Lecturas Matemáticas
Volumen **29** *(2008), páginas 13–22*
ISSN 0120–1980

Solving algebraic Riccati equation using Kovarik's method Part I: Theoretical results

H. Esmaeili
Bu-Ali Sina University, Hamedan, Iran

Abstract. One of the common methods to solve algebraic Riccati equations is using matrix sign function to calculate stable invariant subspaces. In this paper, we show that it is possible to compute the matrix sign function using the quadratically convergent Kovarik's method. Since the explicit inversion of a matrix, in every iteration, must be computed, this will introduce unstability in numerical results. The modification for Kovarik's method is presented here in such way, that the elimination of matrix inversion is considered and, therefore, computational process is limited for every iteration to matrix-by-matrix multiplications.

Key words and phrases. Matrix sign function, Kovarik's method, Algebraic Riccati equation.

2000 AMS Mathematics Subject Classification. 65F15, 15A30.

Resumen. Uno de los métodos comunes para resolver ecuaciones algebraicas de Ricatti es usar la función signo matricial para calcular subespacios invariantes estables. En este artículo mostramos que es posible calcular esta función usando el método cuadráticamente convergente de Kovarick.

1. Introduction

The algebraic Riccati equations of the form

$$A^T X + XA - XRX + G = 0 \tag{1}$$

appears in various control and filtering problems of continuous time systems, where A, R and G are given $n \times n$ matrices, R and G are symmetric positive semidefinite. A special case of (1) is the algebraic Bernouli equation

$$A^T X + XA - XRX = 0$$

with $G = 0$. In applications, the desired solution X of (1) must be symmetric positive semidefinite and stabilizing in the sense that all eigenvalues of $A - RX$ have negative real parts. Under mild technical assumptions on the problems, the existence and uniqueness of such a solution is guaranteed [24].

There have been offered various numerical methods to solve algebraic Riccati equation, which can be reviewed in [19]. The key of numerical technique to solve (1) is to convert the problem to a stable invariant subspace problem of the $2n \times 2n$ Hamiltonian matrix

$$H = \begin{pmatrix} A^T & G \\ R & -A \end{pmatrix},$$

i.e., finding the invariant subspace corresponding to the eigenvalues of H with negative real parts. We observe that X satisfies (1) if and only if

$$\begin{bmatrix} A^T & G \\ R & -A \end{bmatrix} \begin{bmatrix} X & -I_n \\ I_n & 0 \end{bmatrix} = \begin{bmatrix} X & -I_n \\ I_n & 0 \end{bmatrix} \begin{bmatrix} -(A - RX) & -R \\ 0 & (A - RX)^T \end{bmatrix}. \tag{2}$$

The matrix H is said to be Hamiltonian if $JH = (JH)^T$, where $J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}$. If λ is an eigenvalue of the Hamiltonian matrix H, then $-\overline{\lambda}$ is too. It is well-known that H has precisely n eigenvalues with negative real parts. If columns of $(U^T \;\; V^T)^T$ span the desired invariant subspace, then U is invertible and the solution X of (1) is given by $X = -VU^{-1}$ [16].

The numerical solution of the Riccati equation by the computation of the stable invariant subspace has received considerable interest over decades. The existing methods include the Schur vector method [16], the Hamiltonian QR-algorithm [1, 6], the SR algorithm [5], and the matrix sign function method [2, 7, 23]. Among all these existing algorithms, the matrix sign function method is apparently the most suitable method for parallel implementation [17]. However, in the inner loop iteration of the method, it is required to compute matrix inversion, which is potentially numerically unstable when the matrix is ill-conditioned to inversion [9].

The purpose of this article is to calculate matrix sign function, and therefore to solve algebraic Riccati equation, using Kovarik's method [15]. Kovarik's method is quadratically convergent and in every iteration of which, the explicit inverse of a matrix must be calculated. Here, we present a modification for Kovarik's method in such way every iteration needs only matrix-by-matrix multiplication. This modification is comparable to Kovarik's method. In section 2, we introduce matrix sign function and Newton method for its computation. In section 3, the Kovarik's method and some of its convergence property is discussed. We show, in section 4, how to calculate matrix sign function by

Kovarik's method (in the future work, we will turn to numerical results of the method).

