SOLUTION OF QUADRATIC MATRIX EQUATIONS BY LEAST-SQUARES METHOD*

By CHARLES C. LEE (University of North Dakota)

AND

H. P. NIU (University of Missouri, Kansas City)

Abstract. A least-square method is described for obtaining the solutions $A$ and $B$ for the matrix equations $Y^2 - YD + S = 0$ and $Y^2 - DY + S = 0$ respectively. No limitation is set on $A$, $B$, $D$, and $S$ except that they be square matrices. An illustrated example, including computing procedures, is discussed. The mathematical solutions presented should prove useful for solving the eigenvalue problem $(\lambda^2I + \lambda D + S)X = 0$, especially when the dimension of the matrices is large.

Introduction. The quadratic matrix equation $Y^2 - YD + S = 0$ is of fundamental importance to many engineering problems such as $n$-degree viscous damping systems. It is well known that an $n$-degree polynomial matrix equation with nonsingular leading coefficient matrix has at least $n$ solutions. It is extremely difficult to obtain solutions for the $n$-degree polynomial matrix equation; however, in practical applications the quadratic matrix equation is the most important one.

McFarland [1] introduced continued fractions for solving general quadratic equations. Fair [2] used the method to obtain a solution for the quadratic matrix equation. On the basis of Fair's investigation, the solution must be nonsingular, and the method converges very slowly if it is complex, which is likely to be true in practice. The present paper attempts to apply the least-square method [3, 4] to solving this quadratic matrix equation. Solving

\[ Y^2 - YD + S = 0 \]

and

\[ Y^2 - DY + S = 0 \]

is equivalent to finding $A$ and $B$ such that

\[ A + B = D \]

and

\[ AB = S, \]

where $A$ and $B$ are respectively the solutions to (1) and (2).

The solutions $A$ and $B$ may be applied to solve many eigenvalue problems of the type

\[ (\lambda^2I + \lambda D + S)X = 0. \]

* Received December 30, 1970; revised version received May 25, 1971.
It can be proved that

\[(\lambda^2 I + \lambda D + S) = (\lambda I + A)(\lambda I + B)\]  

(6)

is always valid. Consequently, the eigenvalues of \(A\) and \(B\) are the eigenvalues of Eq. (5).

A general quadratic matrix equation

\[T^2M + TC + K = 0\]  

(7)

can be transformed to the form of Eq. (1) by pre-multiplying Eq. (7) by \(M^{-1}\) and letting \(Y = M^{-1}TM, D = M^{-1}C\) and \(S = M^{-1}K\). Similarly, a general second-order eigenvalue problem \((\lambda^2 M + \lambda C + K)X = 0\) can also be transformed to the same form as Eq. (5).

There exist at least two sets of solutions for the system of equations (3) and (4). However, one set of solutions will be sufficient for application.

**Formulation—least square method.** Let

\[A_k = D_k + \sum_{m=1}^{k-1} R_m,\]  

(8)

\[B_k = (D - D_k) - \sum_{m=1}^{k-1} R_m,\]  

(9)

and

\[E_k = S - (A_k + R_k)(B_k - R_k),\]  

(10)

where \(A_k\) and \(B_k\) are the \(k\)th approximation of \(A\) and \(B\), the \(R_m\)'s are the correction matrices and \(E_k\) is the \(k\)th error matrix. \(D_k\) is an arbitrary complex matrix. If \(\lim_{k \to \infty} E_k = 0\), then \(A = \lim_{k \to \infty} A_k\) and \(B = \lim_{k \to \infty} B_k\) are the solutions to Eqs. (1) and (2) respectively. \(E_k\) can be also expressed as

\[E_k = E_{k-1} + A_k R_k - R_k B_k + R_k^2.\]  

(11)

It is seen from the above that the error matrix in Eq. (11) depends on the choice of the proper matrix \(R_k\). In this method, \(R_k\) is taken as a single-element matrix; i.e., all elements of \(R_k\) are zeroes except one specified element \(r_{m,n}\). Due to the fact that when (i) \(m \neq n\), \(R_k B_k\) has only one non-zero row, \(m\)th row, \(A_k R_k\) has only one non-zero column, \(n\)th column, and \(R_k^2 = 0\); but when (ii) \(m = n\), \(R_k B_k\) has only one non-zero row, \(m\)th row, \(A_k R_k\) has only one non-zero column, \(m\)th column, and \(R_k^2\) has only one non-zero element, \(r_{m,m}\), it is convenient to treat the problem by considering four different cases.

