0$ (since $(q_i(c^*))^T p = 0$ and $A(c^*)q_i(c^*) = \lambda_i^* q_i(c^*)$), we have

$$\begin{aligned} \|(A(c) - A(c^*))q_i(c)\|^2 &= (\lambda_i(c) - \lambda_i^*)^2 \cos^2 \phi + \|(\lambda_i(c)I - A(c^*))p\|^2 \sin^2 \phi \\ &\geq \|(\lambda_i(c)I - A(c^*))p\|^2 \sin^2 \phi. \end{aligned} \quad (30)$$

As p is orthogonal to $q_i(c^*)$, we can express p by $p = \sum_{j \neq i} \eta_{ij} q_j(c^*)$. Then by (27) and the fact that $\sum_{j \neq i} \eta_{ij}^2 = 1$, we have

$$\begin{aligned} \|(\lambda_i(c)I - A(c^*))p\|^2 &= \left\| \sum_{j \neq i} (\lambda_i(c) - \lambda_j^*) \eta_{ij} q_j(c^*) \right\|^2 \\ &= \sum_{j \neq i} (\lambda_i(c) - \lambda_j^*)^2 \eta_{ij}^2 \geq \nu^2 \sum_{j \neq i} \eta_{ij}^2 = \nu^2. \end{aligned}$$

Hence by (30) and (5), we have

$$\sin^2 \phi \leq \frac{1}{\nu^2} \|(A(c) - A(c^*))q_i(c)\|^2 \leq \frac{\mu^2}{\nu^2} \|c - c^*\|^2.$$

Therefore, we have

$$\begin{aligned} \|q_i(c) - q_i(c^*)\|^2 &= 2(1 - (q_i(c))^T q_i(c^*)) = 2(1 - \cos \phi) \\ &\leq 2 \sin^2 \phi \leq \frac{2\mu^2}{\nu^2} \|c - c^*\|^2. \quad \square \end{aligned}$$

5 Preliminary Lemmas

In this section, we will estimate how close the approximate eigenvectors $\{q_i^k\}_{i=1}^n$ and $\{q_i^{k+1}\}_{i=1}^n$ obtained by Method II are to the eigenvectors $\{q_i(c^{k+1})\}_{i=1}^n$ of $A(c^{k+1})$. These estimates will be useful in establishing (23) and (24) for the case of $k + 1$. We begin however by showing that the error $\|c^k - c^*\|$ in each iteration is non-increasing.

Lemma 5 *Let c^1 be such that $\|c^1 - c^*\| < \epsilon$. If (23) and (24) are true for all positive integers less than k , then*

$$\|c^{k+1} - c^*\| \leq \|c^k - c^*\| \leq \dots \leq \|c^1 - c^*\|. \quad (31)$$

Proof: By using (24) and noting that $\rho\epsilon < 1$. \square

By (31), we see that $\|c^{k+1} - c^*\| \leq \|c^1 - c^*\| < \epsilon \leq \nu/\mu$. Thus by (28), the eigenvalues $\{\lambda_i(c^{k+1})\}_{i=1}^n$ of $A(c^{k+1})$ are distinct. In particular, the corresponding set of normalized eigenvectors $\{q_i(c^{k+1})\}_{i=1}^n$ forms an orthonormal basis of \mathbb{R}^n . Moreover, by (29), we see that for any k , we can always choose $q_i(c^{k+1})$ such that if $\|c^{k+1} - c^*\| \leq \epsilon$, then we have

$$\|q_i(c^{k+1}) - q_i(c^*)\| \leq \frac{\sqrt{2}\mu}{\nu} \|c^{k+1} - c^*\|. \quad (32)$$

Our next three lemmas are to estimate the coefficients of q_i^k and q_i^{k+1} when expressed under this basis.

