Japan Journal of Industrial and Applied Mathematics (2020) 37:831–849
https://doi.org/10.1007/s13160-020-00421-3

ORIGINAL PAPER

A further study on a nonlinear matrix equation

Jie Meng[1] · Hongjia Chen[2] · Young-Jin Kim[1] · Hyun-Min Kim[3]

Received: 2 October 2019 / Revised: 26 March 2020 / Published online: 1 May 2020

Abstract
The nonlinear matrix equation $X^p = R + M^T(X^{-1} + B)^{-1}M$, where p is a positive integer, M is an arbitrary $n \times n$ real matrix, R and B are symmetric positive semidefinite matrices, is considered. When $p = 1$, this matrix equation is the well-known discrete-time algebraic Riccati equation (DARE), we study the convergence rate of an iterative method which was proposed in Meng and Kim (J Comput Appl Math 322:139–147, 2017). For the generalized case $p \geq 1$, a structured condition number based on the classic definition of condition number is defined and its explicit expression is obtained. Finally, we give some numerical examples to show the sharpness of the structured condition number.

Keywords Matrix equation · Symmetric positive definite · Cyclic reduction · Structured condition number

Mathematics Subject Classification 15A24 · 65F10 · 65H10

✉ Hyun-Min Kim
hyunmin@pusan.ac.kr

Jie Meng
mengjie@pusan.ac.kr

Hongjia Chen
chen@mma.cs.tsukuba.ac.jp

Young-Jin Kim
kyj2806@gmail.com

[1] Finance·Fishery·Manufacture Industrial Mathematics Center on Big Data, Pusan National University, Busan 46241, Republic of Korea

[2] Department of Mathematics, School of Science, Nanchang University, Nanchang 30031, People's Republic of China

[3] Department of Mathematics and Finance·Fishery·Manufacture Industrial Mathematics Center on Big Data, Pusan National University, Busan 46241, Republic of Korea

1 Introduction

We consider the nonlinear matrix equation

$$X^p = R + M^T(X^{-1} + B)^{-1}M, \tag{1}$$

where p is a positive integer, $M \in \mathbb{R}^{n\times n}$, B and R are symmetric positive semidefinite matrices.

For the special case $p = 1$, Eq. (1) is exactly

$$X = R + M^T(X^{-1} + B)^{-1}M, \tag{2}$$

which, under certain condition, is a simplified symmetric form of the well-known discrete-time algebraic Riccati equation (DARE)

$$X = M^T XM - M^T XE(G + E^T XE)^{-1}E^T XM + C^T C, \tag{3}$$

where $M \in \mathbb{R}^{n\times n}$, $E \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{q\times n}$ and $G \in \mathbb{R}^{m\times m}$ is a symmetric positive definite matrix. It has been proved that if (M, E) in DARE (3) is a stabilizable pair and (C, M) is a detectable pair,[1] then DARE (3) has a unique symmetric positive definite solution, see [7, 12, 22]. Under this assumption, with $B = EG^{-1}E^T$ and $R = C^T C$, the DARE (3) can be rewritten in the symmetric form as Eq. (2).

The symmetric positive definite solution of Eq. (2) or, equivalently, the DARE (3) is of theoretical and practical importance in some control problems, see [1, 3, 8, 13, 15, 16, 23] and the references therein. For finding the unique symmetric positive definite solution, Komaroff [11] proposed a fixed-point iteration

$$X_{k+1} = M^T(X_k^{-1} + B)^{-1}M + R, \quad k = 0, 1, 2, \ldots, \tag{4}$$

and proved that the matrix sequence $\{X_k\}$ converges to the unique positive definite solution when M is nonsingular, $R > 0$ and $B > 0$. Later, Dai and Bai [5] showed that the matrix sequence $\{X_k\}$ still converges even if M is singular, and they modified the fixed-point iteration to a new iteration as

$$\begin{cases} Y_0 = (R^{-1} + B)^{-1}, \\ X_{k+1} = M^T Y_k M + R, \\ Y_{k+1} = Y_k\big(2I - (X_{k+1}^{-1} + B)Y_k\big). \end{cases} \tag{5}$$

Comparing with the fixed-point iteration (4), even though the modified fixed-point iteration (5) avoids computing the inverse of $X_k^{-1} + B$, it still involves the computation of X_k^{-1} at each step, which may lead to numerical instability problems due to the matrix inverse. In [18], under the condition that B is nonsingular, an inversion-free variant iteration which completely avoids the computation of X_k^{-1} is proposed as

[1] (M, E) is a stabilizable pair if $\omega^T E = 0$ and $\omega^T M = \lambda\omega^T$ hold for some constant λ and vector ω, then $|\lambda| < 1$ or $\omega = 0$. (C, M) is a detectable pair if (M^T, C^T) is a stabilizable pair.

$$\begin{cases} X_0 = R + M^T B^{-1} M, \\ Y_0 = (BX_0 B + B)^{-1}, \\ Y_{k+1} = 2Y_k - Y_k BX_k BY_k - Y_k BY_k, \\ X_{k+1} = R + M^T B^{-1} M - M^T Y_{k+1} M. \end{cases} \tag{6}$$

In the process of iteration (6), only the inverse of matrix B is required, and since B is a given positive definite matrix, B^{-1} can be readily obtained.

It was proved in [5] that both the convergence rates of the basic fixed-point iteration (4) and the modified fixed-point iteration (5) remains linear. But, as far as we know, there is no work on the convergence rate of the inversion-free variant iteration (6). In this paper, we study the convergence behaviour of iteration (6).

For the general case where $2 \leq p < \infty$, the existence of a unique positive definite solution of Eq. (1) is a direct consequence of the results by Jung et al. [10]. In [18], two iteration methods for computing the unique positive definite solution, as well as a lower and an upper bound of the solution, are given. Here, as a continuation of the previous results, we develop a structured condition number for the nonlinear matrix equation (1). It is validated by numerical examples that the newly proposed structured condition number successively measures the sensitivity of the solution.

The rest of this paper is organized as follows. In Sect. 2, we derive the convergence rate of the inversion-free variant iteration (6). In Sect. 3, a structured condition number is defined and the explicit expression is derived. In Sect. 4, we give a numerical example to show the sharpness of the proposed structured condition number.

We begin with the notation used throughout this paper. $\mathbb{R}^{n\times n}$ is the set of $n \times n$ matrices with elements on field $\mathbb{R}$. $\mathbb{S}^{n\times n}$ and $\mathbb{P}(n)$ are, respectively, the set of symmetric matrices and positive definite matrices. $\|\cdot\|$ and $\|\cdot\|_F$ are the spectral norm and the Frobenius norm, respectively. For a symmetric matrix H, $\lambda_{\max}(H)$ ($\lambda_{\min}(H)$) denotes the maximal (minimal) eigenvalue of H and $\rho(H)$ is the spectral radius of H. For a matrix $A = (a_1, a_2, \ldots, a_n) = (a_{ij}) \in \mathbb{R}^{n\times n}$ and a matrix B, $\mathrm{vec}(A)$ is a vector defined by $\mathrm{vec}(A) = (a_1^T, \ldots, a_n^T)^T$; $A \otimes B = (a_{ij}B)$ is a Kronecker product. For Hermitian matrices X and Y, $X \geq Y (X > Y)$ means that $X - Y$ is positive semidefinite (definite). I represents the identity matrix of size implied by context. $\|\cdot\|$ will be the spectral norm for square matrices unless otherwise noted.