Assume

\[A_k = [a_{m,n} + a'_{m,n}]\]  

(12)

\[B_k = [b_{m,n} + b'_{m,n}]\]  

(13)

and

\[E_{k-1} = [e_{m,n} + e'_{m,n}],\]  

(14)

where \(a_{m,n}\), \(b_{m,n}\) and \(e_{m,n}\) are the real parts and \(a'_{m,n}\), \(b'_{m,n}\) and \(e'_{m,n}\) are the imaginary parts of \(A_k\), \(B_k\) and \(E_{k-1}\) respectively.

**Case 1.** \(r_{m,n} = \alpha, m = n\) and \(\alpha\) is real. Let
\[ F_k(\alpha) = \sum_p \left[ (e_{m,p} - ab_{m,p})^2 + (e_{p,m} + \alpha a_{p,m})^2 + (e'_{m,p} - \alpha b'_{m,p})^2 + (e'_{p,m} + \alpha a'_{p,m})^2 \right] \]
\[ - (e_{m,m} - ab_{m,m})^2 - (e_{m,m} + \alpha a_{m,m})^2 - (e'_{m,m} - \alpha b'_{m,m})^2 \]
\[ - (e'_{m,m} + \alpha a'_{m,m})^2 + (e_{m,m} - ab_{m,m} + \alpha a_{m,m} + \alpha^2)^2 + (e'_{m,m} - \alpha b'_{m,m} + \alpha a_{m,m})^2, \]
(15)

where \( F_k(\alpha) \) is the square sum of the real and imaginary parts of the \( m \)th row and column of \( E_k \). Then set \( F'(\alpha) = 0 \), minimizing the square sum of error, and rearrange the equation; we have

\[ 2\alpha^3 + 3(a_{m,m} - b_{m,m})\alpha^2 + \left[ \sum_p \left( a^2_{p,m} + a'^2_{p,m} + b^2_{m,p} + b'^2_{m,p} \right) + 2e_{m,m} \right. \]
\[ - 2(a_{m,m}b_{m,m} + a'_{m,m}b'_{m,m})\alpha + \sum_p \left( a_{p,m}e_{p,m} + a'_{p,m}e'_{p,m} - b_{m,p}e_{m,p} - b'_{m,p}e'_{m,p} \right) = 0 \]
(16)

There are three roots, \( \alpha_1 \), \( \alpha_2 \) and \( \alpha_3 \), to Eq. (16), and there exists at least one real root. The \( \alpha \) used to minimize \( F_k(\alpha) \) is real and \( F_k(\alpha) \) is the smallest among \( F_k(\alpha_1) \), \( F_k(\alpha_2) \) and \( F_k(\alpha_3) \).

**Case 2.** \( r_{m,n} = \beta_j \), \( m = n \) and \( \beta \) is real. Let

\[ F_k(\beta) = \sum_p \left[ (e_{m,p} - \beta a_{p,m})^2 + (e_{p,m} + \beta a_{p,m})^2 + (e'_{m,p} - \beta b'_{m,p})^2 + (e'_{p,m} + \beta a'_{p,m})^2 \right] \]
\[ + (e_{m,m} - \beta a_{m,m})^2 - (e_{m,m} + \beta a_{m,m})^2 - (e'_{m,m} - \beta b'_{m,m})^2 \]
\[ - (e'_{m,m} + \beta a_{m,m})^2 + (e_{m,m} - \beta b_{m,m} + \beta a_{m,m} + \beta^2)^2 \]
\[ + (e'_{m,m} + \beta a_{m,m} - \beta b_{m,m})^2, \]
(17)

where \( F_k(\beta) \) is the square sum of the real and imaginary parts of the \( m \)th column and row of \( E_k \). From \( F_k'(\beta) = 0 \),

\[ 2\beta^3 + 3(a_{m,m} - b_{m,m})\beta^2 + \left[ \sum_p \left( a^2_{p,m} + a'^2_{p,m} + b^2_{m,p} + b'^2_{m,p} \right) - 2e_{m,m} \right. \]
\[ - 2(a_{m,m}b_{m,m} + a'_{m,m}b'_{m,m})\beta + \sum_p \left( a_{p,m}e_{p,m} - a'_{p,m}e_{p,m} - b_{m,p}e_{m,p} - b'_{m,p}e_{m,p} \right) \]
\[ + b'_{m,p}e_{m,p} = 0 \]
(18)

can be obtained. Eq. (18) has three roots, \( \beta_1 \), \( \beta_2 \) and \( \beta_3 \). \( \beta \) must be real and \( F_k(\beta) \) is the smallest among \( F_k(\beta_1) \), \( F_k(\beta_2) \) and \( F_k(\beta_3) \).