Lemma 6 *Let c^1 be such that $\|c^1 - c^*\| < \epsilon$. Suppose that (23) and (24) are true for all positive integers less than k . If we write*

$$q_i^k = \sum_{j=1}^n \alpha_{ij} q_j(c^{k+1}), \quad i = 1, 2, \dots, n, \quad (33)$$

then

$$|\alpha_{ij}| \leq \frac{(\gamma\nu + \sqrt{2}\mu)}{\nu} \|c^k - c^*\|, \quad 1 \leq i \neq j \leq n, \quad (34)$$

and

$$\alpha_{ii} \geq \frac{1}{2}, \quad 1 \leq i \leq n. \quad (35)$$

Proof: Since $\{q_j(c^{k+1})\}_{j=1}^n$ are orthonormal, we have

$$\|q_i^k - q_i(c^{k+1})\| = \left\| \sum_{j=1}^n \alpha_{ij} q_j(c^{k+1}) - q_i(c^{k+1}) \right\| = \left(\sum_{j \neq i} \alpha_{ij}^2 + (1 - \alpha_{ii})^2 \right)^{\frac{1}{2}}.$$

As q_i^k is a unit vector, $\sum_{j=1}^n \alpha_{ij}^2 = 1$. Hence $\sum_{j \neq i} \alpha_{ij}^2 + (1 - \alpha_{ii})^2 = 2 - 2\alpha_{ii}$.

Therefore we have

$$\|q_i^k - q_i(c^{k+1})\| = (2 - 2\alpha_{ii})^{1/2}, \quad 1 \leq i \leq n.$$

However, according to (23), (31) and (32), we also have,

$$\begin{aligned} \|q_i^k - q_i(c^{k+1})\| &\leq \|q_i^k - q_i(c^*)\| + \|q_i(c^*) - q_i(c^{k+1})\| \\ &\leq \gamma \|c^k - c^*\| + \frac{\sqrt{2}\mu}{\nu} \|c^{k+1} - c^*\| \\ &\leq \frac{(\gamma\nu + \sqrt{2}\mu)}{\nu} \|c^k - c^*\|. \end{aligned}$$

Therefore, we have

$$\sum_{j \neq i} \alpha_{ij}^2 + (1 - \alpha_{ii})^2 \leq \frac{(\gamma\nu + \sqrt{2}\mu)^2}{\nu^2} \|c^k - c^*\|^2,$$

and

$$\alpha_{ii} \geq 1 - \frac{(\gamma\nu + \sqrt{2}\mu)^2}{2\nu^2} \|c^k - c^*\|^2, \quad 1 \leq i \leq n. \quad (36)$$

Since $\epsilon \leq (\gamma + \sqrt{2}\mu/\nu)^{-1}$, the lemma follows. \square

Lemma 7 *Let c^1 be such that $\|c^1 - c^*\| < \epsilon$. Suppose that (23) and (24) are true for all positive integers less than k . If we write*

$$q_i^{k+1} = \sum_{j=1}^n \beta_{ij} q_j(c^{k+1}), \quad i = 1, \dots, n, \quad (37)$$

then

$$|\beta_{ij}| \leq \frac{2|\alpha_{ij}|\mu}{\nu} \|c^{k+1} - c^*\| \leq \frac{2\mu}{\nu} \|c^{k+1} - c^*\|, \quad 1 \leq i \neq j \leq n, \quad (38)$$

and

$$1 - |\beta_{ii}| \leq \frac{2\mu^2}{\nu^2} \|c^{k+1} - c^*\|^2, \quad 1 \leq i \leq n. \quad (39)$$

Proof: By (15) and (33), we have for all $i = 1, \dots, n$,

$$v_i^k = (A(c^{k+1}) - \lambda_i^* I)^{-1} q_i^k = \sum_{j=1}^n \frac{\alpha_{ij}}{\lambda_j(c^{k+1}) - \lambda_i^*} q_j(c^{k+1}).$$

Using (16), (17) and let $\tau_i^{k+1} = \text{sign}((u_i^{k+1})^T q_i^k)$, then we have

$$q_i^{k+1} = \tau_i^{k+1} \frac{v_i^k}{\|v_i^k\|} = \left(\sum_{j=1}^n \frac{\alpha_{ij}^2}{(\lambda_j(c^{k+1}) - \lambda_i^*)^2} \right)^{-\frac{1}{2}} \sum_{j=1}^n \frac{\tau_i^{k+1} \alpha_{ij}}{\lambda_j(c^{k+1}) - \lambda_i^*} q_j(c^{k+1}).$$