2 Convergence rate

In this section, we study the convergence rate of the inversion-free variant iteration (6). Since B and R are positive semidefinite matrices, there is a matrix $E \in \mathbb{R}^{m\times q}$ and a matrix $C \in \mathbb{R}^{\ell\times m}$ such that $B = EE^T$ and $R = C^T C$. Then Eq. (2) is equivalent to the discrete-time algebraic Riccati equation

$$X = M^T XM - M^T XE(I + E^T XE)^{-1} E^T XM + C^T C. \tag{7}$$

Throughout the rest of this section, we assume that (M, E) is a stabilizable pair and (C, M) is a detectable pair. Then the DARE (7), or equivalently, Eq. (2), has a unique positive definite solution X_+ and

$$\rho((I + BX_+)^{-1}M) < 1, \tag{8}$$

see [7, 12, 22].

If both R and B are positive definite matrices, let $B = Z\Lambda Z^T$ ($Z^TZ = I$, $\Lambda = diag(\lambda_i(B))$) be a spectral decomposition, and let $R = C^TC$ where C is a matrix with independent columns. It can be easily proved that (M, Z) is a stabilizable pair and (C, M) is a detectable pair. Then Eq. (2) has a unique positive definite solution X_+ and $\rho((I + BX_+)^{-1}M) < 1$.

Lemma 2.1 ([14, p.21]) *Let T be a (nonlinear) operator from a Banach space E into itself and $x^* \in E$ be a solution of $x = Tx$. If T is Fréchet differentiable at x^* with $\rho(T'_{x^*}) < 1$, then the iterates $x_{k+1} = Tx_k$ $(k = 0, 1, \ldots)$ converge to x^*, provided that x_0 is sufficiently close to x^*. Moreover, for any $\epsilon > 0$,*

$$\|x_k - x^*\| \le c(x_0;\epsilon)(\rho(T'_{x^*}) + \epsilon)^k,$$

with $\|\cdot\|$ is the norm in E and $c(x_0;\epsilon)$ is a constant independent of k.

Theorem 2.2 *For the inversion-free variant iteration (6), we have*

$$\limsup_{k\to\infty} \sqrt[k]{\|X_k - X_+\|} \le \left(\rho((I + BX_+)^{-1}M)\right)^2.$$

Proof Define an operator T on $\mathbb{R}^{n\times n}$ by

$$T(Y) = 2Y - Y(BRB + BM^TB^{-1}MB + B)Y + YBM^TYMBY,$$

where $Y \in \mathbb{R}^{n\times n}$. Then for the inversion-free variant iteration (6), we have

$$Y_{k+1} = T(Y_k), \quad k = 0, 1, \ldots,$$

and $(BX_+B + B)^{-1}$ is a fixed point of T. Setting $Z = BRB + BM^TB^{-1}MB + B$ for convenience, we have $T(Y) = 2Y - YZY + YBM^TYMBY$, then the Fréchet derivative $T'_Y : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is

$$T'_Y(H) = 2H - HZY - YZH + HBM^TYMBY + YBM^TYMBH + YBM^THMBY.$$

Let $Y = (BX_+B + B)^{-1}$, it follows

$$\begin{aligned} T'_{(BX_+B+B)^{-1}}(H) = {} & 2H - HZ(BX_+B + B)^{-1} - (BX_+B + B)^{-1}ZH \\ & + HBM^T(BX_+B + B)^{-1}MB(BX_+B + B)^{-1} \\ & + (BX_+B + B)^{-1}BM^T(BX_+B + B)^{-1}MBH \\ & + (BX_+B + B)^{-1}BM^THMB(BX_+B + B)^{-1}. \end{aligned}$$

Since B is supposed to be nonsingular, by Woodbury identity, we obtain $(B + X_+^{-1})^{-1} = B^{-1} - (BX_+B + B)^{-1}$. Hence

$$
\begin{aligned}
&- HZ(BX_+B+B)^{-1} + HBM^T(BX_+B+B)^{-1}MB(BX_+B+B)^{-1} \\
&\quad = -H(BRB + BM^TB^{-1}MB + B)(BX_+B+B)^{-1} + HBM^T(BX_+B+B)^{-1}M(I+BX_+)^{-1} \\
&\quad = -HB(R+B^{-1})(I+BX_+)^{-1} + HBM^T((BX_+B+B)^{-1} - B^{-1})M(I+BX_+)^{-1} \\
&\quad = -HB(R+B^{-1})(I+BX_+)^{-1} - HBM^T(B+X_+^{-1})^{-1}M(I+BX_+)^{-1} \\
&\quad = -HB(B^{-1}+R+M^T(B+X_+^{-1})^{-1}M)(I+BX_+)^{-1} \\
&\quad = -HB(B^{-1}+X_+)(I+BX_+)^{-1} \\
&\quad = -H,
\end{aligned} \tag{9}
$$

where (9) is obtained by using the fact that X_+ is a solution of Eq. (2). Similarly, it can be obtained that

$$-(BX_+B+B)^{-1}ZH + (BX_+B+B)^{-1}BM^T(BX_+B+B)^{-1}MBH = -H.$$

It follows that

$$
\begin{aligned}
T'_{(BX_+B+B)^{-1}}(H) &= 2H - H - H + (BX_+B+B)^{-1}BM^THMB(BX_+B+B)^{-1} \\
&= (I+X_+B)^{-1}M^THM(I+BX_+)^{-1}.
\end{aligned}
$$

Denote the eigenvalues of $S = (I+X_+B)^{-1}M^T$ by $\lambda_i, i = 1, \dots n$, and let them be ordered such that their moduli are nonincreasing, i.e.,

$$|\lambda_1| \geq |\lambda_2| \dots \geq |\lambda_n|.$$

Since the "vec" form of the Fréchet derivative $T'_{(BX_+B+B)^{-1}}$ is $S \otimes S$, then the eigenvalues of $T'_{(BX_+B+B)^{-1}}$ are $\lambda_i\lambda_j$ for $i,j = 1, \dots, n$. Hence, we have $\rho\big(T'_{(BX_+B+B)^{-1}}\big) = \lambda_1^2$, that is, $\rho\big(T'_{(BX_+B+B)^{-1}}\big) = \big(\rho((I+BX_+)^{-1}M)\big)^2$.

According to [18, Theorem 3.6], it can be proved that the sequence $\{Y_k\}$ converges to $(BX_+B+B)^{-1}$ which is the fixed point of T. By Lemma 2.1, we have

$$\limsup_{k\to\infty} \sqrt[k]{\|Y_k - (BX_+B+B)^{-1}\|} \leq \big(\rho\big((I+BX_+)^{-1}M\big)\big)^2.$$

Moreover,

$$
\begin{aligned}
\|X_k - X_+\| &= \|R + M^T(B^{-1} - Y_k)M - R - M^T(X_+^{-1}+B)^{-1}M\| \\
&= \|M^T(B^{-1} - Y_k)M - M^TB^{-1}M + M^T(BX_+B+B)^{-1}M\| \\
&= \|M^T((BX_+B+B)^{-1} - Y_k)M\| \\
&\leq \|M\|^2\|Y_k - (BX_+B+B)^{-1}\|,
\end{aligned}
$$

which implies that

$$\limsup_{k\to\infty} \sqrt[k]{\|X_k - X_+\|} \leq \big(\rho((I+BX_+)^{-1}M)\big)^2.$$

□

It was proved in [5] that both the convergence rates of the basic fixed-point iteration (4) and the modified fixed-point iteration (5) satisfy

$$\limsup_{k\to\infty} \sqrt[k]{\|X_k - X_+\|} \le \left(\rho((I + BX_+)^{-1}M)\right)^2.$$

It can seen that if $\rho((I + BX_+)^{-1}M)$ is very close to 1, the convergence of the inversion-free variant iteration (6), the modified fixed-point iteration (5) and the fixed-point iteration (4) may be very slow. Here we say a few words about the cyclic reduction, which has a quadratic convergence rate and can be applied to Eq. (2).

