Journal of Computational Mathematics, Vol.21, No.2, 2003, 201–206.

# A NOTE ON THE BACKWARD ERRORS FOR INVERSE EIGENVALUE PROBLEMS*1)

Xinguo Liu Zhengjian Bai
(*Department of Mathematics, Ocean University of China, Qingdao 266071*)

**Abstract**

In this note, we consider the backward errors for more general inverse eigenvalue problems by extending Sun's approach. The optimal backward errors are defined for diagonalization matrix inverse eigenvalue problem with respect to an approximate solution, and the upper and lower bounds are derived for the optimal backward errors. The results may be useful for testing the stability of practical algorithms.

*Key words*: Inverse eigenvalue problem; Optimal backward error; Upper bound; Lower bound

## 1. Introduction

Since the mid-1950's, inverse eigenvalue problems have been important subjects in numerical algebra and scientific and engineering computation. Many different kinds of inverse eigenvalue problems have arisen from various applications, including control theory, structural machanics, geology, molecular spectroscopy, and so on. The relevant conditions for the solvability, perturbation analysis, and numerical methods can be found in the literature (see, e.g., [3,4]).

In this note we use $\mathcal{C}^{n\times n}$ to denote the set of $n\times n$ complex matrices, the notation $A^H$ denotes the conjugate transpose of $A$, $\lambda(A)$ is the spectrum of $A$. $\|\cdot\|_F$ denotes Frobenius norm. The relation $A \simeq B$ means that the same order matrices A and B have the same eigenvalues. $\mathcal{D}_n$ stands for the set of diagonal matrices of order n (See [5] for other symbols).

An inverse eigenvalue problem, roughly speaking, is how to determine the elements of a matrix from its spectral data. One of the most important problems is the following inverse eigenvalue problem.

**Problem A.** Given $A_0, A_1, \cdots, A_n \in \mathcal{C}^{n\times n}$ and complex numbers $\lambda_1,\cdots,\lambda_n$, find $c = (c_1,\cdots,c_n)$ with complex components such that

$$A = A_0 + \sum_{k=1}^{n} c_k A_k \simeq \Lambda,$$

where $\Lambda = diag(\lambda_1,\cdots,\lambda_n)$.

If we take $A_k = e_k e_k^T (k = 1,\cdots,n)$, where $e_k$ denote the $k$th column of the identity matrix of order $n$, this problem is known as the additive inverse eigenvalue problem.

Let $\widehat{c} = (\widehat{c}_1,\cdots,\widehat{c}_n)$ be an approximate solution to Problem A. In general, there are many backward perturbations $\Delta A_0,\cdots,\Delta A_n \in \mathcal{C}^{n\times n}$, and $\Delta\Lambda \in \mathcal{D}_n$ such that

$$A_0 + \Delta A_0 + \sum_{k=1}^{n} \widehat{c}_k \left(A_k + \Delta A_k\right) \simeq \Lambda + \Delta\Lambda.$$

* Received May 19, 2000.
1)This work was supported by NNSF of China.

It may be well asked: How close is the nearest Problem A for which $\widehat{c}$ is the solution?

There are various approaches to define backward errors for measuring the distance between the original problem and the perturbed problems. We define the backward error $\eta^{(\theta)}(\widehat{c})$ by

$$\eta^{(\theta)}(\widehat{c}) = \min_{(\Delta A_0, \Delta A_1, \cdots, \Delta A_n, \Delta\Lambda)\in\mathcal{G}} \left\{ \|\Delta A_0\|_F^2 + \sum_{k=1}^{n} \theta_k^2 \|\Delta A_k\|_F^2 + \theta_{n+1}^2 \|\Delta\Lambda\|_F^2 \right\}^{1/2},$$

where the set $\mathcal{G}$ is defined by

$$\mathcal{G} = \left\{ \begin{array}{l} (\Delta A_0, \Delta A_1, \cdots, \Delta A_n, \Delta\Lambda) | \Delta A_0, \cdots, \Delta A_n \in \mathcal{C}^{n\times n}, \Delta\Lambda \in \mathcal{D}_n \\ A_0 + \Delta A_0 + \sum\limits_{k=1}^{n} \widehat{c}_k (A_k + \Delta A_k) \simeq \Lambda + \Delta\Lambda \\ \Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n) \end{array} \right\},$$

and $\theta = (\theta_1, \cdots, \theta_{n+1})$, in which $\theta_1, \cdots, \theta_{n+1}$ are positive parameters.