Comparing this with (37), we see that

$$\beta_{ij} = \frac{\tau_i^{k+1} \alpha_{ij}}{\lambda_j(c^{k+1}) - \lambda_i^*} \left(\sum_{j=1}^n \frac{\alpha_{ij}^2}{(\lambda_j(c^{k+1}) - \lambda_i^*)^2} \right)^{-\frac{1}{2}}, \quad 1 \leq i, j \leq n. \quad (40)$$

Therefore

$$|\beta_{ij}| = \frac{|\alpha_{ij}(\lambda_i(c^{k+1}) - \lambda_i^*)|}{|\alpha_{ii}(\lambda_j(c^{k+1}) - \lambda_i^*)|} \left(1 + \sum_{j \neq i} \frac{\alpha_{ij}^2 (\lambda_i(c^{k+1}) - \lambda_i^*)^2}{\alpha_{ii}^2 (\lambda_j(c^{k+1}) - \lambda_i^*)^2} \right)^{-\frac{1}{2}}, \quad 1 \leq i, j \leq n. \quad (41)$$

By (25), (27) and (35), we see that

$$|\beta_{ij}| \leq \frac{|\alpha_{ij}(\lambda_i(c^{k+1}) - \lambda_i^*)|}{|\alpha_{ii}(\lambda_j(c^{k+1}) - \lambda_i^*)|} \leq \frac{2|\alpha_{ij}|\mu}{\nu} \|c^{k+1} - c^*\|, \quad 1 \leq i \neq j \leq n,$$

which is the first inequality in (38). The second inequality in (38) just follows by using the fact $|\alpha_{ij}| \leq 1$. From (41), we also have

$$|\beta_{ii}| = \left(1 + \sum_{j \neq i} \frac{\alpha_{ij}^2 (\lambda_i(c^{k+1}) - \lambda_i^*)^2}{\alpha_{ii}^2 (\lambda_j(c^{k+1}) - \lambda_i^*)^2} \right)^{-\frac{1}{2}}, \quad 1 \leq i \leq n. \quad (42)$$

Notice that the function $g(\zeta) \equiv \zeta/2 - 1 + (1 + \zeta)^{-1/2}$ has $g(0) = 0$ and $g'(\zeta) \geq 0$ for all $\zeta \geq 0$. Therefore we have

$$1 - (1 + \zeta)^{-1/2} \leq \frac{\zeta}{2}, \quad \forall \zeta \geq 0.$$

Applying this inequality to (42), we then have for all $i = 1, \dots, n$,

$$1 - |\beta_{ii}| \leq \frac{1}{2} \left(\sum_{j \neq i} \frac{\alpha_{ij}^2 (\lambda_i(c^{k+1}) - \lambda_i^*)^2}{\alpha_{ii}^2 (\lambda_j(c^{k+1}) - \lambda_i^*)^2} \right).$$

Using (25), (27) and (35), we see that

$$1 - |\beta_{ii}| \leq \frac{2\mu^2}{\nu^2} \|c^{k+1} - c^*\|^2 \sum_{j \neq i} \alpha_{ij}^2.$$

Since q_i^k is a unit vector, $\sum_{j \neq i} \alpha_{ij}^2 \leq 1$, see (33). Thus (39) follows. \square

In the last lemma of this section, we investigate the sign of β_{ii} .