2. Matrix sign function

Let us first briefly recall the matrix sign function method for computing the stable invariant subspace. Suppose

$$H = S\begin{pmatrix} J_- & 0 \\ 0 & J_+ \end{pmatrix} S^{-1}$$

be the Jordan canonical form of the matrix H, where the eigenvalues of J_- are the eigenvalues of H with negative real parts and the eigenvalues of J_+ are the eigenvalues of H with positive real parts. Then the matrix sign function $\operatorname{sign}(H)$ is defined by [2, 23]

$$\operatorname{sign}(H) = S\begin{pmatrix} -I_n & 0 \\ 0 & +I_n \end{pmatrix} S^{-1}.$$

If H has an eigenvalue with zero real part, then $\operatorname{sign}(H)$ will not be defined. It follows immediately from the definition that if $\operatorname{sign}(H)$ is defined and C is nonsingular, then $\operatorname{sign}(C^{-1}HC) = C^{-1}\operatorname{sign}(H)C$. The matrix

$$P_- = \frac{1}{2}(I_{2n} - \operatorname{sign}(H)) = S\begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix} S^{-1}$$

is the projection on the stable invariant subspace corresponding to the eigenvalues with negative real parts. Then the first n columns $(Q_{11}^T \;\; Q_{21}^T)^T$ of Q in the rank revealing QR decomposition of P_- span the stable invariant subspace, and the desired solution X of (1) is given by $X = Q_{21}Q_{11}^{-1}$.

The matrix sign function method was first introduced by Roberts [23] for solving the algebraic Riccati equation. However, it was soon extended to solving the spectral decomposition problem [4]. A survey can be found in [3, 18].

Since the matrix sign function $\operatorname{sign}(H)$ satisfies $\operatorname{sign}^2(H) = I_{2n}$ [10, 11], we may use the Newton method

$$H_0 = H, \quad H_{j+1} = \frac{1}{2}\left(H_j + H_j^{-1}\right), \quad j = 0, 1, 2, \ldots, \tag{3}$$

to compute $\operatorname{sign}(H)$. It can be shown that the iteration is globally and ultimately quadratically convergent with $\lim_{j\to\infty} H_j = \operatorname{sign}(H)$, provided H has no pure imaginary eigenvalues [12, 23]. The Newton iteration is terminated when $\|H_{j+1} - H_j\|_1 \le \tau\|H_j\|_1$, where τ is a given tolerance value. Starting (3) with the Hamiltonian matrix $H_0 = H$ makes each H_j Hamiltonian, too. Unfortunately, in finite precision arithmetic, the ill conditioning of a matrix H_j with respect to inversion and rounding errors, may destroy the convergence of the Newton iteration (3), or cause convergence to the wrong answer. There exist different scaling schemes to speedup the convergence of the iteration, and make it more suitable for parallel computation [14, 19].

Byers [7] improved the method (3) by changing the inversion of nonsymmetric matrices to symmetric ones for a Hamiltonian matrix. More results can be found in Laub's review article [17]. There have been some modifications for Newton's method to remove the necessity of inverse calculation [14, 10, 7]. For example, if $\|H^2 - I_{2n}\| < 1$, then we can use Newton–Schulz iteration

$$H_0 = H, \quad H_{j+1} = \frac{1}{2} H_j \left(3I_{2n} - H_j^2\right), \quad j = 0, 1, 2, \ldots,$$

to avoid the use of the matrix inverse.

If H has no eigenvalue with zero real part, then it is shown that the matrix sign function satisfies [11, 13]

$$\mathrm{sign}(H) = (H^2)^{-1/2} H.$$

In section 4, we use of this definition for computing matrix sign function by Kovarik's method.

3. Kovarik's method

Kovarik [15] proposed his algorithm for approximate orthogonalization of a finite linearly independent set of vectors in a Hilbert space. His algorithm is some kind of iterative version of the classical Gram-Schmidt one and also some of its direct applications have been derived for variational finite element formulation of elliptic problems and least squares. Kovarik showed that the approximate orthogonalization method has quadratic convergence. The main difficulty with this method is the necessity of computing the inverse of a matrix explicitly in every iteration. Many years after Kovarik, Popa [22] adapted and extended his algorithm for a set of arbitrary vectors in $\mathcal{R}^n$, and proved that the transformed matrix columns, in addition to rows, are "quasi-orthogonal".