**Problem B.** Estimate $\eta^{(\theta)}(\widehat{c})$.

For the following case, the explicit expression of $\eta^{(\theta)}(\widehat{c})$ is derived by Sun[2]. $\lambda_1, \cdots, \lambda_n$ are real numbers, $A_0, A_1, \cdots, A_n$ are real symmetric matrices, $\Delta A_k (k = 0, 1, \cdots, n)$ are real symmetric matrices, $\Delta\Lambda$ is real digonnal matrix, and $\widehat{c} = (\widehat{c}_1, \cdots, \widehat{c}_n)$ is an approximate solution with real components to Problem A in such a case. One of main tools which were used by Sun is the Mirsky inequality and the spectral decomposition theorem. However, for more general matrices, these results are not true any more. Hence the purpose of this note is to extend Sun's method to consider more general Problem B. In general, the eigenvalues of real matrices are complex numbers, and the solvability of Problem A has been discussed[7]. So, in this note, our discussion is not limited to the real case.

## 2. Main Results

We first prove two preliminary lemmas.

**Lemma 1**[1]**.** Let $A, B \in \mathcal{C}^{n\times n}$, and $A$ be a normal matrix, $\lambda(A) = \{\lambda_j\}_{j=1}^n, \lambda(B) = \{\mu_j\}_{j=1}^n$. Then, there exists a permutation $\pi(1), \cdots, \pi(n)$ of $\{1, 2, \cdots, n\}$ such that

$$\sqrt{\sum_{j=1}^{n} \left|\lambda_j - \mu_{\pi(j)}\right|^2} \le \sqrt{n}\, \|A - B\|_F .$$

**Lemma 2.** Let $A, B \in \mathcal{C}^{n\times n}, A$ be a diagonalizable matrix. i.e., $A = U\Lambda U^{-1}$, where U is nonsingular, $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n), \lambda(B) = \{\mu_j\}_{j=1}^n$. Then, there exists a permutation $\pi(1), \cdots, \pi(n)$ of $\{1, 2, \cdots, n\}$ such that

$$\sqrt{\sum_{j=1}^{n} \left|\lambda_j - \mu_{\pi(j)}\right|^2} \le \kappa_2(U)\sqrt{n}\, \|A - B\|_F ,$$

where $\kappa_2(U) \equiv \|U\|_2 \|U^{-1}\|_2$.

*Proof.* By

$$U^{-1}(A - B)U = \Lambda - U^{-1}BU,$$

where $\Lambda$ is a normal matrix, and

$$\lambda(U^{-1}BU) = \lambda(B).$$

By Lemma 1, we obtain that there exists a permutation $\pi(1), \cdots, \pi(n)$ of $\{1, 2, \cdots, n\}$ such that

$$\begin{aligned}\sqrt{\sum_{j=1}^{n} \left|\lambda_j - \mu_{\pi(j)}\right|^2} &\leq \sqrt{n} \left\|U^{-1}(A-B)U\right\|_F \\ &\leq \sqrt{n}\|U\|_2\|U^{-1}\|_2\|A-B\|_F \\ &= \kappa_2(U)\sqrt{n}\,\|A-B\|_F .\end{aligned}$$

Now we consider Problem B.

Let $\widetilde{A} = A_0 + \sum_{k=1}^{n} \widehat{c}_k A_k$, and assume that $\widetilde{A}$ is diagonalizable, i.e.

$$\widetilde{A} = X\widetilde{\Lambda}X^{-1}, \widetilde{\Lambda} = \mathrm{diag}(\widetilde{\lambda}_1, \cdots, \widetilde{\lambda}_n),$$

where $X$ is a nonsingular matrix.

Let

$\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n), \kappa = \sqrt{n}\kappa_2(X), \kappa_2(X) = \|X\|_2\|X^{-1}\|_2,$
$\Theta = \mathrm{diag}(\theta_1, \cdots, \theta_{n+1}), \mathcal{P}^{n\times n} = \{P|P$ is a permutation matrix of order $n\},$
$g = (|\widehat{c}_1|, \cdots, |\widehat{c}_n|, \frac{1}{\kappa})^T, g_0 = (|\widehat{c}_1|, \cdots, |\widehat{c}_n|, 1)^T, \rho_0 = \min_{P\in\mathcal{P}^{n\times n}} \|P\Lambda - \widetilde{\Lambda}\|_F, \rho = \frac{\rho_0}{\kappa}.$