Suppose B is a positive definite matrix, applying the Woodbury identity to Eq. (2) yields

$$X = R + M^T B^{-1} M - M^T (BXB + B)^{-1} M,$$

from which we get

$$B^{-1} + X = B^{-1} + R + M^T B^{-1} M - M^T (BXB + B)^{-1} M. \tag{10}$$

Let $Y = B^{-1} + X, Q = B^{-1} + R + M^T B^{-1} M, A = B^{-1} M$, then (10) can be written as

$$Y = Q - A^T Y^{-1} A. \tag{11}$$

If X_+ is the unique positive definite solution of Eq. (2), then $Y_+ = B^{-1} + X_+$ solves Eq. (11). It has been proved in [6] that if Eq. (11) has a positive definite solution, then it has a maximal one and a minimal one. We show that Y_+ is the maximal solution of Eq. (11).

Theorem 2.3 *If X_+ is the unique positive definite solution of Eq. (2), then $Y_+ = B^{-1} + X_+$ is the maximal positive definite solution of Eq. (11).*

Proof For all $|\lambda| < 1$, it holds

$$\begin{aligned} Y_+ + \lambda A &= X_+ + B^{-1} + \lambda B^{-1} M \\ &= B^{-1}(I + BX_+)(I + \lambda(I + BX_+)^{-1} M). \end{aligned}$$

Since X_+ is the unique positive definite solution of Eq. (2), it follows from (8) that $\rho((I + BX_+)^{-1}M) < 1$. Thus, $(I + \lambda(I + BX_+)^{-1}M)$ is invertible for all $|\lambda| < 1$. Hence, $Y_+ + \lambda A$ is invertible for all $|\lambda| < 1$. According to [6, Theorem 3.4], we know that Y_+ is the maximal positive definite solution of Eq. (11). □

Meini [17] showed that the cyclic reduction algorithm can be applied to find the maximal and minimal positive definite solutions of Eq. (11). Since we have proved that $Y_+ = X_+ + B^{-1}$ is the maximal solution of Eq. (11) if X_+ is the unique positive definite solution of Eq. (2), it seems that it is possible to apply cyclic reduction algorithm to Eq. (2). Indeed, if we transplant Meini's result to Eq. (11), we get

$$\begin{cases} A_0 = B^{-1}M, \\ Q_0 = Y_0 = B^{-1} + R + M^T B^{-1} M, \\ A_{k+1} = A_k Q_k^{-1} A_k, \\ Q_{k+1} = Q_k - A_k Q_k^{-1} A_k^T - A_k^T Q_k^{-1} A_k, \\ Y_{k+1} = Y_k - A_k^T Q_k^{-1} A_k. \end{cases}$$

Theorem 3.1 in [17] shows that the matrices $\{Q_k\}$ and $\{Y_k\}$ are positive definite and it holds that $0 < Q_{k+1} \le Q_k$ and $0 < Y_{k+1} \le Y_k$ for all k. Moreover, Theorem 3.2 in [17] shows that the first block entry $\{Y_k\}$ quadratically converges to $Y_+ = X_+ + B^{-1}$ if $\rho(Y_+^{-1}A) < 1$.

Since $Y_+ = B^{-1} + X_+$, the matrix sequence $\{Y_k\}$ converges to Y_+, which implies that the sequence $X_k = Y_k - B^{-1}$ converges to X_+. If we replace Y_k by $X_k + B^{-1}$ and after rearrangement, the iteration based on cyclic reduction for Eq. (2) can be stated as

$$\begin{cases} A_0 = B^{-1}M, \\ Q_0 = B^{-1} + R + M^T B^{-1} M, \\ X_0 = R + M^T B^{-1} M, \\ A_{k+1} = A_k Q_k^{-1} A_k, \\ Q_{k+1} = Q_k - A_k Q_k^{-1} A_k^T - A_k^T Q_k^{-1} A_k, \\ X_{k+1} = X_k - A_k^T Q_k^{-1} A_k. \end{cases} \tag{12}$$

Theorem 2.4 *Suppose $R > 0$ and $B > 0$, then the matrix sequence $\{X_k\}$ generated by iteration (12) converges to X_+, where X_+ is the unique positive definite solution of Eq. (2), and it holds that $0 < X_{k+1} \le X_k$ and for any $\epsilon > 0$, $\|X_k - X_+\| = O((\sigma + \epsilon)^{2\cdot 2^k})$, where $\sigma = \rho((I + BX_+)^{-1}M)$.*

Proof It has been proved by Meini [17] that $0 < Y_{k+1} \le Y_k$, which implies $0 < X_{k+1} \le X_k$. Since both R and B are positive definite matrices, then Eq. (2) has a unique positive definite solution X_+ and $\rho((I + BX_+)^{-1}M) < 1$, which gives

$$\begin{aligned} \rho(Y_+^{-1}A) &= \rho((B^{-1} + X_+)^{-1}B^{-1}M) \\ &= \rho((I + BX_+)^{-1}M) < 1, \end{aligned}$$

it follows from [17, Theorem 3.2] that $\|Y_k - Y_+\| = O((\sigma + \epsilon)^{2\cdot 2^k})$, thus $\|X_k - X_+\| = \|Y_k - Y_+\| = O((\sigma + \epsilon)^{2\cdot 2^k})$. □

Remark 2.5 Suppose $B > 0$, let $B = Z\Lambda Z^T$ ($Z^T Z = I$, $\Lambda = diag(\lambda_i(B))$) be a spectral decomposition, and let $R = C^T C$ where C is a matrix with possibly dependent columns. The condition $R > 0$ in Theorem 2.4 is not necessary if (M, Z) is a stabilizable pair and and (C, M) is a detectable pair.

3 A structured condition number

In this section, based on the classic definition of the condition number defined in Rice [20], we define a new structured condition number at X_+ of the nonlinear matrix equation (1), where X_+ is the unique positive definite solution to Eq. (1). The explicit expression of the structured condition number is obtained.

Let $M(\tau) = M + \tau E$, $B(\tau) = B + \tau G$, and $R(\tau) = R + \tau H$, where τ is a real parameter, $E, G, H \in \mathbb{R}^{n\times n}$, and G and H are symmetric matrices. We consider the equation

$$X^p = M(\tau)^T(X^{-1} + B(\tau))^{-1}M(\tau) + R(\tau).$$

Let $F(X, \tau) = X^p - M(\tau)^T(X^{-1} + B(\tau))^{-1}M(\tau) - R(\tau)$, and X_+ is the unique symmetric positive definite solution of Eq. (1). Then

1. $F(X_+, 0) = 0$;
2. $F(X, \tau)$ is differentiable arbitrarily many times in the neighborhood of $(X_+, 0)$ since X^p and X^{-1} are the polynomial or rational function of the elements of X.
3. $\frac{\partial F}{\partial X}\big|_{(X_+,0)} = \sum_{k=0}^{p-1} X_+^{p-1-k} \otimes X_+^k - (M^T \otimes M^T)\big((I + X_+B)^{-1} \otimes (I + X_+B)^{-1}\big)$.