Lemma 8 *Suppose β_{ij} are defined as in (37), then*

$$0 \leq \beta_{ii} \leq 1, \quad 1 \leq i \leq n.$$

Proof: From (42), we immediately have $\beta_{ii} \leq 1$. By (17), we have

$$(q_i^{k+1})^T q_i^k = \text{sign}((u_i^{k+1})^T q_i^k)(u_i^{k+1})^T q_i^k \geq 0. \quad (43)$$

By (33), (37) and the orthogonality of $q_i(c^{k+1})$, we also have

$$(q_i^{k+1})^T q_i^k = \sum_{j=1}^n \alpha_{ij} \beta_{ij} = \alpha_{ii} \beta_{ii} + \sum_{j \neq i} \alpha_{ij} \beta_{ij}.$$

From (34), (35), (38), (39) and the fact $\epsilon \leq \nu/(2\mu)$, we have

$$|\alpha_{ii} \beta_{ii}| \geq \frac{1}{4}, \quad i = 1, \dots, n, \quad (44)$$

and

$$\left| \sum_{j \neq i} \alpha_{ij} \beta_{ij} \right| \leq \frac{2n\mu(\gamma\nu + \sqrt{2}\mu)}{\nu^2} \|c^{k+1} - c^*\| \|c^k - c^*\|.$$

As $\epsilon < \min\{\nu(\gamma\nu\mu + \sqrt{2}\mu^2)^{-1}, \nu/(16n)\}$, therefore, we get

$$\left| \sum_{j \neq i} \alpha_{ij} \beta_{ij} \right| \leq \frac{1}{8}. \quad (45)$$

Since $\alpha_{ii} \beta_{ii} \geq 0$ by (43) and (35), we see that $\beta_{ii} \geq 0$. \square

6 The Mathematical Induction

In this section, we will prove that (23) and (24) are true for the case of $k + 1$.

We will also show that the Jacobian matrix J_{k+1} defined by Method II in (13) is nonsingular if $J(c^*)$ is nonsingular. In particular, c^{k+2} can be computed in (14). Firstly, we give the constants that have been used before.

The constants γ and ρ in (23) and (24) are well defined by

$$\gamma = \frac{(2 + \sqrt{2})\mu}{\nu},$$

and

$$\rho = \min \left\{ \rho_1, \frac{4n\mu(2 + \sqrt{2})}{\nu} \max_{1 \leq i \leq n} |\lambda_i^*| \|J^{-1}(c^*)\|_F \right\}.$$

Summary all the requirements on ϵ , see Lemma 2, Lemma 5, (36), (44), (45) and (48), also in order to make sure the first step of Method II has quadratic convergence, we take

$$\epsilon = \min \left\{ \epsilon_0, \epsilon_1, \rho^{-1}, \frac{\nu}{2\mu}, \frac{\nu}{16n}, \frac{\nu}{\mu}(\gamma\nu + \sqrt{2}\mu)^{-1}, (4\sqrt{n}\mu\gamma\|J(c^*)^{-1}\|_F)^{-1} \right\}.$$

We are now ready to prove that (23) and (24) are true for the case of $k + 1$.

Lemma 9 *Let c^1 be such that $\|c^1 - c^*\| < \epsilon$. Suppose that (23) and (24) are true for all positive integers less than k . Then (23) holds for $k + 1$, i.e.*

$$\|q_i^{k+1} - q_i(c^*)\| \leq \gamma\|c^{k+1} - c^*\|, \quad 1 \leq i \leq n. \quad (46)$$

Proof: Since $\{q_j(c^{k+1})\}_{j=1}^n$ are orthonormal, we have by (37),

$$\|q_i^{k+1} - q_i(c^{k+1})\| = \left\| \sum_{j=1}^n \beta_{ij} q_j(c^{k+1}) - q_i(c^{k+1}) \right\| = \left(\sum_{j \neq i} \beta_{ij}^2 + (1 - \beta_{ii})^2 \right)^{\frac{1}{2}},$$

for $i = 1, \dots, n$. As q_i^{k+1} is a unit vector, $\sum_{j=1}^n \beta_{ij}^2 = 1$. Hence

$$\sum_{j \neq i} \beta_{ij}^2 + (1 - \beta_{ii})^2 = 2 - 2\beta_{ii}.$$

Therefore we have

$$\|q_i^{k+1} - q_i(c^{k+1})\| = (2 - 2\beta_{ii})^{1/2}, \quad 1 \leq i \leq n.$$