Suppose $m \leq n$ and M is a $m \times n$ real matrix of rank r. Approximate orthogonalization method of Kovarik is to transform M to a matrix with approximate orthogonal rows (see (8)). This method starts with $M_0 = M$ and produces the sequences of K_k and M_k, $k \geq 0$, as follow:

$$K_k = (I_m - M_k M_k^T)(I_m + M_k M_k^T)^{-1}, \quad M_{k+1} = (I_m + K_k) M_k, \quad k \geq 0. \tag{4}$$

Kovarik showed that if the rows of matrix M are linearly independent and

$$M_\star = \left[(MM^T)^{1/2}\right]^{-1} M$$

then,

(a) the rows of $M_\star$ are orthogonal;
(b) the sequence of matrices M_k, $k \geq 0$, defined in (4), are convergent to $M_\star$. In addition,

$$\|K_0\| < 1$$

and

$$\|M_\star - M_k\|_2 \le \|K_0\|_2^{2^k}, \quad \forall\, k \ge 1. \tag{5}$$

The inequality (5) shows that Kovarik's method is convergent of the second order.

Based on the fact that rows of M are linearly independent, we can conclude that Gram's matrix MM^T is symmetric and positive definite. Therefore, the matrix $M_\star$ is well-defined. On the other hand, if rows of matrix M are not linearly independent, there still exists matrix $(MM^T)^{1/2}$ but not invertible. In this case, we regard "natural" generalization of $M_\star$ as

$$M_\infty = \left[(MM^T)^{1/2}\right]^+ M\,.$$

Here, B^+ shows the Moore-Penrose pseudo–inverse of matrix B [9]. It is proved [22] that M_k matrices, in this case, converge to M_∞ and the rows of M_∞ are approximately orthogonal.

We know that $I + M_k M_k^T$ is invertible if and only if [9]

$$\|M_k M_k^T\|_2 < 1. \tag{6}$$

For $k = 0$, the condition of (6) turns into

$$\|MM^T\|_2 < 1. \tag{7}$$

If (7) is the case, it is proved that (6) holds for all $k \ge 1$. On the other hand, the assumption (7) is not restrictive. It is enough to scale matrix M, say, as the following:

$$M^{new} := \frac{1}{\sqrt{\|M\|_1 \|M\|_\infty + 1}}\, M.$$

Hence, without loss of generality, we suppose that matrix M satisfies (7). Therefore, by putting $\delta = 1 - \lambda_{\min}$, where $\lambda_{\min}$ is the smallest nonzero eigenvalue of MM^T, we have

$$M_\infty := \lim_{k\to\infty} M_k = \left[(MM^T)^{1/2}\right]^+ M$$

and

$$\|M_k - M_\infty\|_2 \le \delta^{2^k}, \quad \forall\, k \ge 0.$$

Suppose that SVD factorization of matrix M is

$$U^T M V = (\sigma_1, \dots, \sigma_r, 0, \dots, 0),$$

where

$$\sigma_1 \ge \dots \ge \sigma_r > 0, \quad r = \mathrm{rank}\,(M).$$

By putting

$$\tilde{I} = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}_{m\times m},$$

we can prove [22] that the following "approximately orthogonal" relationship satisfies among the rows of M_∞:

$$\langle (M_\infty)_i, (M_\infty)_j \rangle = \langle (U)_i \tilde{I}, (U)_j \rangle. \tag{8}$$

If the rows of matrix M are linearly independent, then $\tilde{I} = I$ and (8) is the same classical orthogonal relation. Popa showed that there is a similar relationship between the columns of M_∞ and

$$\lim_{k\to\infty} \kappa_2(M_k) = \kappa_2(M_\infty) = 1.$$

Despite the quadratic convergence of Kovarik's algorithm, there is a difficult computational aspect related to the matrix inversion in (4) at each of its iterations. Several modifications have been proposed for Kovarik's method, all of which try to eliminate the necessity to explicitly compute the inverse. These are upon using some approximations for $(I + A_k A_k^T)^{-1}$, based on Taylor's series of particular functions, and are commonly linearly convergent [20, 21, 22].