Remark 3.1 The introduction of $\partial F/\partial X$ is similar to [21]. According to [21], we know that $\frac{\partial AXB}{\partial X} = B^T \otimes A$, $\frac{\partial X^{-1}}{\partial X} = -(X^{-1})^T \otimes X^{-1}$, it follows from the chain rule that

$$\begin{aligned}\frac{\partial F(X)}{\partial X} &= \sum_{k=0}^{p-1} X^{p-1-k} \otimes X^k \\ &\quad - (M(\tau)^T \otimes M(\tau)^T)\big((X^{-1} + B(\tau))^{-1} \otimes (X^{-1} + B(\tau))^{-1}\big)(X^{-1} \otimes X^{-1}) \\ &= \sum_{k=0}^{p-1} X^{p-1-k} \otimes X^k - (M(\tau)^T \otimes M(\tau)^T)\big((I + XB(\tau))^{-1} \otimes (I + XB(\tau))^{-1}\big)\end{aligned}$$

Since $M(0) = M, B(0) = B$, we have

$$\frac{\partial F}{\partial X}\Big|_{(X_+,0)} = \sum_{k=0}^{p-1} X_+^{p-1-k} \otimes X_+^k - (M^T \otimes M^T)\big((I + X_+B)^{-1} \otimes (I + X_+B)^{-1}\big).$$

$\frac{\partial F}{\partial X}\big|_{(X_+,0)}$ is invertible under certain conditions. For example, if $p \geq 2$ and at least one of R and M is nonsingular, it follows from Theorem 2.3 in [18] that $X_+ \geq \alpha_2 I$, where α_2 is the unique positive root of function $g_2(x) = x^p - \lambda_{\min}(M^TM)(\lambda_{\min}(B) + x^{-1})^{-1} - \lambda_{\min}(R)$. Suppose $\|M\|^2 < p\alpha_2^{p-1}$, then by Theorem 3.3.16 in Horn and Johnson [9] we get

$$\begin{aligned}\sigma_{\min}\Big(\frac{\partial F}{\partial X}\big|_{(X_+,0)}\Big) &\geq \lambda_{\min}\Big(\sum_{k=0}^{p-1} X_+^{p-1-k}\otimes X_+^k\Big)\\ &\quad - \Big\|(M^T\otimes M^T)\big((I+X_+B)^{-1}\otimes(I+X_+B)^{-1}\big)\Big\|\\ &\geq p\alpha_2^{p-1} - \|M\|^2 > 0,\end{aligned} \tag{13}$$

which implies that $\det\Big(\frac{\partial F}{\partial X}\big|_{(X_+,0)}\Big) \neq 0$.

In what follows, we suppose that $\frac{\partial F}{\partial X}\big|_{(X_+,0)}$ is invertible. Then, according to implicit function theory [19], there is $\delta > 0$ such that if $\tau \in (-\delta, \delta)$, there is a unique $X(\tau)$ satisfying

1. $F(X(\tau), \tau) = 0, X(0) = X_+$.
2. $X(\tau)$ is differentiable arbitrarily many times with respect to τ.

For

$$X(\tau)^p = M(\tau)^T(X(\tau)^{-1} + B(\tau))^{-1}M(\tau) + R(\tau), \tag{14}$$

by taking derivative for both sides of (14) with respect to τ at $\tau = 0$, we arrive at

$$\begin{aligned}&\sum_{k=0}^{p-1} X_+^k\dot{X}(0)X_+^{p-1-k} - M^T(I+X_+B)^{-1}\dot{X}(0)(I+BX_+)^{-1}M\\ &\quad = H + E^T(X_+^{-1}+B)^{-1}M + M^T(X_+^{-1}+B)^{-1}E - M^T(X_+^{-1}+B)^{-1}G(X_+^{-1}+B)^{-1}M.\end{aligned}$$

Let

$$\begin{aligned}J_1 &= \sum_{k=0}^{p-1} X_+^{p-1-k}\otimes X_+^k,\\ J_2 &= (M^T\otimes M^T)((I+BX_+)^{-T}\otimes(I+X_+B)^{-1}),\\ L_1 &= I\otimes(M^T(X_+^{-1}+B)^{-1}) + (((X_+^{-1}+B)^{-1}M)^T\otimes I)\Pi,\\ L_2 &= -((X_+^{-1}+B)^{-1}M)^T\otimes(M^T(X_+^{-1}+B)^{-1}),\\ L &= [L_1 \quad L_2 \quad I_{n^2}],\\ v &= (\mathrm{vec}(E)^T, \quad \mathrm{vec}(G)^T, \quad \mathrm{vec}(H)^T)^T,\end{aligned}$$

where $\Pi \in \mathbb{R}^{n^2\times n^2}$ is the vec permutation satisfying $\Pi\,\mathrm{vec}(E) = \mathrm{vec}(E^T)$. Then we have

$$(J_1 - J_2)\mathrm{vec}(\dot{X}(0)) = L\begin{pmatrix}\mathrm{vec}(E)\\ \mathrm{vec}(G)\\ \mathrm{vec}(H)\end{pmatrix} = Lv.$$

Note that $J_1 - J_2$ is invertible since $J_1 - J_2 = \frac{\partial F}{\partial X}\big|_{(X_+,0)}$ and $\frac{\partial F}{\partial X}\big|_{(X_+,0)}$ is assumed to be invertible. Based on the classic definition of the condition number defined in Rice

[20], we define a condition number at the unique positive definite solution X_+ of Eq. (1):

$$\begin{aligned}\mathscr{K}_{X_+} &= \lim_{\tau\to 0^+} \sup_{E\in\mathbb{R}^{n\times n},G,H\in\mathbb{S}^{n\times n}} \left\{ \frac{\|X(\tau)-X_+\|_F}{\|X_+\|_F} \Big/ \left(\frac{\tau\|(E,G,H)\|_F}{\|(M,B,R)\|_F} \right) \right\} \\ &= \max_{E\in\mathbb{R}^{n\times n},G,H\in\mathbb{S}^{n\times n}} \left\{ \frac{\|\dot{X}(0)\|_F}{\|(E,G,H)\|_F} \cdot \frac{\|(M,B,R)\|_F}{\|X_+\|_F} \right\}.\end{aligned} \tag{15}$$

To obtain the explicit expression of the newly defined structured condition number (15), we first define a linear operator $\mathcal{W}$ on a matrix $Z \in \mathbb{C}^{n\times n}$ with a positive semi-idefinite matrix S and a positive integer p as

$$\mathcal{W}(Z,S,p) = \sum_{k=0}^{p-1} S^k Z S^{p-1-k}.$$

Define another linear operator $\mathcal{L} : \mathbb{S}^{n\times n} \to \mathbb{S}^{n\times n}$ as

$$\mathcal{L}(Z) = \mathcal{W}(Z, X_+, p) - M^T(I + X_+B)^{-1}Z(I + BX_+)^{-1}M$$

with $Z \in \mathbb{S}^{n\times n}$.

Since $J_1 - J_2$ is invertible, it implies that $\mathcal{L}$ is invertible. Moreover, we define the operator $\mathcal{P} : \mathbb{R}^{n\times n} \times \mathbb{S}^{n\times n} \times \mathbb{S}^{n\times n} \to \mathbb{S}^{n\times n}$ as

$$\begin{aligned}\mathcal{P}(E,G,H) = \mathcal{L}^{-1}\big(&H + E^T(X_+^{-1}+B)^{-1}M + M^T(X_+^{-1}+B)^{-1}E \\ &- M^T(X_+^{-1}+B)^{-1}G(X_+^{-1}+B)^{-1}M\big)\end{aligned}$$

with $E \in \mathbb{R}^{n\times n}$ and $G, H \in \mathbb{S}^{n\times n}$.