**Case 3.** \( r_{m,n} = \alpha \), \( m \neq n \) and \( \alpha \) is real. Let

\[ F_k(\alpha) = \sum_p \left[ (e_{p,m} + \alpha a_{p,m})^2 + (e'_{p,m} + \alpha a'_{p,m})^2 + (e_{m,p} - ab_{m,p})^2 \right] \]
\[ + (e'_{m,p} - \alpha b'_{m,p})^2 - (e_{m,m} + \alpha a_{m,m})^2 - (e'_{m,m} + \alpha a'_{m,m})^2 \]
\[ - (e_{m,m} - ab_{m,m})^2 - (e'_{m,m} - \alpha b'_{m,m})^2 + (e_{m,m} + \alpha a_{m,m} - ab_{m,m})^2 \]
\[ + (e'_{m,m} + \alpha a'_{m,m} - ab_{m,m})^2, \]
(19)

where \( F_k(\alpha) \) is the square sum of the real and imaginary parts of the \( m \)th row and \( n \)th column of \( E_k \). Let \( F_k'(\alpha) = 0 \); then the corresponding
can be found.

Case 4. \( r_{m,n} = \beta j, m \neq n \) and \( \beta \) is real. Let

\[
F_k(\beta) = \sum_p \left( (e_{p,n} - \beta a_{p,n})^2 + (e_{p,m} + \beta b_{p,m})^2 + (e'_{p,n} + \beta a'_{p,n})^2 \right) \\
+ (e'_{m,n} - \beta b_{m,n})^2 - (e_{m,n} - \beta a_{m,n})^2 - (e_{m,n} + \beta b_{m,n})^2 \\
- (e'_{m,n} + \beta a_{m,n})^2 - (e_{m,n} - \beta b_{m,n})^2 + (e_{m,n} - \beta a'_{m,n} + \beta b'_{m,n})^2 \\
+ (e'_{m,n} + \beta a_{m,n} - \beta b_{m,n})^2,
\]

where \( F_k(\beta) \) is the square sum of the real and imaginary parts of the \( m \)th row and \( n \)th column of \( E_k \). Let \( F_k(\beta) = 0 \); then the corresponding

\[
\beta = \frac{\sum_p (e_{p,n} a_{p,m} + e'_{p,m} b_{n,p} - e_{m,p} b_{n,p} - e'_{p,n} a_{p,m})}{\sum_p (a_{p,m}^2 + a_{p,m}^2 + b_{n,p}^2 + b_{n,p}^2) - 2(a_{m,n} b_{n,n} + a_{m,n} b_{n,n})}
\]

is derived.

Theoretically, the order of picking \( m \)'s and \( n \)'s is indifferent. For programming purposes it is advisable to work in the following order:

(a) diagonal elements, real parts,
(b) diagonal elements, imaginary parts,
(c) off-diagonal elements, real parts,
(d) off-diagonal elements, imaginary parts.

Example. Find \( Y \) such that \( Y^2 - YD + S = 0 \), if

\[
D = \begin{bmatrix} 4.0 & -2.0 & 0.0 \\ -2.0 & 5.0 & -3.0 \\ 0.0 & -3.0 & 3.0 \end{bmatrix} \quad \text{and} \quad S = \begin{bmatrix} 30.0 & -20.0 & 0.0 \\ -20.0 & 50.0 & -30.0 \\ 0.0 & -30.0 & 30.0 \end{bmatrix}
\]

Solution: Let \( A + B = D \) and \( AB = S \); the solutions \( A \) and \( B \) can be found. Assume \( A = A_R + A_I, B = B_R + B_I \) and \( E = E_R + E_I \), where \( A_R, B_R \) and \( E_R \) are the real parts and \( A_I, B_I \) and \( E_I \) are imaginary parts of \( A, B \) and \( E \) respectively. \( E \) is the error matrix.