**Theorem.**

$$\frac{\rho}{\sqrt{1+g^T\Theta^{-2}g}} \leq \eta^{(\theta)}(\widehat{c}) \leq \frac{\kappa_2(X)\rho_0}{1+g_0^T\Theta^{-2}g_0}.$$

*Proof.* Define the sets $\pounds$ and $\pounds_\rho$ by

$$\pounds \equiv \left\{(\mathit{\Delta} A_1, \cdots, \mathit{\Delta} A_n, \mathit{\Delta}\mathit{\Lambda})|\mathit{\Delta} A_1, \cdots, \mathit{\Delta} A_n \in \mathcal{C}^{n\times n}, \mathit{\Delta}\mathit{\Lambda} \in \mathcal{D}_n\right\},$$

$$\pounds_\rho \equiv \left\{(\mathit{\Delta} A_1, \cdots, \mathit{\Delta} A_n, \mathit{\Delta}\mathit{\Lambda}) \in \pounds| \sum_{k=1}^{n} |\widehat{c}_k| \, \|\mathit{\Delta} A_k\|_F + \frac{1}{\kappa}\|\mathit{\Delta}\mathit{\Lambda}\|_F \leq \rho\right\},$$

and for each fixed $\Delta\Lambda \in \mathcal{D}_n$, define the set $\mathcal{S}_{\Lambda+\Delta\Lambda}$ by

$$\mathcal{S}_{\Lambda+\Delta\Lambda} = \left\{M|M \in \mathcal{C}^{n\times n}, \lambda(M) = \lambda(\Lambda+\Delta\Lambda)\right\}.$$

Then by Lemma 2, we have

$$\begin{aligned}&\min_{(\Delta A_1,\cdots,\Delta A_n,\Delta\Lambda)\in\pounds_\rho} \left\{\sum_{k=1}^{n} \theta_k^2 \|\Delta A_k\|_F^2 + \theta_{n+1}^2 \|\Delta\Lambda\|_F^2 + \min_{M\in\mathcal{S}_{\Lambda+\Delta\Lambda}} \left\|M - (\widetilde{A} + \sum_{k=1}^{n} \widehat{c}_k\Delta A_k)\right\|_F^2\right\} \\ &\geq \min_{(\Delta A_1,\cdots,\Delta A_n,\Delta\Lambda)\in\pounds_\rho} \left\{\sum_{k=1}^{n} \theta_k^2 \|\Delta A_k\|_F^2 + \theta_{n+1}^2 \|\Delta\Lambda\|_F^2 + \left[\rho - (\sum_{k=1}^{n} |\widehat{c}_k| \, \|\Delta A_k\|_F + \tfrac{1}{\kappa}\|\Delta\Lambda\|_F)\right]^2\right\} \\ &= \min_{\substack{\delta_1,\cdots,\delta_{n+1}\geq 0 \\ g^T d \leq \rho}} f(d),\end{aligned}$$

where

$$f(d) = d^T\Theta^2 d + \left(\rho - d^T g\right)^2, d = (\delta_1, \cdots, \delta_{n+1})^T,$$
$$\delta_k = \|\Delta A_k\|_F \quad (k = 1, \cdots, n), \delta_{n+1} = \|\Delta\Lambda\|_F .$$

The gradient $\nabla f$ of $f$ has the expression

$$\nabla f = 2\left(\left(\Theta^2 + gg^T\right)d - g\rho\right).$$

Therefore, a vector $d^*$ satisfies $\nabla f(d^*) = 0$ if and only if

$$d^* = \left(\Theta^2 + gg^T\right)^{-1} g\rho = \frac{\rho\Theta^{-2}g}{1+g^T\Theta^{-2}g},$$

which satisfies

$$g^T d^* \le \rho.$$

The Hessian matrix $H$ of $f$ has the expression

$$H = 2\left(\Theta^2 + gg^T\right),$$

which shows that $H$ is symmetric positive definite. Hence the function $f(d)$ has a unique global minimal point $d^*$, and

$$\min_{\substack{\delta_1,\cdots,\delta_{n+1}\ge 0\\ g^T d\le\rho}} f(d) = f(d^*) = \frac{\rho^2}{1+g^T\Theta^{-2}g}.$$