Taking derivative for both sides of (14) with respect to τ, and then letting $\tau = 0$, we arrive at

$$\begin{aligned}\mathcal{L}(\dot{X}(0)) = &H + E^T(X_+^{-1}+B)^{-1}M + M^T(X_+^{-1}+B)^{-1}E \\ &- M^T(X_+^{-1}+B)^{-1}G(X_+^{-1}+B)^{-1}M.\end{aligned}$$

This yields

$$\dot{X}(0) = \mathcal{P}(E,G,H). \tag{16}$$

Define the operator $\mathcal{D} : \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ as

$$\begin{aligned}\mathcal{D}(W,P,Q) = \mathcal{T}^{-1}\big(&Q + W^T(X_+^{-1}+B)^{-1}M + M^T(X_+^{-1}+B)^{-1}W \\ &- M^T(X_+^{-1}+B)^{-1}P(X_+^{-1}+B)^{-1}M\big),\end{aligned} \tag{17}$$

where $W, P, Q \in \mathbb{R}^{n\times n}$ and $\mathcal{T} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is an invertible linear operator defined by

$$\mathcal{T}(N) = \mathcal{W}(N, X_+, p) - M^T(I + X_+B)^{-1}N(I + BX_+)^{-1}M$$

with $N \in \mathbb{R}^{n\times n}$.

Theorem 3.2 *For matrices $W, P, Q \in \mathbb{R}^{n\times n}$, there exist real matrices $\hat{W}, \hat{P}, \hat{Q} \in \mathbb{R}^{n\times n}$ with $(\hat{W}, \hat{P}, \hat{Q}) \neq 0$ such that*

$$\max_{\substack{W,P,Q\in\mathbb{R}^{n\times n}\\(W,P,Q)\neq 0}} \frac{\|\mathcal{D}(W,P,Q)\|_F}{\|(W,P,Q)\|_F} = \frac{\|\mathcal{D}(\hat{W},\hat{P},\hat{Q})\|_F}{\|(\hat{W},\hat{P},\hat{Q})\|_F}, \tag{18}$$

where $\hat{P}^T = \hat{P}$ or $\hat{P}^T = -\hat{P}$ and $\hat{Q}^T = \hat{Q}$ or $\hat{Q} = -\hat{Q}^T$.

Proof Let (W, P, Q) be a singular "vector" of $\mathcal{D}$ corresponding to the largest singular value, then the maximum in (18) occurs since the operator $\mathcal{D}$ is linear. Note that

$$\|\mathcal{D}(W,P,Q)\|_F = \|\mathcal{D}^T(W,P,Q)\|_F = \|\mathcal{D}(W,P^T,Q^T)\|_F,$$

then (W, P^T, Q^T) is also a singular "vector" of $\mathcal{D}$ corresponding to the largest singular value. If $P = -P^T, Q = -Q^T$, then $\hat{W} = W, \hat{P} = P$ and $\hat{Q} = Q$ are the real matrices satisfying (18). If $P = -P^T, Q \neq -Q^T$, then $\hat{W} = W, \hat{P} = P$ and $\hat{Q} = \frac{Q+Q^T}{2}$ are the kind of matrices that satisfy (18). If $P \neq -P^T, Q = -Q^T$, then there exist $\hat{W} = W, \hat{P} = \frac{P+P^T}{2}$ and $\hat{Q} = Q$ satisfying (18). If $P \neq -P^T, Q \neq -Q^T$, then $\hat{W} = W$, $\hat{P} = \frac{P+P^T}{2}$ and $\hat{Q} = \frac{Q+Q^T}{2}$ are the matrices satisfying (18). □

Lemma 3.3 [4] *If two Hermitian matrices $W_1, W_2 \in \mathbb{C}^{n\times n}$ satisfy $W_2 \geq W_1 \geq -W_2$, then $\|W_1\|_F \leq \|W_2\|_F$.*

We say the matrices M and B are (D, p)-stable if $\mathcal{W}(Z, D, p) - M^T(I + D^T B)^{-1}Z(I + BD)^{-1}M \geq 0$ implies $Z \geq 0$.

Theorem 3.4 *Under the condition that $J_1 - J_2$ is invertible and matrices M and B are (X_+, p)-stable. The structured condition number at X_+ of Eq. (1), where X_+ is the unique positive definite solution to Eq. (1), is*

$$\mathcal{K}_{X_+} = \|(J_1 - J_2)^{-1}L\| \frac{\|(M,B,R)\|_F}{\|X_+\|_F}. \tag{19}$$

Proof It follows from (17) that

$$\max_{\substack{W,P,Q\in\mathbb{R}^{n\times n}\\(W,P,Q)\neq 0}} \frac{\|\mathcal{D}(W,P,Q)\|_F}{\|(W,P,Q)\|_F} = \|(J_1 - J_2)^{-1}L\|. \tag{20}$$

Then, according to (15) and (16), it suffices to prove

$$\max_{\substack{E\in\mathbb{R}^{n\times n},G,H\in\mathbb{S}^{n\times n}\\(E,G,H)\neq 0}} \frac{\|\mathcal{P}(E,G,H)\|_F}{\|(E,G,H)\|_F} = \max_{\substack{W,P,Q\in\mathbb{R}^{n\times n}\\(W,P,Q)\neq 0}} \frac{\|\mathcal{D}(W,P,Q)\|_F}{\|(W,P,Q)\|_F}. \tag{21}$$

It is clear that

$$\max_{\substack{E\in\mathbb{R}^{n\times n},G,H\in\mathbb{S}^{n\times n}\\(E,G,H)\neq 0}} \frac{\|\mathcal{P}(E,G,H)\|_F}{\|(E,G,H)\|_F} \le \max_{\substack{W,P,Q\in\mathbb{R}^{n\times n}\\(W,P,Q)\neq 0}} \frac{\|\mathcal{D}(W,P,Q)\|_F}{\|(W,P,Q)\|_F}. \tag{22}$$

What left to prove is

$$\max_{\substack{W,P,Q\in\mathbb{R}^{n\times n}\\(W,P,Q)\neq 0}} \frac{\|\mathcal{D}(W,P,Q)\|_F}{\|(W,P,Q)\|_F} \le \max_{\substack{E\in\mathbb{R}^{n\times n},G,H\in\mathbb{S}^{n\times n}\\(E,G,H)\neq 0}} \frac{\|\mathcal{P}(E,G,H)\|_F}{\|(E,G,H)\|_F}.$$

By Theorem 3.2, there exists $(\hat{W},\hat{P},\hat{Q})\neq 0$, where $\hat{W},\hat{P},\hat{Q}\in\mathbb{R}^{n\times n}$, $\hat{P}$ and $\hat{Q}$ are either symmetric or skew-symmetric matrices, such that

$$\frac{\|\mathcal{D}(\hat{W},\hat{P},\hat{Q})\|_F}{\|(\hat{W},\hat{P},\hat{Q})\|_F} = \max_{\substack{W,P,Q\in\mathbb{R}^{n\times n}\\(W,P,Q)\neq 0}} \frac{\|\mathcal{D}(W,P,Q)\|_F}{\|(W,P,Q)\|_F}.$$

There are four cases according to the symmetry or anti-symmetry property of matrices $\hat{P}$ and $\hat{Q}$.

Case 1: $\hat{P}=\hat{P}^T$ and $\hat{Q}=\hat{Q}^T$. Let $\hat{E}=\hat{W}$, $\hat{G}=\hat{P}$ and $\hat{H}=\hat{Q}$, it yields

$$\frac{\|\mathcal{P}(\hat{E},\hat{G},\hat{H})\|_F}{\|(\hat{E},\hat{G},\hat{H})\|_F} = \max_{\substack{W,P,Q\in\mathbb{R}^{n\times n}\\(W,P,Q)\neq 0}} \frac{\|\mathcal{D}(W,P,Q)\|_F}{\|(W,P,Q)\|_F}. \tag{23}$$

Case 2: $\hat{P}^T=-\hat{P}\neq 0$ and $\hat{Q}^T=\hat{Q}$. In this case, both $(\hat{W},\hat{P},\hat{Q})$ and $(\hat{W},-\hat{P},\hat{Q})$ satisfy (18). As a linear combination of them, $(0,\hat{P},0)\neq 0$ also satisfies (18).