We have offered [8] a single parameter class of modifications for Kovarik's method by using a special quadratic polynomial interpolation which is not needing any inverse calculation but only matrix-by-matrix multiplications. We have proved that the convergence of the mentioned class is linear and its asymptotic error constant is a function of parameter. We chose the best value of parameter such that the convergence is monotonic and asymptotic error constant is very small. Despite the fact that Kovarik's method and our modification are convergent of the second and first order, respectively, numerical experiments show that our modification works as efficiently as Kovarik's method in iterations and costs. On the other hand, our modification provides good results with less work and is free from numerical problems.

4. **Kovarik's method to calculate matrix sign function**

We consider Kovarik's method for Hamiltonian matrix

$$H_0 = H = \begin{pmatrix} A^T & G \\ R & -A \end{pmatrix}$$

as the following:

$$K_k = (I_{2n} - H_k^2)(I_{2n} + H_k^2)^{-1}, \quad H_{k+1} = (I_{2n} + K_k) H_k, \quad k \geq 0. \tag{9}$$

Suppose that Jordan canonical form of matrix H_0 is as

$$S^{-1} H_0 S = \begin{pmatrix} J_-^{(0)} & 0 \\ 0 & J_+^{(0)} \end{pmatrix} = \Sigma^{(0)}.$$

Inductively, assume that Jordan canonical form of matrix H_k can be written as

$$S^{-1} H_k S = \begin{pmatrix} J_-^{(k)} & 0 \\ 0 & J_+^{(k)} \end{pmatrix} = \Sigma^{(k)}.$$

In this case,

$$\begin{aligned} S^{-1}H_{k+1}S = S^{-1}(I_{2n}+K_k)H_kS = S^{-1}H_kS + S^{-1}(I_{2n}-H_k^2)(I_{2n}+H_k^2)^{-1}H_kS \\ = S^1H_kS + S^{-1}(I_{2n} - H_kSS^{-1}H_k)SS^{-1}(I_{2n} + H_kSS^{-1}H_k)^{-1}SS^{-1}H_kS \\ = S^{-1}H_kS + S^{-1}(I_{2n} - H_kSS^{-1}H_k)S[S^{-1}(I_{2n} + H_kSS^{-1}H_k)S]^{-1}S^{-1}H_kS \\ = S^{-1}H_kS + (I_{2n} - S^{-1}H_kSS^{-1}H_kS)[I_{2n} + S^{-1}H_kSS^{-1}H_kS]^{-1}S^{-1}H_kS \\ = \Sigma^{(k)} + [I_{2n} - (\Sigma^{(k)})^2][I_{2n} + (\Sigma^{(k)})^2]^{-1}\Sigma^{(k)} = \Sigma^{(k+1)} \end{aligned}$$

is the Jordan canonical form of the matrix H_{k+1}, in which

$$\Sigma^{(k+1)} = \Sigma^{(k)} + [I_{2n} - (\Sigma^{(k)})^2][I_{2n} + (\Sigma^{(k)})^2]^{-1}\Sigma^{(k)}.$$

Therefore,

$$J_-^{(k+1)} = J_-^{(k)} + \left(I_n - (J_-^{(k)})^2\right)\left(I_n + (J_-^{(k)})^2\right)^{-1} J_-^{(k)}$$

$$J_+^{(k+1)} = J_+^{(k)} + \left(I_n - (J_+^{(k)})^2\right)\left(I_n + (J_+^{(k)})^2\right)^{-1} J_+^{(k)}.$$

