Numer. Math. (2012) 120:671–700
DOI 10.1007/s00211-011-0421-0

Numerische Mathematik

Accurate solutions of M-matrix algebraic Riccati equations

Jungong Xue · Shufang Xu · Ren-Cang Li

Received: 27 September 2010 / Published online: 12 October 2011

Abstract This paper is concerned with the relative perturbation theory and its entrywise relatively accurate numerical solutions of an M-matrix Algebraic Riccati Equations (MARE)

$$XDX - AX - XB + C = 0$$

by which we mean the following conformally partitioned matrix

$$\begin{pmatrix} B & -D \\ -C & A \end{pmatrix}$$

is a nonsingular or an irreducible singular M-matrix. It is known that such an MARE has a unique minimal nonnegative solution Φ. It is proved that small relative perturbations to the entries of A, B, C, and D introduce small relative changes to the entries of the nonnegative solution Φ. Thus the smaller entries Φ do not suffer bigger relative errors than its larger entries, unlike the existing perturbation theory for (general) Algebraic Riccati Equations. We then discuss some minor but crucial implementation

J. Xue
School of Mathematical Science, Fudan University, Shanghai 200433, People's Republic of China
e-mail: xuej@fudan.edu.cn

S. Xu
School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China
e-mail: xsf@math.pku.edu.cn

R.-C. Li (✉)
Department of Mathematics, University of Texas at Arlington, Arlington, P.O. Box 19408, TX 76019, USA
e-mail: rcli@uta.edu

changes to three existing numerical methods so that they can be used to compute Φ as accurately as the input data deserve. Current study is based on a previous paper of the authors' on M-matrix Sylvester equation for which $D = 0$.

Mathematics Subject Classification (2000) 15A24 · 65F30 · 65G99 · 65H10

1 Introduction

An *M-Matrix Algebraic Riccati Equation*[1] (MARE) is the matrix equation

$$XDX - AX - XB + C = 0, \tag{1.1}$$

in which A, B, C, and D are matrices whose sizes are determined by the partitioning

$$W = \begin{matrix} \\ m \\ n \end{matrix} \begin{matrix} \begin{matrix} m & n \end{matrix} \\ \begin{pmatrix} B & -D \\ -C & A \end{pmatrix} \end{matrix}, \tag{1.2}$$

and W is a nonsingular or an irreducible singular M-matrix. This kind of Riccati equations arise in applied probability and transportation theory and have been attracting a lot of attention recently. See [9–11,13–15,18] and the references therein. It is shown in [10,11] that (1.1) has a unique minimal nonnegative solution Φ, i.e.,

$$\Phi \leq X \quad \text{for any other nonnegative solution } X \text{ of (1.1).}$$

When $D = 0$, MARE (1.1) degenerates to an M-matrix Sylvester equation (MSE)

$$AX + XB = C, \tag{1.3}$$

which has a unique solution that is nonnegative if A and B are M-matrices and one of them is also nonsingular.

Throughout this article, A, B, C, and D are reserved for the coefficient matrices of MARE (1.1) for which

$$\boxed{W \text{ defined by (1.2) is a nonsingular } M\text{-matrix or an irreducible singular } M\text{-matrix.}} \tag{1.4}$$

This assumption on W is the same as made in [9]. Their perturbed ones are denoted, respectively, by the same letters with a *tilde*, e.g., A is perturbed to $\widetilde{A}$, and the perturbed (1.1) is

$$\widetilde{X}\widetilde{D}\widetilde{X} - \widetilde{A}\widetilde{X} - \widetilde{X}\widetilde{B} + \widetilde{C} = 0. \tag{1.5}$$

[1] Previously it was called a Nonsymmetric Algebraic Riccati Equation, a name that seems to be too broad to be descriptive.

Our first goal in this paper is to perform an entrywise perturbation analysis for the minimal nonnegative solution Φ. Specifically, we seek bounds on the entrywise relative errors in the solution caused by small entrywise relative perturbations to the coefficient matrices A, B, C, and D. Our results suggest each and every entry of the solution, no matter how tiny it may be, is determined to a relative accuracy that is comparable to the entrywise relative accuracy residing in these coefficient matrices.

Previously related results in [9] bound the norm of the solution error: *if W is nonsingular or W is singular and irreducible with $u_1^{\mathrm{T}} v_1 \neq u_2^{\mathrm{T}} v_2$, and if $\|\widetilde{W} - W\|$ is sufficiently small, then there exists a constant $\beta > 0$ (dependent on W) such that*

$$\|\widetilde{\Phi} - \Phi\| \leq \beta \|\widetilde{W} - W\|, \tag{1.6}$$

where $\|\cdot\|$ is some matrix norm, each symbol without and with a *tilde* denotes the original and its corresponding perturbed one, and positive vectors u_i and v_i are defined in Theorem 2.1 later. In two aspects, the outcome of our analysis improves (1.6). First this normwise bound does not distinguish smaller entries from larger ones in the sense that (1.6) gives one bound for the absolute errors in all entries, regardless of their magnitudes. Secondly, this result only says β's existence and gives no useful information[2] as to how β relates to W. Our results, on the other hand, bound the entrywise relative errors in the solution directly and are explicitly expressed in ceratin parameters computable from W and Φ.

Following the analysis, we demonstrate that certain fixed point iterations [10], the structure-preserving doubling algorithm (SDA) [13], and the Newton method [9,11], all after some minor but crucial implementation changes, can deliver solutions with entrywise relative accuracy as the input data deserve. This contrasts favorably to many other methods (see [3,4,8–11,13] and references therein) most of which are backward stable in the normwise sense and cannot produce solutions with guaranteed entrywise relative accuracy as determined by the input data. This is our second goal.