It is clear that there is a real orthogonal matrix U and $p_i>0$ $(i=1,2,\ldots,k)$ such that

$$\hat{P}=U\left[\begin{pmatrix}0 & p_1\\ -p_1 & 0\end{pmatrix}\oplus\cdots\oplus\begin{pmatrix}0 & p_k\\ -p_k & 0\end{pmatrix}\oplus 0\right]U^T. \tag{24}$$

Let $\hat{E}=0,\hat{H}=0$ and

$$\hat{G}=U\left[\begin{pmatrix}-p_1 & 0\\ 0 & -p_1\end{pmatrix}\oplus\cdots\oplus\begin{pmatrix}-p_k & 0\\ 0 & -p_k\end{pmatrix}\oplus 0\right]U^T, \tag{25}$$

we have

$$\hat{G}\in\mathbb{S}^{n\times n},\ \|\hat{G}\|_F=\|\hat{P}\|_F,\ \hat{G}\le 0,\ -\hat{G}\ge i\hat{P}\ge\hat{G}.$$

Let $\hat{Z}=\mathcal{D}(0,\hat{P},0)\in\mathbb{R}^{n\times n}$ and $\tilde{Z}=\mathcal{P}(0,\hat{G},0)\in\mathbb{S}^{n\times n}$, then

$$\mathcal{L}(\tilde{Z})=-M^T(X_+^{-1}+B)^{-1}\hat{G}(X_+^{-1}+B)^{-1}M$$

and

$$\mathcal{T}(i\hat{Z})=-M^T(X_+^{-1}+B)^{-1}i\hat{P}(X_+^{-1}+B)^{-1}M.$$

It follows that

$$\mathcal{L}(\tilde{Z}) + \mathcal{T}(i\hat{Z}) = -M^T(X_+^{-1} + B)^{-1}(\hat{G} + i\hat{P})(X_+^{-1} + B)^{-1}M \geq 0 \tag{26}$$

and

$$\mathcal{L}(\tilde{Z}) - \mathcal{T}(i\hat{Z}) = -M^T(X_+^{-1} + B)^{-1}(\hat{G} - i\hat{P})(X_+^{-1} + B)^{-1}M \geq 0. \tag{27}$$

Since M and B are (X_+, p)-stable, we have from (26) and (27) that $\tilde{Z} + i\hat{Z} \geq 0$ and $\tilde{Z} - i\hat{Z} \geq 0$, which leads to

$$\tilde{Z} \geq i\hat{Z} \geq -\tilde{Z}.$$

As a direct consequence of Lemma 3.3, we have $\|i\hat{Z}\|_F \leq \|\tilde{Z}\|_F$.

Then,

$$\begin{aligned}\frac{\|\mathcal{D}(0,\hat{P},0)\|_F}{\|(0,\hat{P},0)\|_F} &= \frac{\|\hat{Z}\|_F}{\|(0,\hat{P},0)\|_F} = \frac{\|i\hat{Z}\|_F}{\|(0,i\hat{P},0)\|_F} \\ &\leq \frac{\|\tilde{Z}\|_F}{\|(0,\hat{G},0)\|_F} = \frac{\|\mathcal{P}(0,\hat{G},0)\|_F}{\|(0,\hat{G},0)\|_F}.\end{aligned}$$

Now a symmetric matrix $\hat{G} \in \mathbb{S}^{n\times n}$ is found and satisfies

$$\max_{\substack{W,P,Q\in\mathbb{R}^{n\times n}\\(W,P,Q)\neq 0}} \frac{\|\mathcal{D}(W,P,Q)\|_F}{\|(W,P,Q)\|_F} = \frac{\|\mathcal{D}(0,\hat{P},0)\|_F}{\|(0,\hat{P},0)\|_F} \leq \frac{\|\mathcal{P}(0,\hat{G},0)\|_F}{\|(0,\hat{G},0)\|_F}. \tag{28}$$

Case 3: $\hat{P}^T = \hat{P}$ and $\hat{Q}^T = -\hat{Q} \neq 0$. In this case, $(0,0,\hat{Q})$ satisfies (18). It is clear that there is a real orthogonal matrix V and $q_i > 0$ $(i = 1, 2, \ldots, \ell)$ such that

$$\hat{Q} = V\left[\begin{pmatrix} 0 & q_1 \\ -q_1 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & q_\ell \\ -q_\ell & 0 \end{pmatrix} \oplus 0\right]V^T. \tag{29}$$

Let

$$\hat{H} = V\left[\begin{pmatrix} q_1 & 0 \\ 0 & q_1 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} q_\ell & 0 \\ 0 & q_\ell \end{pmatrix} \oplus 0\right]V^T, \tag{30}$$

then we have

$$\hat{H} \in \mathbb{S}^{n\times n},\ \|\hat{H}\|_F = \|\hat{Q}\|_F,\ \hat{H} \geq 0,\ \hat{H} \geq i\hat{Q} \geq -\hat{H}.$$

Let $\hat{Z} = \mathcal{D}(0,0,\hat{Q}) \in \mathbb{R}^{n\times n}$ and $\tilde{Z} = \mathcal{P}(0,0,\hat{H}) \in \mathbb{S}^{n\times n}$, it yields

$$\mathcal{L}(\tilde{Z}) + \mathcal{T}(i\hat{Z}) = \hat{H} + i\hat{Q} \geq 0 \tag{31}$$

and

$$\mathcal{L}(\tilde{Z}) - \mathcal{T}(i\hat{Z}) = \hat{H} - i\hat{Q} \geq 0 \tag{32}$$

Since M and B are (X_+, p)-stable, we have from (31) and (32) that $\tilde{Z} + i\hat{Z} \geq 0$ and $\tilde{Z} - i\hat{Z} \geq 0$, which leads to

$$\tilde{Z} \geq i\hat{Z} \geq -\tilde{Z}.$$

Analogously to the last part in Case 2, we find a symmetric matrix $\hat{H} \in \mathbb{S}^{n\times n}$ such that

$$\max_{\substack{W,P,Q\in\mathbb{R}^{n\times n}\\(W,P,Q)\neq 0}} \frac{\|\mathcal{D}(W,P,Q)\|_F}{\|(W,P,Q)\|_F} = \frac{\|\mathcal{D}(0,0,\hat{Q})\|_F}{\|(0,0,\hat{Q})\|_F} \leq \frac{\|\mathcal{P}(0,0,\hat{H})\|_F}{\|(0,0,\hat{H})\|_F}. \tag{33}$$

Case 4: $\hat{P}^T = -\hat{P} \neq 0$ and $\hat{Q}^T = -\hat{Q} \neq 0$. In this case, $(0, \hat{P}, \hat{Q})$ satisfies (18). Obviously, $\hat{P}$ has a form as (24) and $\hat{Q}$ has a form as (29). Choose $\hat{G}$ and $\hat{H}$ to have the form (25) and (30), respectively.