\[
A_R = \begin{bmatrix} 0.2000000000D + 01 & -0.5085521289D + 00 & -0.2627694325D + 00 \\ -0.1491447871D + 01 & 0.2500000000D + 01 & -0.1300853953D + 01 \\ 0.2627694325D + 00 & -0.1699146978D + 01 & 0.1500000000D + 01 \end{bmatrix}
\]

\[
A_I = \begin{bmatrix} -0.2882338745D + 01 & -0.2584494448D + 01 & 0.3268767162D + 01 \\ -0.2584494448D + 01 & 0.4380289005D + 01 & -0.3737220748D + 01 \\ 0.3268767162D + 01 & -0.3737220748D + 01 & -0.3770905143D + 00 \end{bmatrix}
\]
\[ B_r = D - A_r \text{,} \quad B_t = -A_t \]

\[
E_r = \begin{bmatrix}
0.16908299D - 13 & 0.17952250D - 13 & 0.22663556D - 13 \\
0.17622398D - 13 & -0.15554814D - 13 & 0.21458441D - 15 \\
0.22844725D - 13 & 0.20852945D - 13 & 0.34690585D - 13 \\
-0.16999174D - 15 & -0.81385618D - 14 & 0.38100561D - 14 \\
0.72041499D - 15 & 0.24701507D - 15 & 0.12714956D - 13 \\
-0.40905453D - 14 & -0.12292335D - 14 & -0.25058198D - 15
\end{bmatrix}
\]

\[ E_I = \begin{bmatrix}
0.16908299D - 13 & 0.17952250D - 13 & 0.22663556D - 13 \\
0.17622398D - 13 & -0.15554814D - 13 & 0.21458441D - 15 \\
0.22844725D - 13 & 0.20852945D - 13 & 0.34690585D - 13 \\
-0.16999174D - 15 & -0.81385618D - 14 & 0.38100561D - 14 \\
0.72041499D - 15 & 0.24701507D - 15 & 0.12714956D - 13 \\
-0.40905453D - 14 & -0.12292335D - 14 & -0.25058198D - 15
\end{bmatrix}
\]

\[ Y = A \text{ is a solution of } Y^2 - YD + S = 0. \text{ B is not a solution for } Y^2 - YD + S = 0 \text{ in general, unless } A \text{ and } B \text{ commute. It can be proved that } B \text{ is also a solution for this example.} \]

**Discussion and conclusion.** (1) In many applications the governing matrix equation is

\[ \left( \lambda^2 M + \lambda C + K \right) X = 0 \quad (23) \]

where \( M, C \) and \( K \) are real and symmetric matrices. The eigenvalues of Eq. (23) are in pairs of complex conjugate, i.e., if \( a + bj \) is one eigenvalue, so is \( a - bj \). It is desirable to keep the transformed version of Eq. (23) symmetric. Let \( D = M^{-1/2}CM^{-1/2} \), \( S = M^{-1/2}KM^{-1/2} \) and \( Z = M^{-1/2}XM^{-1/2} \); then

\[ \left( \lambda^2 I + \lambda D + S \right) Z = 0 \quad (24) \]

can be achieved. In Eq. (24) \( D \) and \( S \) are symmetric and in most applications they are real matrices. \( M^{-1/2} \) is the square root of \( M^{-1} \). Since \( M^{-1} \) can be transformed to \( M^{-1} = UGU^{-1} \), where \( U \) is a square unitary matrix, \( U^{-1} \) is the inverse of \( U \) and \( G \) is a diagonal matrix, \( M^{-1/2} \) can be defined as \( UG^{1/2}U^{-1} \), where \( G^{1/2} \) is a diagonal matrix with diagonal elements as square roots of diagonal elements of \( G \)’s. It is desirable to find \( A \) and \( B \) such that

\[ \left( \lambda I + A \right) \left( \lambda I + B \right) = \lambda^2 I + \lambda D + S \quad (25) \]

and eigenvalues of \( A \) are complex conjugates of eigenvalues of \( B \). Using then the facts that \( \text{tr} A \) and \( \text{tr} B \) are complex conjugate of each other and that \( \text{tr} A + \text{tr} B = \text{tr} D \), it is possible to assume \( A_1 = \frac{1}{2}D - Ij \) and \( B_1 = \frac{1}{2}D + Ij \) without any correction from Eq. (16) necessary.

(2) Introducing imaginary parts in \( A_1 \) and \( B_1 \) is essential when \( D \) and \( S \) are real.

(3) One may envision that there are at least two sets of solutions satisfying \( A + B = D \) and \( AB = S \); therefore it is necessary to reselect \( A_1 \) and \( B_1 \) after several iterations. Improper selection of \( A_1 \) and \( B_1 \) may lead to \( \lim_{k \to \infty} E_k = E \), where \( E \) is a non-zero matrix. It is recommended that the last few iterations be printed so that a new set of \( A_1 \) and \( B_1 \) can be determined.

(4) Once the proper \( A_1 \) and \( B_1 \) have been found, the least-square method will give very accurate solutions.
References