This paper is organized as follows. Section 2 establishes a relative perturbation theory for the MARE. It is done with the help of the corresponding theory for an MSE in [21]. Some of the proofs for our results are long and complicated and so deferred to Sect. 3. Section 4 explains that three types of methods—fixed point iterations, SDA, and even the Newton method—after some minor but crucial implementation changes can be used to solve MARE (1.1) with the predicted relative accuracy by our theory. Numerical examples are given in Sect. 5 to demonstrate our theory and the effectiveness of the algorithms. Finally, we give our concluding remarks in Sect. 6.

Notation We will follow the notation as specified at the end of Sect. 1 in [21].

2 Entry-wise perturbation analysis

We shall present two kinds of first order error analysis. The first one given in Sect. 2.2 follows the standard approach: represent any perturbed matrix $\widetilde{Z}$ by $Z + \Delta Z$, expand

[2] It, however, may be possible to express β in a more explicit manner in terms of certain eigenvalues and eigenvectors related to W, following the proof in [9].

the perturbed MARE, and seek a bound on $\Delta\Phi$ through ignoring any term that involves two of ΔZ's. The obtained bound is sharp and suitable for practical error estimations. The second approach given in Sect. 2.3 is more complicated and gives less sharp bounds. But we argue that the resulting bounds yield more insightful information for theoretical understanding as to why the minimal nonnegative solution Φ should retain entrywise relative accuracy even for its smallest entries in the face of entrywise relative perturbations to W. It is made possible by revealing the important roles played by certain spectral radii to be defined in (2.10). More discussions on pros and cons for the bounds by both approaches will be given in Sect. 2.4.

2.1 Setting the stage

Recall if (1.4) holds, then the associated MARE (1.1) has a unique minimal nonnegative solution Φ [9]. Some properties of Φ are summarized in Theorem 2.1 below.

Theorem 2.1 **([9,10])** *Assume* (1.4).

(a) *MARE* (1.1) *has a unique minimal nonnegative solution* Φ*;*
(b) *If* W *is irreducible, then* $\Phi > 0$ *and* $A - \Phi D$ *and* $B - D\Phi$ *are irreducible* M*-matrices;*
(c) *If* W *is nonsingular, then* $A - \Phi D$ *and* $B - D\Phi$ *are nonsingular* M*-matrices;*
(d) *Suppose* W *is irreducible and singular. Let* $u_1, v_1 \in \mathbb{R}^m$ *and* $u_2, v_2 \in \mathbb{R}^n$ *be positive vectors such that*

$$W \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0, \quad \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}^{\mathrm{T}} W = 0. \tag{2.1}$$

1. *If* $u_1^{\mathrm{T}} v_1 > u_2^{\mathrm{T}} v_2$*, then* $B - D\Phi$ *is a singular* M*-matrix with*[3] $(B - D\Phi)v_1 = 0$ *and* $A - \Phi D$ *is a nonsingular* M*-matrix;*
2. *If* $u_1^{\mathrm{T}} v_1 < u_2^{\mathrm{T}} v_2$*, then* $B - D\Phi$ *is a nonsingular* M*-matrix and* $A - \Phi D$ *is a singular* M*-matrix;*
3. *If* $u_1^{\mathrm{T}} v_1 = u_2^{\mathrm{T}} v_2$*, then both* $B - D\Phi$ *and* $A - \Phi D$ *are singular* M*-matrices.*

Suppose that (1.1) is perturbed to (1.5) with small entrywise relative errors such that

$$|\widetilde{A} - A| \le \epsilon |A|, \ |\widetilde{B} - B| \le \epsilon |B|, \ |\widetilde{C} - C| \le \epsilon C, \ |\widetilde{D} - D| \le \epsilon D, \tag{2.2}$$

where $0 \le \epsilon < 1$. When the associated

$$\widetilde{W} = \begin{pmatrix} \widetilde{B} & -\widetilde{D} \\ -\widetilde{C} & \widetilde{A} \end{pmatrix}$$

[3] [10, Theorem 4.8] says in this case $D\Phi v_1 = Dv_2$ which leads to $(B - D\Phi)v_1 = Bv_1 - Dv_2 = 0$.

is also a nonsingular M-matrix or an irreducible singular M-matrix, (1.5) has a unique minimal nonnegative solution $\widetilde{\Phi}$, too. In this section, we shall seek to bound the entrywise relative error in $\widetilde{\Phi}$.

Rewrite (1.1), after substituting $X = \Phi$, as

$$(A - \Phi D)\Phi + \Phi(B - D\Phi) = C - \Phi D\Phi, \tag{2.3}$$

and define Φ_1 and Φ_2 by

$$(A - \Phi D)\Phi_1 + \Phi_1(B - D\Phi) = C, \tag{2.4}$$
$$(A - \Phi D)\Phi_2 + \Phi_2(B - D\Phi) = \Phi D\Phi. \tag{2.5}$$

Then $\Phi = \Phi_1 - \Phi_2$. When W is a nonsingular or an irreducible singular M-matrix with $u_1^{\mathrm{T}} v_1 \neq u_2^{\mathrm{T}} v_2$,

$$P_\Phi \stackrel{\text{def}}{=} I_m \otimes (A - \Phi D) + (B - D\Phi)^{\mathrm{T}} \otimes I_n \tag{2.6}$$

is a nonsingular M-matrix by [9, Theorem 1.1]. This P_Φ is also a matrix representation of the following linear operator

$$\mathcal{L}_\Phi \,:\, X \to (A - \Phi D)X + X(B - D\Phi). \tag{2.7}$$

Since $P_\Phi^{-1} \geq 0$, $\mathcal{L}_\Phi^{-1}(X_1) \leq \mathcal{L}_\Phi^{-1}(X_2)$ for $X_1 \leq X_2$, and especially $\mathcal{L}_\Phi^{-1}(X) \geq 0$ for $X \geq 0$. We also have

$$\Phi_1 \geq \Phi_2 \geq 0, \;\; \Phi_1 \geq \Phi \geq 0$$

because $\Phi = \Phi_1 - \Phi_2 \geq 0$.