Let $\hat{Z} = \mathcal{D}(0, \hat{P}, \hat{Q}) \in \mathbb{R}^{n\times n}$ and $\tilde{Z} = \mathcal{P}(0, \hat{G}, \hat{H}) \in \mathbb{S}^{n\times n}$, it yields

$$\mathcal{L}(\tilde{Z}) + \mathcal{T}(i\hat{Z}) = \hat{H} + i\hat{Q} - M^T(X_+^{-1} + B)^{-1}(\hat{G} + i\hat{P})(X_+^{-1} + B)^{-1}M \geq 0 \tag{34}$$

and

$$\mathcal{L}(\tilde{Z}) - \mathcal{T}(i\hat{Z}) = \hat{H} - i\hat{Q} - M^T(X_+^{-1} + B)^{-1}(\hat{G} - i\hat{P})(X_+^{-1} + B)^{-1}M \geq 0. \tag{35}$$

Analogously to the last part in Case 2 again, we now find the symmetric matrices $\hat{G}, \hat{H} \in \mathbb{S}^{n\times n}$ such that

$$\max_{\substack{W,P,Q\in\mathbb{R}^{n\times n}\\(W,P,Q)\neq 0}} \frac{\|\mathcal{D}(W,P,Q)\|_F}{\|(W,P,Q)\|_F} = \frac{\|\mathcal{D}(0,\hat{P},\hat{Q})\|_F}{\|(0,\hat{P},\hat{Q})\|_F} \leq \frac{\|\mathcal{P}(0,\hat{G},\hat{H})\|_F}{\|(0,\hat{G},\hat{H})\|_F}. \tag{36}$$

From (23), (28), (33) and (36), we can conclude that

$$\max_{\substack{W,P,Q\in\mathbb{R}^{n\times n}\\(W,P,Q)\neq 0}} \frac{\|\mathcal{D}(W,P,Q)\|_F}{\|(W,P,Q)\|_F} \leq \max_{\substack{E\in\mathbb{R}^{n\times n},G,H\in\mathbb{S}^{n\times n}\\(E,G,H)\neq 0}} \frac{\|\mathcal{P}(E,G,H)\|_F}{\|(E,G,H)\|_F},$$

which, together with (22), leads to (21). It follows from (15), (16), (20) and (21) that (19) holds. □

The condition that $J_1 - J_2$ is invertible in Theorem 3.4 highly relies on X_+. Since the value of X_+ may not be trivial in most cases, a sufficient condition to determine whether $J_1 - J_2$ is invertible is necessary. As we have mentioned, for $p \geq 2$, at least one of R and M is nonsingular and $\|M\|^2 < p\alpha_2^{p-1}$, it can be proved analogously to (13) that $J_1 - J_2$ is invertible. We thus get the following corollary.

Corollary 3.5 *Suppose $p \geq 2$ is a positive integer, at least one of R and M is nonsingular, $\|M\|^2 < p\alpha_2^{p-1}$ and matrices M and B are (X_+, p)-stable. The structured condition number at the unique positive definite solution X_+ of Eq. (1) is*

$$\mathcal{K}_{X_+} = \|(J_1 - J_2)^{-1}L\| \frac{\|(M,B,R)\|_F}{\|X_+\|_F}.$$

For $p = 1$, if X_+ is the unique positive definite solution of the DARE (2), then $\rho((I + BX_+)^{-1}M) < 1$. It follows that

$$J_1 - J_2 = I - ((I + BX_+)^{-1}M)^T \otimes (M^T(I + X_+B)^{-1})$$

is invertible. Moreover, since $\rho((I + BX_+)^{-1}M) < 1$, for a Hermitian matrix Z, $Z - M^T(I + X_+B)^{-1}Z(I + BX_+)^{-1}M \geq 0$ implies $Z \geq 0$. We have the following theorem.

Theorem 3.6 *If $p = 1$, X_+ is the unique positive definite solution of the DARE (2), the structured condition number at X_+ is*

$$\mathcal{K}_{X_+} = \|(I - J)^{-1}L\| \frac{\|(M, B, R)\|_F}{\|X_+\|_F},$$

where $J = (M^T \otimes M^T)((I + BX_+)^{-T} \otimes (I + X_+B)^{-1})$.

4 Numerical examples

In this paper, we apply the iteration method proposed in [18] for finding the minimal nonnegative solution of Eq. (1) and of the perturbed equation. Our experiments were done in MATLAB R2017b with machine precision around 10^{-16} and the iterations terminate if the relative residual $\rho_{res}(X_k)$ satisfies

$$\rho_{res}(X_k) = \frac{\|fl(X_k^p - R - M^T(B + X_k^{-1})^{-1}M)\|_F}{\|X_k^p\|_F + \|R\|_F + \|M^T\|_F\|(B + X_k^{-1})^{-1}\|_F\|M\|_F} \leq 10^{-15}.$$

Example 4.1 We consider the matrix equation (2) with the elements of matrix M being defined by the following scheme:

(I) For a real $0 < \alpha < 1$.
(II) For $i = 1, \ldots, n$:

(a) for $j = i, \ldots, n$, set $m_{ij} = i^2 + 1/j$;
(b) compute $s_1 = \sum_{j=1}^{i-1} m_{ij}$, $s_2 = \sum_{j=i}^{n} m_{ij}$;
(c) for $j = i, \ldots, n$, set

$$m_{ij} = m_{ij}\frac{1 - \alpha - s_1}{s_2}, \quad m_{ji} = m_{ij}.$$

$$\text{Let } R = I_n, B = \begin{pmatrix} 3 & -1 & & & \\ -1 & 3 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 3 & -1 \\ & & & -1 & 3 \end{pmatrix}_{n\times n}.$$

Table 1 Comparison of iterations and relative errors

α	CR (12)		IF-FIP (6)		M-FIP (5)		FIP (4)	
	Iter.	Err.	Iter.	Err.	Iter.	Err.	Iter.	Err.
n = 10								
0.001	4	1.0578e−15	15	2.1610e−15	15	7.2910e−15	15	2.7712e−15
0.1	4	2.2456e−16	14	3.8373e−15	15	1.2272e−15	14	4.3197e−15
0.3	4	4.8185e−17	12	4.0445e−15	12	7.8957e−15	12	3.5088e−15
0.5	4	8.3634e−17	10	1.5460e−15	10	2.2760e−15	10	1.2071e−15
0.7	3	4.0194e−15	7	8.2267e−15	8	1.9330e−16	7	5.1789e−15
0.9	3	6.9783e−17	5	7.0066e−17	5	7.2011e−17	5	9.4514e−16
n = 30								
0.001	4	2.6982e−15	15	5.3810e−15	16	2.5137e−15	15	7.0611e−15
0.1	4	5.7222e−16	14	9.0921e−15	15	3.1650e−15	15	1.5407e−15
0.3	4	7.4828e−17	12	8.4578e−15	13	1.4409e−15	12	7.0378e−15
0.5	4	5.6717e−17	10	2.7399e−15	10	3.8326e−15	10	2.2372e−15
0.7	3	5.3859e−15	8	2.2548e−16	8	2.5759e−16	7	6.0986e−15
0.9	3	5.7761e−17	5	7.5399e−17	5	7.6715e−17	5	1.4611e−15
n = 60								
0.001	4	2.7234e−15	15	5.4028e−15	16	2.6327e−15	15	7.1580e−15
0.1	4	5.8673e−16	14	9.0370e−15	15	3.2381e−15	15	1.9351e−15
0.3	4	7.7674e−17	12	8.1529e−15	13	1.4267e−15	12	6.7812e−15
0.5	4	5.9180e−17	10	2.5320e−15	10	3.5201e−15	10	2.0553e−15
0.7	3	4.6619e−15	7	9.4303e−15	8	2.2783e−16	7	5.1544e−15
0.9	3	6.4121e−17	5	6.2699e−17	5	6.3447e−17	5	1.0775e−15

Since both R and B are positive definite matrices, the DARE (2) has a unique positive definite solution X_+ and $\rho((I + BX_+)^{-1}M) < 1$. This implies that iteration (12) converges quadratically.