If

$$J_+^{(k)} = \operatorname{diag}(\sigma_1^{(k)}, \sigma_2^{(k)}, \ldots, \sigma_n^{(k)}),$$
$$J_-^{(k)} = \operatorname{diag}(\eta_1^{(k)}, \eta_2^{(k)}, \ldots, \eta_n^{(k)}),$$

then,

$$\sigma_j^{(k+1)} = \left[1 + \frac{1 - (\sigma_j^{(k)})^2}{1 + (\sigma_j^{(k)})^2}\right]\sigma_j^{(k)} = \frac{2\sigma_j^{(k)}}{1 + (\sigma_j^{(k)})^2}, \quad j = 1, \ldots, n, \quad k \geq 0, \tag{10}$$

$$\eta_j^{(k+1)} = \left[1 + \frac{1 - (\eta_j^{(k)})^2}{1 + (\eta_j^{(k)})^2}\right]\eta_j^{(k)} = \frac{2\eta_j^{(k)}}{1 + (\eta_j^{(k)})^2}, \quad j = 1, \ldots, n, \quad k \geq 0. \tag{11}$$

Hence, the convergence of Kovarik's method is reduced to that of real numbers sequences (10) and (11). On the other hand, as we need only the sign of real parts of the eigenvalues, without loss of generality, we can suppose that the absolute value of all eigenvalues of H_0 is less than one and H_0 has no pure imaginary eigenvalues. In addition, to calculate the projection matrix P_-, we need to have those eigenvalues with negative real parts. Therefore, we focus only on sequence (11). This sequence starts from an initial value $\eta_j^{(0)}$, with $0 < \eta_j^{(0)} < 1$. If η_j is the limit of sequence (11), then $\eta_j = 2\eta_j/(1+\eta_j^2)$, so that it follows $\eta_j = \pm 1$. As sequence (11) is strictly ascending for $\eta_j^{(0)} \in (0, 1)$, we have $\eta_j = 1$.

Kovarik's method (9) needs the calculation of matrix inverse in every iteration. Various modifications have been offered which have tried to eliminate this precondition. These modifications are obtained with approximation of

$1/(1+(\eta_j^{(k)})^2)$ (and therefore, $(I_{2n}+H_k^2)^{-1}$). For example, the approximation of

$$\frac{1}{1+(\eta^{(k)})^2} \approx \sum_{s=0}^{q_k} \left(-(\eta_j^{(k)})^2\right)^s$$

has been used in [20] resulting in the modification of

$$K_k = (I_{2n}-H_k^2)\sum_{s=0}^{q_k}(-H_k^2)^s, \quad H_{k+1} = (I_{2n}+K_k)H_k, \quad \forall\, k \geq 0$$

for Kovarik's method. Here, q_k is a positive odd arbitrary integer which must be chosen large enough to obtain a good convergence. Moreover, for q_k to be even, it's likely not to be convergent.

We have offered a single parameter modification class in [8] as

$$K_k = (I_m - M_kM_k^T)(I_m - \alpha M_kM_k^T), \quad M_{k+1} = (I_m+K_k)M_k, \quad k \geq 0 \tag{12}$$

for Kovarik's method in which $\alpha \in (0,1]$ is a parameter. We proved that every modification in this class is linearly convergent. In addition, the choice of $\alpha = 0.507$ gives us

$$K_k = (I_m - M_kM_k^T)(I_m - 0.507M_kM_k^T), \quad M_{k+1} = (I_m+K_k)M_k, \quad k \geq 0 \tag{13}$$

that is the best linearly convergent modification with a strict convergence and asymptotic error constant of 0.014. Numerical experiments show that the number of iterations of the above modification is the same as that of Kovarik's method so that the calculation time is less. In addition, calculation cost in every iteration is restricted to matrix-by-matrix multiplications and is less than that for Kovarik's method. We note that the above modification does not need any calculation of matrix inverse. Finally, the following method is suggested to calculate matrix sign function:

$$K_k = (I_{2n}-H_k^2)(I_{2n}-0.507H_k^2), \quad H_{k+1} = (I_{2n}+K_k)H_k, \quad k \geq 0. \tag{14}$$

In future work, we will express the numerical results of this modification.

Conclusion

In this paper, we study the matrix sign function method to solve algebraic Riccati equation. We suggest an algorithm, based on Kovarik's approximate orthogonalization method, to calculate the matrix sign function. It is theoretically observed that Kovarik's method can be used to calculate matrix sign function, and in turn, to solve algebraic Riccati equation. Numerical tests will be analyzed in future works.

References

[1] G. Ammar, P. Benner, & V. Mehrmann, *A multishift algorithm for the numerical solution of algebric Riccati equations. Elect. Trans. on Numer. Anal., Kent State University*, **1** (1993), 33–48.