Proposition 2.1 *Suppose* (1.4). $\Phi_{(i,j)} = 0$ *if and only if* $(\Phi_1)_{(i,j)} = 0$.

The proof of this proposition is deferred to Sect. 3. Define, with the understanding that the indeterminant $0/0$ is regarded as 0,

$$\kappa = \max_{i,j} \frac{(\Phi_1)_{(i,j)}}{\Phi_{(i,j)}}. \tag{2.8}$$

It is evident that $\kappa \geq 1$. Proposition 2.1 ensures also $\kappa < \infty$.

Split A and B as

$$A = D_1 - N_1, \quad D_1 = \mathrm{diag}(A), \tag{2.9a}$$
$$B = D_2 - N_2, \quad D_2 = \mathrm{diag}(B). \tag{2.9b}$$

Correspondingly

$$A - \Phi D = D_1 - N_1 - \Phi D, \quad B - D\Phi = D_2 - N_2 - D\Phi,$$

and set

$$\lambda_1 = \rho(D_1^{-1}(N_1 + \Phi D)), \quad \lambda_2 = \rho(D_2^{-1}(N_2 + D\Phi)), \quad \lambda = \max\{\lambda_1, \lambda_2\}, \tag{2.10}$$

$$\tau_1 = \frac{\min_i A_{(i,i)}}{\max_j B_{(j,j)}}, \qquad \tau_2 = \frac{\min_j B_{(j,j)}}{\max_i A_{(i,i)}}. \tag{2.11}$$

If W is nonsingular, then $A - \Phi D$ and $B - D\Phi$ are nonsingular M-matrices by Theorem 2.1; so $\lambda_1 < 1$ and $\lambda_2 < 1$ [20, Theorem 3.15 on p.90] and thus $0 \le \lambda < 1$. If W is an irreducible singular M-matrix, then by Theorem 2.1

1. if $u_1^{\mathrm{T}} v_1 > u_2^{\mathrm{T}} v_2$, then $\lambda_1 < 1$ and $\lambda_2 = 1$;
2. if $u_1^{\mathrm{T}} v_1 < u_2^{\mathrm{T}} v_2$, then $\lambda_1 = 1$ and $\lambda_2 < 1$;
3. if $u_1^{\mathrm{T}} v_1 = u_2^{\mathrm{T}} v_2$, then $\lambda_1 = \lambda_2 = 1$.

The third case $u_1^{\mathrm{T}} v_1 = u_2^{\mathrm{T}} v_2$ is rather extreme and for which P_Φ is singular. It is argued in [9] that for the case for sufficiently small $\|\widetilde{W} - W\|$ there exists a constant β such that

1. $\|\widetilde{\Phi} - \Phi\| \le \beta \|\widetilde{W} - W\|^{1/2}$;
2. $\|\widetilde{\Phi} - \Phi\| \le \beta \|\widetilde{W} - W\|$ if $\widetilde{W}$ is also singular.

This β like the one in (1.6), also due to [9], is known by its existence. Our current analysis below does not work for this extreme case for which it remains to be an open problem whether our approach here can be extended to yield more informative error bounds.

2.2 Standard first order error analysis

The standard first order error analysis is often very useful in revealing the sensitivity of the interested problem because it usually produces asymptotically best possible first order error bounds. When applied to MARE, it gives a first order error bound that depends on the solution of another Sylvester equation.

Lemma 2.1 *Suppose* (1.4) *and that* $\widetilde{W}$ *is a nonsingular M-matrix or an irreducible singular M-matrix. If* (2.2) *holds for some* $0 \le \epsilon < 1$, *then* $\Phi_{(i,j)} = 0$ *if and only if* $\widetilde{\Phi}_{(i,j)} = 0$.

Proof The conclusion is evident if W is irreducible because then $\Phi > 0$ and $\widetilde{\Phi} > 0$ by Theorem 2.1(b). In general, consider the iterations: $Z_0 = \widetilde{Z}_0 = 0$ and for $k \ge 0$,

$$\begin{aligned} D_1 Z_{k+1} + Z_{k+1} D_2 &= C + Z_k D Z_k + N_1 Z_k + Z_k N_2, \\ \widetilde{D}_1 \widetilde{Z}_{k+1} + \widetilde{Z}_{k+1} \widetilde{D}_2 &= \widetilde{C} + \widetilde{Z}_k \widetilde{D} \widetilde{Z}_k + \widetilde{N}_1 \widetilde{Z}_k + \widetilde{Z}_k \widetilde{N}_2. \end{aligned}$$

The conditions of this lemma implies that $I_m \otimes A + B^{\mathrm{T}} \otimes I_n$ and $I_m \otimes \widetilde{A} + \widetilde{B}^{\mathrm{T}} \otimes I_n$ are nonsingular M-matrices. Therefore [10]

$$Z_0 \le Z_1 \le Z_2 \le \cdots, \quad \lim_{k\to\infty} Z_k = \Phi, \quad \text{and}$$

$$\widetilde{Z}_0 \le \widetilde{Z}_1 \le \widetilde{Z}_2 \le \cdots, \quad \lim_{k\to\infty} \widetilde{Z}_k = \widetilde{\Phi}.$$

It is sufficient to prove that Z_k and $\widetilde{Z}_k$ have the same zero-nonzero pattern for each k, i.e., $(Z_k)_{(i,j)} = 0$ if and only if $(\widetilde{Z}_k)_{(i,j)} = 0$. This can be done by induction on k. No proof is necessary for $k = 0$. Assume that this is true for $k = \ell$. Consider now $k = \ell + 1$. Since W and $\widetilde{W}$ have the same zero-nonzero pattern and, by the induction hypothesis, Z_ℓ and $\widetilde{Z}_\ell$ have the same zero-nonzero pattern, we conclude that $C + Z_\ell D Z_\ell + N_1 Z_\ell + Z_\ell N_2 = D_1 Z_{\ell+1} + Z_{\ell+1} D_2$ and $\widetilde{C} + \widetilde{Z}_\ell \widetilde{D} \widetilde{Z}_\ell + \widetilde{N}_1 \widetilde{Z}_\ell + \widetilde{Z}_\ell \widetilde{N}_2 = \widetilde{D}_1 \widetilde{Z}_{\ell+1} + \widetilde{Z}_{\ell+1} \widetilde{D}_2$ must have the same zero-nonzero pattern, and so does $Z_{\ell+1}$ and $\widetilde{Z}_{\ell+1}$. □