Using this equality and (29), we then have, for all $i = 1, \dots, n$,

$$\begin{aligned} \|q_i^{k+1} - q_i(c^*)\| &\leq \|q_i^{k+1} - q_i(c^{k+1})\| + \|q_i(c^{k+1}) - q_i(c^*)\| \\ &\leq (2 - 2\beta_{ii})^{\frac{1}{2}} + \frac{\sqrt{2}\mu}{\nu} \|c^{k+1} - c^*\|. \end{aligned}$$

Because β_{ii} is positive, by substituting (39) in this inequality, (46) follows.

□

In the next lemma, we show that J_{k+1} is invertible and hence we can compute c^{k+2} in (14).

Lemma 10 *Suppose that $J(c^*)$ is nonsingular and that c^1 is such that $\|c^1 - c^*\| < \epsilon$. If (23) and (24) are true for all positive integers less than k , then J_{k+1} defined in (13) is nonsingular with*

$$\|J_{k+1}^{-1}\|_F \leq 2\|J(c^*)^{-1}\|_F. \quad (47)$$

In particular, we can solve for c^{k+2} in (14).

Proof: For all $1 \leq i \neq j \leq n$, by the definitions of $J(c^*)$ and J_{k+1} in (10) and (13) respectively, we have

$$\begin{aligned} |[J_{k+1}]_{i,j} - [J(c^*)]_{i,j}| &= |(q_i^{k+1})^T A_j q_i^{k+1} - q_i(c^*) A_j q_i(c^*)| \\ &= \left| \sum_{l,m=1}^n [A_j]_{l,m} ([q_i^{k+1}]_l [q_i^{k+1}]_m - [q_i(c^*)]_l [q_i(c^*)]_m) \right| \\ &\leq \sum_{l,m=1}^n |[A_j]_{l,m} [q_i^{k+1}]_l ([q_i^{k+1}]_m - [q_i(c^*)]_m)| \\ &\quad + \sum_{l,m=1}^n |[A_j]_{l,m} [q_i(c^*)]_m ([q_i^{k+1}]_l - [q_i(c^*)]_l)|. \end{aligned}$$

Using the Cauchy-Schwartz inequality and after some simplifications, we get

$$|[J_{k+1}]_{i,j} - [J(c^*)]_{i,j}| \leq \|A_j\|_F (\|q_i^{k+1}\| \|q_i^{k+1} - q_i(c^*)\| + \|q_i(c^*)\| \|q_i^{k+1} - q_i(c^*)\|).$$

Recall that $\{q_i^{k+1}\}_{i=1}^n$ and $\{q_i(c^*)\}_{i=1}^n$ are all unit vectors, we finally have

$$|[J_{k+1}]_{i,j} - [J(c^*)]_{i,j}| \leq 2\|A_j\|_F \|q_i^{k+1} - q_i(c^*)\|, \quad 1 \leq i \neq j \leq n.$$

Thus by (6) and (23),

$$\|J_{k+1} - J(c^*)\|_F^2 \leq \sum_{i,j=1}^n (4\|A_j\|_F^2 \|q_i^{k+1} - q_i(c^*)\|^2) \leq 4n\mu^2\gamma^2 \|c^{k+1} - c^*\|^2.$$

By (31) and the fact that $\|c^1 - c^*\| < \epsilon \leq (4\sqrt{n}\mu\gamma\|J(c^*)^{-1}\|_F)^{-1}$, we then have

$$\|J_{k+1} - J(c^*)\|_F^2 \leq \frac{4n\mu^2\gamma^2}{16n\mu^2\gamma^2\|J(c^*)^{-1}\|_F^2} \leq \frac{1}{4\|J(c^*)^{-1}\|_F^2}. \quad (48)$$