[2] Z. Bai, J. Demmel, *Using the matrix sign function to compute invariant subspaces* SIAM J. Matrix Anal. Applic. **19** (1998), 205–225.

[3] Z. Bai & J. Demmel, *Design of a parallel nonsymmetric eigenroutine toolbox*, Part I. In R. F. Sincovec et al. (Editors), *Proceedings of the Sixth SIAM Conference on Parallel Processing for Scientific Computing*. SIAM, 1993.

[4] A. N. Beavers & E. D. Denman, *A computaional method for eigenvalue and eigenvectors of a matrix with real eigenvalues*. Numer. Math. **21** (1973), 389–396.

[5] Bunse-Gerstner & V. Mehrmann, *A symplectic QR-like algorithm for the solution of the real algebric Riccati equation*. IEEE Trans. Autom. Control, **AC-31** (1986), 1104–1113.

[6] R. Byers, *A Hamiltonian QR algorithm*. SIAM J. Sci. Stat. Comp. **7** (1986), 212–229.

[7] R. Byers, *Solving the algeraic Riccati equation with he matrix sign function*. Lin. Alg. Appl. **85** (1987), 267–279.

[8] H. Esmaeili, *A modification for Kovarik's method*. Lecturas Matemáticas **27** (2006), 1–12.

[9] G. H. Golub & C. F. van Loan, *Matrix computations*. The Jhon's Hopkins University Press, Baltimore, 1983.

[10] N. J. Higham, *Computing the polar decomposition –with applications*. SIAM J. Sci. Stat. Comput. **7** (1986), 1160-1174.

[11] N. J. Higham, *The matrix sign decomposition and its relation to the polar decomposition*. Lin. Alg. Appl. **20** (1994), 212–213.

[12] J. Howland, *The sign matrix and the separation of matrix eigenvalues* Lin. Alg. Appl. **49** (1983), 221–232.

[13] C. S. Kenney & A. J. Laub, *The matrix sign function*. IEEE Trans. Automat. Control **40** (1995), 1330–1348.

[14] C. Kenney & A. Laub, *Rational iteration methods for the matrix sign function*. SIAM J. Mat. Anal. Appl. **21** (1991), 487–494.

[15] Z. Kovarik, *Some iterative methods for improving orthogonality*. SIAM J. Numer. Anal. **7** (1970), 386–389.

[16] A. J. Laub, *A Schur method for solving algebric Riccati equations*. IEEE Trans. Auto. Control **AC-24** (1979), 913–921.

[17] A. Laub, *Invariant subspace method for the numerical solution of Riccati equations*. In S. Bittanti, A.J. Laub & J. C. Willems (editors),*Riccati Equations*, Berlin, Springer–Verlag, 1990.

[18] C–C. Lin & E. Zmijewwski, *A parallel algorithm for computing the eigenvalues of an unsymetric matrix on an SIMD mesh of processors*. Department of Computer Science TRCS 91–15, University of California, Santa Barbara, CA, July 1991.

[19] V. Mehrmann, *The linear quadratic control problem: Theory and numerical algorithms*. Habilitationsschrif Universität Bielefeld FRG, 1987.

[20] C. Popa, *A method for improving orthogonality of rows and columns of matrices*. International J. Computer Mathematics **77** (2001), 469–480.

[21] C. Popa, *Extention of an approximate orthogonalization algorithm to arbitrary rectangular matrices*. Lin. Alg. Appl. **33** (2001), 181–192.

[22] C. Popa, *Modified Kovarik algorithm for approximate orthogonalization of arbitrary matrices*. International J. Computer Mathematics **80** (2003), 519–525.

[23] J. Roberts, *Linear model reduction and solution of the algebric Riccati equation* Inter. J. Control **32** (1980), 677–687.

[24] M. Wonham, *Linear Multivariable Control Theory: A Geometric Approach*. Springer–Verlag, New York, 2nd edition, 1979.

(Recibido en abril de 2007. Aceptado para publicación en abril de 2008)

H. Esmaeili, Department of Mathematics
Bu-Ali Sina University, Hamedan, Iran
e-mail: Esmaeili@basu.ac.ir