For different values of the matrix size n and parameter α, we apply the four iterations to Eq. (2). Table 1 reports the number of iterations and the residual errors. From Table 1 we can see that the iteration based on cyclic reduction for Eq. (2) uses much less iterations for obtaining the unique positive definite solution.

Example 4.2 This is an example for the DARE (2) and is taken from [2]. The coefficient matrices are constructed as follows. Let $M_0 = diag(0, 1, 3)$, $V = I - \frac{2}{3}vv^T$, $v^T = [1, 1, 1]$. Then

$$M = VM_0V,\ B = \epsilon I,\ R = \epsilon I,$$

where ϵ is a real parameter.

Suppose the coefficients are perturbed, in MATLAB commands, as follows.

Table 2 Comparison of the perturbation bound and relative error

j	$\frac{\|\tilde{X}_+ - X_+\|_F}{\|X_+\|_F}$	$\mathcal{K}_{X_+}\Delta$
p = 1		
2	1.2840e−03	4.3523e−03
4	3.5654e−06	1.2435e−05
6	8.7542e−08	2.6473e−07
p = 3		
2	1.2507e−03	3.9725e−02
4	1.2660e−05	3.8459e−04
6	9.0601e−08	2.2824e−06
p = 5		
2	2.5651e−04	2.1810e−02
4	2.7971e−06	2.7595e−04
6	5.8162e−08	4.8110e−06

$$\begin{aligned}\tau &= rand(1) * 10^{-j};\\ \tilde{M} &= M + \tau * rand(3);\\ \tilde{R} &= R + \tau * eye(3);\\ \tilde{B} &= B + \tau * eye(3),\end{aligned}$$

Set $\Delta M = \tilde{M} - M$, $\Delta R = \tilde{R} - R$ and $\Delta B = \tilde{B} - B$. Using the structured condition number (19), we obtain a perturbation bound $\mathcal{K}_{X_+}\Delta$, where $\Delta = \frac{\|(\Delta M, \Delta R, \Delta B)\|_F}{\|(M,R,B)\|_F}$. We compute the relative error $\frac{\|\tilde{X}_+ - X_+\|_F}{\|X_+\|_F}$. Let $\epsilon = 2$, for $j = 2, 4, 6$ and $p = 1, 3, 5$, we compare the computed perturbation bound with the relative error. The results are shown in Table 2.

For this example, Table 2 shows that the estimated perturbation bound is very close to the exact relative perturbation errors, which implies that the structured condition number successfully indicated the sensitivity of the minimal positive definite solution X_+ to the perturbations in coefficients.

5 Conclusion

In this paper, the convergence behaviour of an already existing iterative method is studied. For the general case $p \geq 1$, we define a structured condition number at the unique positive definite solution X_+ of Eq. (1), which is validated by numerical examples that the newly proposed structured condition number measures the sensitivity of the solution well.

Acknowledgements This work was partially supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A3B04033516), the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) (NRF-2017R1A5A1015722) and the Basic Science Research Program through the National Research Foundation of Korea (NRF-2019R1I1A1A01062548) funded by the

Ministry of Education, and the National Natural Science Foundation of China under grant No.11961048, NSF of Jiangxi Province with No.20181ACB20001. The authors thank the anonymous referees for providing very useful suggestions for improving this paper.

Compliance with ethical standards

Conflict of interest No potential conflict of interest was reported by the authors.

References

1. Anderson, W.N., Kleindorfer, G.D., Kleindorfer, P.R., Woodroofe, M.B.: Consistent estimates of the parameters of a linear system. Ann. Math. Stat. **40**, 2064–2075 (1969)
2. Benner, P., Laub, A.J., Mehrmann, V.: A collection of benchmark examples for the numerical solution of algebraic Riccati equations II: Discrete-time case, Tech. Report SPC 95-23, Fak. f. Mathematik, TU Chemnitz Zwickau, 09107 Chemnitz, FRG (1995)
3. Bougerol, Ph: Kalman filtering with random coefficients and contractions. SIAM J. Control Optim. **31**, 942–959 (1993)
4. Byers, R., Nash, S.: On the singular vectors of the Lyapunov operator. SIAM J. Algebraic Discrete Methods **8**, 59–66 (1987)
5. Dai, H., Bai, Z.-Z.: On eigenvalue bounds and iteration methods for discrete algebraic Riccati equations. J. Comput. Math. **29**, 341–366 (2011)
6. Engwerda, J.C., Ran, A.C.M., Rijkeboer, A.L.: Necessary and sufficient conditions for the existence of a positive definite solution of the matrix equation $X + A^*X^{-1}A = Q$. Linear Algebra Appl. **186**, 255–275 (1993)
7. Gudmundsson, T., Kemmey, C., Laub, A.J.: Scaling of the discrete-time algebraic Riccati equation to enhance stability of the Schur solution method. IEEE Trans. Autom. Control **37**, 513–518 (1992)
8. Guo, C.-H., Lancaster, P.: Iterative solution of two matrix equations. Math. Comput. **68**, 1589–1603 (1999)
9. Horn, R.A., Johnson, C.R.: Topics in Matrix Analysis. Cambridge University Press, Cambridge (1991)
10. Jung, C., Kim, H.-M., Lim, Y.: On the solution of the nonlinear matrix equation $X^n = f(X)$. Linear Algebra Appl. **430**, 2042–2052 (2009)
11. Komaroff, N.: Iterative matrix bounds and computational solutions to the discrete algebraic Riccati equation. IEEE Trans. Autom. Control **39**, 1676–1678 (1994)
12. Konstantinov, M., Petkov, P., Christov, N.: Perturbation analysis of the discrete Riccati equation. Kybernetika **29**, 18–29 (1993)
13. Kouikoglou, V.S., Phillis, Y.A.: Trace bounds on the covariances of continuous-time systems with multiplicative noise. IEEE Trans. Autom. Control **38**, 138–142 (1993)
14. Krasnosel'skii, M.A., Vainikko, G.M., Zabreiko, P.P., Ruticki, YaB, Stet'senko, V.Ya.: Approximate Solution of Operator Equations. Wolters-Noordhoff Publishing, Gronongen (1972)
15. Lancaster, P., Rodman, L.: Algebraic Riccati Equations. Oxford Science Publications, Oxford University Press, Oxford (1995)
16. Lee, C.-H.: Simple stabilizability criteria and memoryless state feedback control design for time-delay systems with time-varying perturbations. IEEE Trans. Circuits Syst. I Fundam. Theory Appl. **45**, 1211–1215 (1998)
17. Meini, B.: Efficient computation of the extreme solutions of $X + A^*X^{-1}A = Q$ and $X - A^*X^{-1}A = Q$. Math. Comput. **71**, 1189–1204 (2002)
18. Meng, J., Kim, H.-M.: The positive definite solution of the nonlinear matrix equation $X^p = A + M(B + X^{-1})^{-1}M^*$. J. Comput. Appl. Math. **322**, 139–147 (2017)
19. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equation in Several Variables. Academic Press, New York (1970)
20. Rice, J.R.: A theory of condition. SIAM J. Numer. Anal. **3**, 287–310 (1996)
21. Sun, J.-G.: Eigenvalues and eigenvectors of a matrix dependent on several parameters. J. Comput. Math. **3**, 351–364 (1985)

22. Wonham, W.M.: Linear Multivariable Control: A Geometric Approach, 2nd edn. Springer, New York (1979)
23. Wang, S.-S., Chen, B.-S., Lin, T.-P.: Robust stability of uncertain time-delay systems. Int. J. Control **46**, 963–976 (1987)