Write

$$J(c^*)^{-1}J_{k+1} = I - J(c^*)^{-1}(J(c^*) - J_{k+1}) \equiv I - E,$$

then by (48), we have

$$\|E\|_F \leq \|J(c^*)^{-1}\|_F \|J(c^*) - J_{k+1}\|_F < \frac{1}{2}.$$

Hence by applying Lemma 2.3.3 in Golub and van Loan [2], we see that $J(c^*)^{-1}J_{k+1}$ is nonsingular (which implies that J_{k+1} is nonsingular) and

$$\|J_{k+1}^{-1}J(c^*)\|_F = \|(I - E)^{-1}\|_F \leq \frac{1}{1 - \|E\|_F} < 2.$$

Thus, $\|J_{k+1}^{-1}\|_F = \|J_{k+1}^{-1}J(c^*)J(c^*)^{-1}\|_F \leq 2\|J(c^*)^{-1}\|_F$. \square

Finally, we prove that (24) is true for the case of $k + 1$.

Lemma 11 *Let c^1 be such that $\|c^1 - c^*\| < \epsilon$. Suppose that (23) and (24) are true for all positive integers less than k . Then (24) holds for $k + 1$, i.e.,*

$$\|c^{k+2} - c^*\| \leq \rho \|c^{k+1} - c^*\|^2.$$

Proof: Let $Q(c^*) \equiv [q_1(c^*), \dots, q_n(c^*)]$ and $Q_{k+1} \equiv [q_1^{k+1}, \dots, q_n^{k+1}]$. Clearly, by (4), $Q(c^*)$ is orthogonal. Define

$$I + V \equiv Q(c^*)^T Q_{k+1}. \quad (49)$$

By (46), we have

$$\begin{aligned}\|V\|_F &= \|Q(c^*)^T Q_{k+1} - I\|_F = \|Q_{k+1} - Q(c^*)\|_F \\ &= \left(\sum_{i=1}^n \|q_i^{k+1} - q_i(c^*)\|^2 \right)^{1/2} \leq \gamma \sqrt{n} \|c^{k+1} - c^*\|.\end{aligned}\quad (50)$$

Using (49), we can write

$$\begin{aligned}Q_{k+1}^T A(c^*) Q_{k+1} &= (Q_{k+1}^T Q(c^*)) (Q(c^*)^T A(c^*) Q(c^*)) (Q(c^*)^T Q_{k+1}) \\ &= (I + V)^T \Lambda^* (I + V) \\ &= \Lambda^* + \Lambda^* V + V^T \Lambda^* + V^T \Lambda^* V,\end{aligned}\quad (51)$$

where $\Lambda^* = \text{diag}(\lambda_1^*, \dots, \lambda_n^*)$.

From the definition of J_k in (13), we see that

$$[J_{k+1} \cdot c^*]_i = \sum_{j=1}^n (q_i^{k+1})^T A_j q_j^{k+1} \cdot c_j^* = (q_i^{k+1})^T A(c^*) q_i^{k+1}, \quad 1 \leq i \leq n,$$

i.e., the vector $J_{k+1} \cdot c^*$ gives the main diagonal of $Q_{k+1}^T A(c^*) Q_{k+1}$. Thus comparing the main diagonal entries of matrices on both side of (51), we get

$$J_{k+1} \cdot c^* = (\lambda_1^*, \lambda_2^*, \dots, \lambda_n^*)^T - w, \quad (52)$$

where $w = \text{diag}(\Lambda^* V + V^T \Lambda^* + V^T \Lambda^* V)$, i.e.

$$[w]_i = 2\lambda_i^* [V]_{i,i} + \sum_{j=1}^n \lambda_j^* [V]_{i,i}^2, \quad 1 \leq i \leq n. \quad (53)$$

Using (14) in (52), we then have

$$J_{k+1}(c^* - c^{k+2}) = w.$$

By (47), this becomes

$$\|c^{k+2} - c^*\| \leq \|J_{k+1}^{-1}\| \|w\| \leq 2\|J(c^*)^{-1}\|_F \|w\|. \quad (54)$$