Theorem 2.2 *Suppose that W in (1.2) is a nonsingular M-matrix or an irreducible singular M-matrix with $u_1^{\mathrm{T}} v_1 \ne u_2^{\mathrm{T}} v_2$. Suppose (2.2) and that $\widetilde{W}$ is an M-matrix. For sufficiently small ϵ, we have*

$$|(\Phi - \widetilde{\Phi}) \oslash \Phi| \le 2\epsilon \Upsilon \oslash \Phi + O(\epsilon^2) \tag{2.12}$$

$$\le 2\gamma\epsilon \, \mathbf{1}_n \mathbf{1}_m^{\mathrm{T}} + O(\epsilon^2), \tag{2.13}$$

where $\oslash$ denotes the entrywise division, Υ and γ are defined by

$$(A - \Phi D)\Upsilon + \Upsilon(B - D\Phi) = D_1\Phi + \Phi D_2, \quad \gamma = \max_{i,j} (\Upsilon \oslash \Phi)_{(i,j)}. \tag{2.14}$$

Proof Write $\widetilde{Z} = Z + (\Delta Z)$ for $Z = A, B, C, D$, and Φ, where ΔZ is the perturbation to Z. Substitute them into $\widetilde{\Phi}\widetilde{D}\widetilde{\Phi} - \widetilde{A}\widetilde{\Phi} - \widetilde{\Phi}\widetilde{B} + \widetilde{C} = 0$ and notice $\Phi D \Phi - A\Phi - \Phi B + C = 0$ to get

$$\begin{aligned}&(A - \Phi D)(\Delta\Phi) + (\Delta\Phi)(B - D\Phi)\\ &\quad = \widetilde{\Phi}(\Delta D)\widetilde{\Phi} + (\Delta\Phi)D(\Delta\Phi) - (\Delta A)\widetilde{\Phi} - \widetilde{\Phi}(\Delta B) + (\Delta C).\end{aligned}$$

Since $\mathcal{L}_\Phi$ is invertible, this equation implies $\Delta\Phi = O(\epsilon)$ for sufficiently tiny ϵ. Therefore

$$\begin{aligned}|\Delta\Phi| &\le \mathcal{L}_\Phi^{-1}\big(|\widetilde{\Phi}(\Delta D)\widetilde{\Phi} + (\Delta\Phi)D(\Delta\Phi) - (\Delta A)\widetilde{\Phi} - \widetilde{\Phi}(\Delta B) + (\Delta C)|\big)\\ &\le \epsilon\, \mathcal{L}_\Phi^{-1}\big(\Phi D\Phi + |A|\Phi + \Phi|B| + C\big) + O(\epsilon^2)\\ &= 2\epsilon\, \mathcal{L}_\Phi^{-1}\big(D_1\Phi + \Phi D_2\big) + O(\epsilon^2)\end{aligned} \tag{2.15}$$

which yields (2.13) since Φ and $\widetilde{\Phi}$ have the same zero-nonzero pattern by Lemma 2.1. □

The following proposition proves that (2.12) is sharp and γ is finite.

Proposition 2.2 *In* Theorem 2.2, $\limsup_{\epsilon\to 0} \dfrac{|\Delta\Phi|}{\epsilon} = 2\Upsilon$ *and* $\gamma < \infty$.

Proof In the proof of Theorem 2.2, if we take

$$\Delta A = \epsilon|A|, \quad \Delta B = \epsilon|B|, \quad \Delta C = -\epsilon C, \quad \Delta D = -\epsilon D, \tag{2.16}$$

then $\widetilde{W}$ is a nonsingular M-matrix for sufficiently tiny positive ϵ since $\widetilde{W} > W$. Examine each inequality sign in (2.15) to see that up to the first order $|\Delta\Phi| = 2\epsilon\, \mathcal{L}_{\Phi}^{-1}\big(D_1\Phi + \Phi D_2\big)$, and therefore the limit is 2Υ.

To show that $\gamma < \infty$ under the conditions of Theorem 2.2, it suffices to show that $\Phi_{(i,j)} = 0$ implies $\Upsilon_{(i,j)} = 0$. In fact, $\Phi_{(i,j)} = 0$ implies $\widetilde{\Phi}_{(i,j)} = 0$ by Lemma 2.1 and thus $\Delta\Phi_{(i,j)} = 0$, and therefore $\Upsilon_{(i,j)} = 0$ by the limit formula. □

Remark 2.1 The following iterative scheme: $\Upsilon_0 = \Phi/2$ and for $k \geq 0$

$$D_1\Upsilon_{k+1} + \Upsilon_{k+1}D_2 = (N_1 + \Phi D)\Upsilon_k + \Upsilon_k(N_2 + D\Phi) + D_1\Phi + \Phi D_2 \tag{2.17}$$

produces a sequence $\{\Upsilon_i\}$ that monotonically convergent to Υ because P_Φ is a nonsingular M-matrix. That we start with $\Upsilon_0 = \Phi/2$ is because $\Upsilon \geq \Phi/2$. So it is numerically simple to use (2.13): iterate (2.17) enough steps until Υ_k has one or more correct decimal digits in each of its entries. The subsequently estimated γ by the approximate Υ through (2.14) should give an adequate estimate for entrywise relative errors. But this can be costly sometimes when (2.17) is slowly convergent and/or some entries of Υ are of much tinier magnitudes than others (because convergence to entries of different magnitudes is not uniform in general). □

As a consequence of Theorem 2.2 and Proposition 2.2, we deduce immediately the so-called *componentwise condition number* for MARE (1.1): 2γ in the sense of [7]. Since an MARE is an example of a nonsymmetric algebraic Riccati equation (NARE), this result of ours is a special case of Lin and Wei [16] who were interested in general NAREs. In comparing ours to the one in [16], we have taken advantage of (1.1) being an MARE, followed the simple and standard way for derivation, and then arrived at a simpler expression. We have also shown that 2γ is finite. This is very important because for a general NARE, it is conceivable that its componentwise condition number could be infinite (perhaps more often than not).