On the other hand, if $(\Delta A_1,\cdots,\Delta A_n,\Delta\Lambda)\in \pounds\backslash\pounds_\rho$, then

$$\sum_{k=1}^{n} |\widehat{c}_k|\, \|\Delta A_k\|_F + \frac{1}{\kappa}\|\Delta\Lambda\|_F > \rho,$$

i.e.

$$\rho < g^T d.$$

Therefore

$$\rho^2 < \left(d^T g\right)^2 \le \left(d^T\Theta^2 d\right)\left(g^T\Theta^{-2}g\right) \le \left(d^T\Theta^2 d\right)\left(1+g^T\Theta^{-2}g\right),$$

i.e.

$$d^T\Theta^2 d > \frac{\rho^2}{1+g^T\Theta^{-2}g}.$$

From the discussion, we obtain

$$\begin{aligned}
\left[\eta^{(\theta)}(\widehat{c})\right]^2 &= \inf_{(\Delta A_0,\cdots,\Delta A_n,\Delta\Lambda)\in\mathcal{G}} \left\{\|\Delta A_0\|_F^2 + \sum_{k=1}^{n}\theta_k^2\|\Delta A_k\|_F^2 + \theta_{n+1}^2\|\Delta\Lambda\|_F^2\right\}\\
&= \inf_{(\Delta A_0,\cdots,\Delta A_n,\Delta\Lambda)\in\pounds} \left\{d^T\Theta^2 d + \min_{M\in\mathcal{S}_{\Lambda+\Delta\Lambda}}\left\|M - \left(\widetilde{A}+\sum_{k=1}^{n}\widehat{c}_k\Delta A_k\right)\right\|_F^2\right\}\\
&= \min\left\{\min_{(\Delta A_0,\cdots,\Delta A_n,\Delta\Lambda)\in\pounds_\rho}\left[d^T\Theta^2 d + \min_{M\in\mathcal{S}_{\Lambda+\Delta\Lambda}}\left\|M-(\widetilde{A}+\sum_{k=1}^{n}\widehat{c}_k\Delta A_k)\right\|_F^2\right],\right.\\
&\qquad \left.\inf_{(\Delta A_0,\cdots,\Delta A_n,\Delta\Lambda)\in\pounds\backslash\pounds_\rho}\left[d^T\Theta^2 d + \min_{M\in\mathcal{S}_{\Lambda+\Delta\Lambda}}\left\|M-(\widetilde{A}+\sum_{k=1}^{n}\widehat{c}_k\Delta A_k)\right\|_F^2\right]\right\}\\
&\ge \frac{\rho^2}{1+g^T\Theta^{-2}g}.
\end{aligned}$$

The first inequality is proved.
For deriving the other inequality, let

$$\widehat{c}_k = |\widehat{c}_k|\,\bar{\sigma_k}(k=1,\cdots,n),$$

where $\bar{\sigma_k}$ denotes the conjugate of the complex number $\sigma_k$.
Let $P_\pi$ be a permutation matrix of order $n$ such that

$$\|P_\pi\Lambda - \widetilde{\Lambda}\|_F = \rho_0.$$

Take the specific permutations $\Delta A_k^*(k=1,\cdots,n)$, and $\Delta\Lambda^*$ defined by

$$\Delta A_k^* = X\Delta\Lambda_k^* X^{-1}, \Delta\Lambda_k^* = \frac{|\widehat{c}_k|\,\sigma_k}{\left(1+g_0^T\Theta^{-2}g_0\right)\theta_k^2}\left(P_\pi\Lambda - \widetilde{\Lambda}\right), k=1,\cdots,n,$$

and

$$\Delta\Lambda^* = -\frac{1}{(1+g_0^T\Theta^{-2}g_0)\theta_{n+1}^2}\left(P_\pi\Lambda - \widetilde{\Lambda}\right),$$

and

$$\Delta A_0^* = X\left(P_\pi\Lambda + \Delta\Lambda^*\right)X^{-1} - \left(\widetilde{A} + \sum_{k=1}^n \widehat{c}_k \Delta A_k^*\right).$$

We have

$$\lambda\left(A_0 + \Delta A_0^* + \sum_{k=1}^n \widehat{c}_k\left(A_k + \Delta A_k^*\right)\right) = \lambda\left(\Lambda + P_\pi^T\Delta\Lambda^*\right).$$