It remains to estimate $\|w\|$. For this, we first note that by the definition of V in (49) and the fact that $Q(c^*)$ is orthogonal, we have

$$I + V + V^T + V^T V = (I + V)^T (I + V) = Q_{k+1}^T Q(c^*) Q(c^*)^T Q_{k+1} = Q_{k+1}^T Q_{k+1}.$$

Since $\{q_j^{k+1}\}_{j=1}^n$ are unit vector, we see that the main diagonal entries of $Q_{k+1}^T Q_{k+1}$ are 1. Hence we see that the main diagonal entries of $V + V^T + V^T V$ are zeros. Therefore, we get

$$[V]_{i,i} = -\frac{1}{2} \sum_{j=1}^n [V]_{j,i}^2, \quad 1 \leq i \leq n.$$

Putting this back into (53), we then have

$$\begin{aligned} \sum_{i=1}^n [w]_i^2 &\leq 2 \left\{ \sum_{i=1}^n (\lambda_i^*)^2 [V]_{ii}^2 + \sum_{i=1}^n \left(\sum_{j=1}^n \lambda_j^* [V]_{ji}^2 \right)^2 \right\} \\ &\leq 4 \max_{1 \leq i \leq n} |\lambda_i^*|^2 \sum_{i=1}^n \left(\sum_{j=1}^n [V]_{ji}^2 \right)^2 \\ &\leq 4 \max_{1 \leq i \leq n} |\lambda_i^*|^2 \left(\sum_{i=1}^n \sum_{j=1}^n [V]_{ji}^2 \right)^2 \\ &\leq 4 \max_{1 \leq i \leq n} |\lambda_i^*|^2 \|V\|_F^4. \end{aligned}$$

Thus by (50), we get

$$\|w\| \leq 2\gamma n \max |\lambda_i^*|^2 \|c^{k+1} - c^*\|^2.$$

The lemma now follows by putting this estimate back into (54). \square

Let us end the proof of the mathematical induction with the case $k = 1$.

Lemma 12 *Let c^1 be such that $\|c^1 - c^*\| < \epsilon$. Then (23) and (24) hold for $k = 1$.*

Proof: As have already remarked in §2, the second iterants c^2 for Methods I and II are the same, since the exact eigenvalues and eigenvectors are computed in the first iteration in both methods. Thus by Theorem 1, and the fact that $\|c^1 - c^*\| < \epsilon_1 < \epsilon$, we have

$$\|c^2 - c^*\| \leq \rho_1 \|c^1 - c^*\|^2 \leq \rho \|c^1 - c^*\|^2.$$

i.e. (24) holds. By (29), we also have, for all $i = 1, \dots, n$,

$$\|q_i^1 - q_i(c^*)\| = \|q_i(c^1) - q_i(c^*)\| \leq \frac{\sqrt{2}\mu}{\nu} \|c^1 - c^*\|. \quad \square$$

Proof of Theorem 2 : Lemma 12 shows that (23) and (24) hold for $k = 1$. Using Lemma 10, Lemma 11 and mathematical induction, the quadratic convergence follows. \square

7 Concluding Remarks

As we have mentioned previously, the iterants generated by our Methods I and II are the same as that generated by Methods I and II in [1]. We have proved that Method II is still quadratically convergent, although it is a quasi-Newton method. In Freidland, Nocedal and Overton [1], they presented numerical experiments which illustrate this quadratic convergence. In that paper, they also considered the case where multiple eigenvalues were given. They proposed several modified Newton-type methods, and the numerical examples show that they are methods with quadratic convergence.

References

- [1] S. Friedland, J. Nocedal and M. L. Overton, *The Formulation and Analysis of Numerical Methods for Inverse Eigenvalue Problems*, SIAM J. Numer. Anal., 24(1987), 634–667.
- [2] G. Golub and C. van Loan, *Matrix Computations*, 2nd Ed., The John Hopkins University Press, Baltimore, 1989.
- [3] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, N.J. 1980.
- [4] J. G. Sun, *Eigenvalues and Eigenvectors of a Matrix Dependent on Several Parameters*, J. Comput. Math., 3(1985), 351–364.
- [5] J. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.