2.3 New first order error analysis

Theorem 2.3 and Corollary 2.1 are for the case in which W is a nonsingular M-matrix, while Theorems 2.4 and 2.5 are about the case in which W is an irreducible singular M-matrix with $u_1^{\mathrm{T}}v_1 \neq u_2^{\mathrm{T}}v_2$.

Throughout this subsection, λ_i for $i = 1, 2$ and λ are defined as in (2.10), κ as in (2.8), τ_i for $i = 1, 2$ as in (2.11), and γ as in Theorem 2.2.

All bounds take the form, for sufficiently small ϵ,

$$|\Phi - \widetilde{\Phi}| \leq \Big[2mn\,\kappa\,\chi\,\epsilon + O(\epsilon^2)\Big]\,\Phi, \tag{2.18}$$

where χ is a constant dependent on λ_i and τ_i. Following the proofs in section 3, we find that the second order term $O\left(\epsilon^2\right)$ in (2.18) is bounded by

$$\kappa\ f(m,n)(\chi+\gamma)^2\epsilon^2 \tag{2.19}$$

for some low degree polynomials f in m and n.

The proofs of the following theorems are rather complicated and thus deferred to Sect. 3.

Theorem 2.3 *Suppose that W in (1.2) is a nonsingular M-matrix, and suppose (2.2). Then (2.18) holds with*

$$\chi=\max\left\{\frac{1+\lambda_1+(1+\lambda_2)\tau_1^{-1}}{1-\lambda_1+(1-\lambda_2)\tau_1^{-1}},\ \frac{1+\lambda_2+(1+\lambda_1)\tau_2^{-1}}{1-\lambda_2+(1-\lambda_1)\tau_2^{-1}}\right\}. \tag{2.20}$$

Corollary 2.1 *Under the conditions of* Theorem 2.3, (2.18) *holds with*

$$\chi=\frac{1+\lambda}{1-\lambda}. \tag{2.21}$$

Proof Notice

$$\frac{1+\lambda_i+(1+\lambda_j)\tau_i^{-1}}{1-\lambda_i+(1-\lambda_j)\tau_i^{-1}}\le\frac{1+\lambda}{1-\lambda}$$

and then apply Theorem 2.3. □

Theorem 2.4 *Suppose that W in (1.2) is an irreducible singular M-matrix with $u_1^{\mathrm{T}}v_1\neq u_2^{\mathrm{T}}v_2$. Suppose*[4] *(2.2) and $\widetilde{W}$ is an M-matrix. Then (2.18) holds with*

$$\chi=2\begin{cases}\dfrac{1+\lambda_1+2\tau_1^{-1}}{1-\lambda_1}, & \text{if } u_1^{\mathrm{T}}v_1>u_2^{\mathrm{T}}v_2,\\ \dfrac{1+\lambda_2+2\tau_2^{-1}}{1-\lambda_2}, & \text{if } u_1^{\mathrm{T}}v_1<u_2^{\mathrm{T}}v_2.\end{cases} \tag{2.22}$$

Theorem 2.5 aims at the perturbation analysis for the Wiener-Hopf factorization of a Markov chain.

Theorem 2.5 *Suppose that $-W$ in (1.2) is the generator of an irreducible Markov chain, i.e., W is an irreducible singular M-matrix with $v_1=\mathbf{1}_m$, $v_2=\mathbf{1}_n$ in (2.1). Suppose (2.2), and that $-\widetilde{W}$ is also the generator of an irreducible Markov chain.*

[4] Because of (2.2), $\widetilde{W}$ too is irreducible if $\epsilon<1$.

Then (2.18) *holds with*[5]

$$\chi = \begin{cases} 2 \times \dfrac{1+\lambda_1}{1-\lambda_1}, & \text{if } u_1^{\mathrm{T}}\mathbf{1}_m > u_2^{\mathrm{T}}\mathbf{1}_n, \\ \dfrac{2(m+n)}{mn} + 2[4(m+n)+1]\dfrac{1+\lambda_2}{1-\lambda_2}, & \text{if } u_1^{\mathrm{T}}\mathbf{1}_m < u_2^{\mathrm{T}}\mathbf{1}_n. \end{cases} \tag{2.23}$$

If also $u_1 = \mathbf{1}_m$ *and* $u_2 = \mathbf{1}_n$ (*i.e.*, $\mathbf{1}_{m+n}^{\mathrm{T}} W = 0$), $m \neq n$, *and* $\mathbf{1}_{m+n}^{\mathrm{T}} \widetilde{W} = 0$, *then*

$$\chi = 2 \times \min_i \frac{1+\lambda_i}{1-\lambda_i}. \tag{2.23$'$}$$

The conditions of Theorem 2.5 make Theorem 2.4 immediately applicable for the case. What distinguishes the two theorems is that χ given by (2.22) contains τ_i while χ by (2.23) contains no τ_i but, as a tradeoff, contains a big factor roughly $8(m+n)$ when $u_1^{\mathrm{T}}\mathbf{1}_m < u_2^{\mathrm{T}}\mathbf{1}_n$.

2.4 Discussions

We have obtained two kinds of first order error bounds in subsections 2.2 and 2.3. Both have their own virtues and shortcomings.