Nevertheless

$$\begin{aligned}
\left[\eta^{(\theta)}(\widehat{c})\right]^2 &\le \|\Delta A_0^*\|_F^2 + \sum_{k=1}^n \theta_k^2\|\Delta A_k^*\|_F^2 + \theta_{n+1}^2\|\Delta\Lambda^*\|_F^2 \\
&= \kappa_2^2(X)\left\{\left\|P_\pi\Lambda + \Delta\Lambda^* - \widetilde{\Lambda} - \sum_{k=1}^n |\widehat{c}_k|^2 \frac{P_\pi\Lambda - \widetilde{\Lambda}}{\left(1+g_0^T\Theta^{-2}g_0\right)\theta_k^2}\right\|_F^2\right. \\
&\quad \left. + \sum_{k=1}^n \frac{|\widehat{c}_k|^2}{(1+g_0^T\Theta^{-2}g_0)^2\theta_k^2}\left\|P_\pi\Lambda - \widetilde{\Lambda}\right\|_F^2 + \frac{\left\|P_\pi\Lambda - \widetilde{\Lambda}\right\|_F^2}{\left(1+g_0^T\Theta^{-2}g_0\right)^2\theta_{n+1}^2}\right\} \\
&= \frac{\kappa_2^2(X)\rho_0^2}{1+g_0^T\Theta^{-2}g_0}.\Box
\end{aligned}$$

**Remark.** The parameters $\theta_1,\cdots,\theta_{n+1}$ allow us some flexibility. For instance, take $\theta_k = \frac{\|A_0\|_F}{\|A_k\|_F} \equiv \theta_k^*$, for $k=1,\cdots,n$, and $\theta_{n+1} = \frac{\|A_0\|_F}{\|\Lambda\|_F} \equiv \theta_{n+1}^*$,
then, we have

$$\begin{aligned}
\frac{1}{\|A_0\|}\eta^{(\theta^*)}(\widehat{c}) &= \eta(\widehat{c}) \\
&= \min_{(\Delta A_0,\Delta A_1,\cdots,\Delta A_n,\Delta\Lambda)\in\mathcal{G}}\left[\left(\frac{\|\Delta A_0\|_F}{\|A_0\|_F}\right)^2 + \sum_{k=1}^n\left(\frac{\|\Delta A_k\|_F}{\|A_k\|_F}\right)^2 + \left(\frac{\|\Delta\Lambda\|_F}{\|\Lambda\|_F}\right)^2\right]^{\frac{1}{2}},
\end{aligned}$$

where $\theta^* = \left(\theta_1^*,\cdots,\theta_{n+1}^*\right)$.

## 3. Remarks

**Remark 1.** In this note, we assume $\widetilde{A}$ to be diagonalizable. It was proved in [6] that if $A$ has a multiple eigenvalue, the related inverse eigenvalue problems (in theory) have no solutions usually. However, when $\lambda_1,\cdots,\lambda_n$ are distinct, and $\rho$ is enough small, so that $\widetilde{\lambda}_1,\cdots,\widetilde{\lambda}_n$ are distict too, then $\widetilde{A}$ must be a diagonalizable matrix (see [6]).

**Remark 2.** The symmetric problem was considered by Sun. In this note we allow $A_k, \Delta A_k(k=1,\cdots,n), A_0, \Delta A_0$ to be complex matrices, $\lambda_1,\cdots,\lambda_n$ be complex numbers. On the one hand, the eigenvalues of a real matrix may be complex numbers. On the other hand, it was proved in [7] that in such a case the related inverse eigenvalue problems are solvable almost everywhere.

## References


[1] J. G. Sun, On the variation of the spectrum of a normal matrix, *Lin. Alg. Appl.*, **246**(1996), 215-223.

[2] J. G. Sun, Backward errors for the inverse eigenvalue problems, *Numer. Math.*, **82**(1999), 339-349.

[3] Shu-Fang Xu, An Introduction to inverse algebraic eigenvalue problems, Peking University Press, 1998.

[4] S. Q. Zhou, H. Dai, The algebraic inverse eigenvalue problems (in Chinese), Henan Science and Technology Press, Zhengzhou, China, 1991.

[5] J. G. Sun, The Perturbation Analysis of Matrix (in Chinese), Academic Press, China, 1987.

[6] J. G. Sun and Q. Ye, The unsolvability of inverse algebraic eigenvalue problems almost everywhere, *J. Comp. Math.*, **4**(1986), 21-226.

[7] S. F. Xu, The solvability of algebraic inverse eigenvalue problems almost everywhere, J. Comp. Math., Supplementary Issue, 1992, 152-157.