The first order error bound by Theorem 2.2 yields the componentwise condition number in the sense of [7] and thus is sharp, but not without sacrifice. Namely, without actually solving the first equation in (2.14), it is hard to imagine that the bounds by Theorem 2.2 is going to be tiny (comparable to ϵ).

The first order error bounds in subsection 2.3 implies that $|(\Phi - \widetilde{\Phi}) \oslash \Phi|$ will be large if $(1-\lambda_1)^{-1}$ and $(1-\lambda_2)^{-1}$ are large, and tiny if $(1-\lambda_1)^{-1}$ or $(1-\lambda_2)^{-1}$ is modest and κ is modest. Such a conclusion cannot be read off from Theorem 2.2. The use of the spectral radii can be beneficial because usually spectral radii are able to expose insight information of matrices and can be estimated with relatively little effort. Also our proofs in the next section lead us to conclude a rough bound (2.19) on the second order terms.

Due to the artifact of our proofs, the first order error bounds in subsection 2.3 turn to have constant factors that are overestimated and consequently less sharp than the bound by Theorem 2.2. In this respect, their values are more theoretical than practical. They also assure us that γ in Theorem 2.2 will be modest if $(1-\lambda_1)^{-1}$ or $(1-\lambda_2)^{-1}$ is modest and κ is modest. In fact, we have

[5] The proof in the next section says we can have a slightly smaller

$$\chi = \frac{2(m+n)}{mn\,\kappa} + 2[4(m+n)+1]\frac{1+\lambda_2}{1-\lambda_2}$$

for the case $u_1^{\mathrm{T}}\mathbf{1}_m < u_2^{\mathrm{T}}\mathbf{1}_n$. We drop κ off from this expression mainly to make it independent of κ as we do for all χ in the other theorems.

Proposition 2.3 *Under the conditions of Theorem 2.2, we have*

$$\kappa \le 2\gamma \le 2mn\,\kappa\,\chi, \tag{2.24}$$

where χ is defined in Theorems 2.3–2.5 according to the different cases of the theorems, κ in (2.8), and γ in (2.14).

Proof It can be seen from (2.4) and (2.14) that $2\Upsilon \ge \Phi_1$ because

$$2(D_1\Phi + \Phi D_2) = \Phi D\Phi + |A|\Phi + \Phi|B| + C \ge C.$$

This implies the first inequality in (2.24). The second inequality is a consequence of Proposition 2.2, together with Theorems 2.2–2.5. □

We now argue that the second inequality in (2.24) in general cannot be improved, modulo a factor cmn, where c is some constant independent of the dimensional parameters m and n. To this end, we consider nonsingular and irreducible M-matrices $A = I - N_1$ and $B = I - N_2$, where $N_i \ge 0$ and $\mathrm{diag}(N_i) = 0$. Thus for $i = 1, 2$

$$\varrho_i = \rho(N_i) < 1.$$

Let $u > 0$ and $y > 0$ be the Perron eigenvectors of N_1 and N_2^{T}, respectively, i.e.,

$$N_1 u = \varrho_1 u, \quad y^{\mathrm{T}} N_2 = \varrho_2 y^{\mathrm{T}}.$$

Scale u and y so that $y^{\mathrm{T}}u = 1$. Pick $\xi > 0$ and $\zeta > 0$ sufficiently small such that W with A and B just defined and $C = \xi uy^{\mathrm{T}}$ and $D = \zeta uy^{\mathrm{T}}$ is an irreducible nonsingular or a singular M-matrix. We now have constructed an MARE (1.1). Following the argument in [10], we find that the following iterative method: $Z_0 = 0$ and for $k \ge 0$,

$$AZ_{k+1} + Z_{k+1}B = C + Z_k D Z_k$$

will produce a monotonically increasing sequence $\{Z_k\}$ that converges to the nonnegative minimal solution Φ of the MARE. It can be proved, e.g., by induction, that $Z_k = \eta_k uy^{\mathrm{T}}$ for some $\eta_k \ge 0$. Therefore $\Phi = \eta uy^{\mathrm{T}}$ for some $\eta \ge 0$. Substitute $\Phi = \eta uy^{\mathrm{T}}$ into (1.1) to see that η is the smallest positive root of

$$\zeta\eta^2 - (2 - \varrho_1 - \varrho_2)\eta + \xi = 0.$$

This gives

$$\eta = \frac{2\xi}{(2 - \varrho_1 - \varrho_2) + \sqrt{(2 - \varrho_1 - \varrho_2)^2 - 4\xi\zeta}}.$$

Now note $\lambda_i = 1 - \varrho_i - \eta\zeta$ for $i = 1, 2$ because $u > 0$, $y > 0$, and

$$(A - \Phi D)u = (1 - \varrho_1 - \eta\zeta)u, \quad y^{\mathrm{T}}(B - D\Phi) = (1 - \varrho_2 - \eta\zeta)y^{\mathrm{T}}.$$

It can be verified that

$$\Phi_1 = \frac{\xi}{2-\lambda_1-\lambda_2}\, uy^{\mathrm{T}}, \quad \Upsilon = \frac{2\eta}{2-\lambda_1-\lambda_2}\, uy^{\mathrm{T}}$$

and consequently

$$\kappa = \frac{\xi}{\eta(2-\lambda_1-\lambda_2)}, \quad \gamma = \frac{2}{2-\lambda_1-\lambda_2}.$$

Since all χ in Theorems 2.3 and 2.4 and Theorem 2.5 for the case $u_1^{\mathrm{T}}\mathbf{1}_m > u_2^{\mathrm{T}}\mathbf{1}_n$ are within a constant factor (independent of m and n) of $(2-\lambda_1-\lambda_2)^{-1}$, it suffices for us to compare γ with $\kappa\,(2-\lambda_1-\lambda_2)^{-1}$ for the purpose of comparing γ with $\kappa\,\chi$. We have

$$\begin{aligned}\frac{\gamma}{\kappa\,(2-\lambda_1-\lambda_2)^{-1}} &= \frac{2}{\kappa} = \frac{2\eta(2-\lambda_1-\lambda_2)}{\xi} \\ &= \frac{4(2-\lambda_1-\lambda_2)}{(2-\varrho_1-\varrho_2)+\sqrt{(2-\varrho_1-\varrho_2)^2-4\xi\zeta}},\end{aligned}$$

and therefore

$$\lim_{\zeta\to 0^+} \frac{\gamma}{\kappa\,(2-\lambda_1-\lambda_2)^{-1}} = 2.$$

This means, modulo a factor cmn, the second inequality in (2.24) cannot be improved. As a by-product, we see that it is possible $\gamma/\kappa \gg 1$.

3 Proofs

Proof of Proposition 2.1 If W in (1.2) is irreducible, then P_Φ is irreducible and $\Phi > 0$ [9] and thus $\Phi_1 \geq \Phi > 0$. For the general case in which W may be reducible, a more complicated argument is needed to prove the conclusion. The proof given below does not distinguish whether W is reducible or not.

Since $\Phi_1 \geq \Phi \geq 0$, $(\Phi_1)_{(i,j)} = 0$ implies $\Phi_{(i,j)} = 0$.

It remains to show that $\Phi_{(i,j)} = 0$ implies $(\Phi_1)_{(i,j)} = 0$. Split A and B as in (2.9). We have

$$D_1\Phi + \Phi D_2 = \Phi D\Phi + N_1\Phi + \Phi N_2 + C.$$

Because D_1 and D_2 are diagonal with positive diagonal entries, $(D_1\Phi+\Phi D_2)_{(i,j)} = 0$ if and only if $\Phi_{(i,j)} = 0$. Therefore,

$$\Phi_{(i,j)} = 0 \quad \Rightarrow \quad (\Phi D\Phi)_{(i,j)} = (N_1\Phi)_{(i,j)} = (\Phi N_2)_{(i,j)} = C_{(i,j)} = 0. \quad (3.1)$$

Consider the following iteration

$$Z_0 = 0, \tag{3.2a}$$
$$D_1 Z_{k+1} + Z_{k+1} D_2 = C + \Phi D Z_k + Z_k D \Phi + N_1 Z_k + Z_k N_2 \quad \text{for } k \geq 0. \tag{3.2b}$$

This corresponds to an iteration to compute Φ_1 based on the so-called *regular splitting* [20] of (2.4) after written equivalently as $P_\Phi \operatorname{vec}(\Phi_1) = \operatorname{vec}(C)$:

$$P_\Phi = [I_n \otimes D_1 + D_2^{\mathrm{T}} \otimes I_m] - [I_n \otimes (N_1 + \Phi D) + (N_2 + D\Phi)^{\mathrm{T}} \otimes I_m].$$

Therefore (3.2) is convergent [20, Theorem 3.15 on p.90] and $\lim_{k\to 0} Z_k = \Phi_1$. We claim that the sequence is monotonically increasing, i.e., $Z_0 \leq Z_1 \leq Z_2 \leq \cdots$, too. Clearly $Z_0 \leq Z_1$. Suppose that $Z_{k-1} \leq Z_k$ for all $k \leq \ell$. We shall prove that $Z_{k-1} \leq Z_k$ holds for $k = \ell + 1$ as well. In fact, we have by (3.2b)

$$D_1(Z_{\ell+1} - Z_\ell) + (Z_{\ell+1} - Z_\ell) D_2 = (N_1 + \Phi D)(Z_\ell - Z_{\ell-1}) + (Z_\ell - Z_{\ell-1}) \times (D\Phi + N_2)$$

which leads to $Z_{\ell+1} - Z_\ell \geq 0$ because $Z_\ell - Z_{\ell-1} \geq 0$ by the induction hypothesis, $N_1 + \Phi D \geq 0$, and $D\Phi + N_2 \geq 0$.

Let $\mathscr{I} = \{(i, j) : \Phi_{(i,j)} = 0\}$. We claim that for any $(i, j) \in \mathscr{I}$, $(Z_k)_{(i,j)} = 0$ for all $k \geq 0$. We prove this again by the induction on k. Clearly it holds for $k = 0$ since $Z_0 = 0$. Suppose that $(Z_k)_{(i,j)} = 0$ for all $k \leq \ell$ and any $(i, j) \in \mathscr{I}$. We shall prove $(Z_k)_{(i,j)} = 0$ for $k = \ell + 1$ and any $(i, j) \in \mathscr{I}$ as well. In the remainder of this proof, (i, j) represents an arbitrary index pair from $\mathscr{I}$. By (3.1), $(\Phi D \Phi)_{(i,j)} = 0$. So must $(\Phi D Z_\ell)_{(i,j)} = 0$ too, because

$$(\Phi D \Phi)_{(i,j)} = \sum_p (\Phi D)_{(i,p)} \Phi_{(p,j)} = 0$$

implies if $(\Phi D)_{(i,p)} \neq 0$, then $\Phi_{(p,j)} = 0$ which implies $(Z_\ell)_{(p,j)} = 0$ by the induction hypothesis. Therefore

$$(\Phi D Z_\ell)_{(i,j)} = \sum_p (\Phi D)_{(i,p)} (Z_\ell)_{(p,j)} = 0.$$

Similarly $(Z_\ell D \Phi)_{(i,j)} = 0$, and $(N_1 Z_\ell)_{(i,j)} = 0$ and $(Z_\ell N_2)_{(i,j)} = 0$ because $(N_1 \Phi)_{(i,j)} = (\Phi N_2)_{(i,j)} = 0$ by (3.1). Equation (3.2b) for $k = \ell$ yields $(D_1 Z_{\ell+1} + Z_{\ell+1} D_2)_{(i,j)} = 0$, i.e., $(Z_{\ell+1})_{(i,j)} = 0$. This completes the proof. □

Lemma 3.1 ([10, Theorem 2.4]) *Assume* (1.4). *Then the minimal nonnegative solution Φ increases entrywise as the entries of C and D increase, and the entries of A and B decreases, provided the corresponding perturbed $\widetilde{W}$ is still a nonsingular M-matrix or an irreducible singular M-matrix.*

The proofs below for the three theorems rely on splitting $\widetilde{\Phi}$ into three pieces as follows. It can be verified that

$$\widetilde{\Phi}(\widetilde{B}-\widetilde{D}\Phi)+(\widetilde{A}-\Phi\widetilde{D})\widetilde{\Phi}=\widetilde{C}-\Phi\widetilde{D}\Phi+(\widetilde{\Phi}-\Phi)\widetilde{D}(\widetilde{\Phi}-\Phi). \tag{3.3}$$

Recall $\Phi=\Phi_1-\Phi_2$, where Φ_1 and Φ_2 are defined in (2.4) and (2.5). Define $\check{\Phi}_1$, $\check{\Phi}_2$, and Δ by

$$(\widetilde{A}-\Phi\widetilde{D})\check{\Phi}_1+\check{\Phi}_1(\widetilde{B}-\widetilde{D}\Phi)=\widetilde{C}, \tag{3.4}$$
$$(\widetilde{A}-\Phi\widetilde{D})\check{\Phi}_2+\check{\Phi}_2(\widetilde{B}-\widetilde{D}\Phi)=\Phi\widetilde{D}\Phi, \tag{3.5}$$
$$(\widetilde{A}-\Phi\widetilde{D})\Delta+\Delta(\widetilde{B}-\widetilde{D}\Phi)=(\widetilde{\Phi}-\Phi)\widetilde{D}(\widetilde{\Phi}-\Phi). \tag{3.6}$$

Then

$$\widetilde{\Phi}=\check{\Phi}_1-\check{\Phi}_2+\Delta. \tag{3.7}$$

The basic idea of our proofs below is to seek bounds on $|\Phi_i-\check{\Phi}_i|$ and $|\Delta|$.

Proof of Theorem 2.3 By Lemma 3.1, it suffices to consider the two extreme cases:

$$\begin{aligned}\widetilde{A}&=(1-\epsilon)D_1-(1+\epsilon)N_1,\\ \widetilde{B}&=(1-\epsilon)D_2-(1+\epsilon)N_2,\ \widetilde{C}=(1+\epsilon)C,\ \widetilde{D}=(1+\epsilon)D;\end{aligned} \tag{3.8}$$
$$\begin{aligned}\widetilde{A}&=(1+\epsilon)D_1-(1-\epsilon)N_1,\\ \widetilde{B}&=(1+\epsilon)D_2-(1-\epsilon)N_2,\ \widetilde{C}=(1-\epsilon)C,\ \widetilde{D}=(1-\epsilon)D.\end{aligned} \tag{3.9}$$

In what follows, we shall consider the case (3.8) only, because the other one can be dealt with similarly. We then have $0\le\Phi\le\widetilde{\Phi}$. By Theorem 2.2,

$$0\le(\widetilde{\Phi}-\Phi)\widetilde{D}(\widetilde{\Phi}-\Phi)\le\left[4\gamma^2\epsilon^2+O(\epsilon^3)\right]\Phi\widetilde{D}\Phi. \tag{3.10}$$

Compare (3.6) to (3.5) to get

$$0\le\Delta\le\left[4\gamma^2\epsilon^2+O(\epsilon^3)\right]\check{\Phi}_2. \tag{3.11}$$

We claim that for $i=1,2$

$$0\le\check{\Phi}_i-\Phi_i\le\left[2mn\,\chi\,\epsilon+O(\epsilon^2)\right]\Phi_i. \tag{3.12}$$

Suppose, for the moment, that $A-\Phi D$ and $B-D\Phi$ are irreducible. Then there exist vectors $u>0$ and $y>0$ such that

$$D_1^{-1}(N_1+\Phi D)u=\lambda_1 u,\quad D_2^{-1}(N_2+D\Phi)^{\mathrm{T}}y=\lambda_2 y.$$

Let $\widetilde{P}_\Phi = I_m \otimes (\widetilde{A} - \Phi\widetilde{D}) + (\widetilde{B} - \widetilde{D}\Phi)^{\mathrm{T}} \otimes I_n$. We have

$$P_\Phi(y \otimes u) = (1-\lambda_1)y \otimes D_1 u + (1-\lambda_2)D_2 y \otimes u, \tag{3.13}$$

$$\begin{aligned}\widetilde{P}_\Phi(y \otimes u) = {} & [(1-\epsilon) - (1+\epsilon)\lambda_1]y \otimes (D_1 u) \\ & + [(1-\epsilon) - (1+\epsilon)\lambda_2](D_2 y) \otimes u.\end{aligned} \tag{3.14}$$

Both right-hand sides are positive for sufficiently small ϵ. Now we look at the entrywise ratio $[\widetilde{P}_\Phi(y \otimes u)] \oslash [P_\Phi(y \otimes u)]$ whose typical kth entry is, by (3.13) and (3.14),

$$\frac{[\widetilde{P}_\Phi(y \otimes u)]_{(k)}}{[P_\Phi(y \otimes u)]_{(k)}} = \frac{[(1-\epsilon) - (1+\epsilon)\lambda_1]A_{(i,i)} + [(1-\epsilon) - (1+\epsilon)\lambda_2]B_{(j,j)}}{(1-\lambda_1)A_{(i,i)} + (1-\lambda_2)B_{(j,j)}} \tag{3.15}$$

for some i and j. The derivative of the right-hand side of (3.15) with respect to $t = A_{(i,i)}/B_{(j,j)}$ is

$$\frac{2\epsilon(\lambda_2 - \lambda_1)}{[(1-\lambda_1)t + (1-\lambda_2)]^2}$$

which implies the right-hand side of (3.15) is an increasing function of $t = A_{(i,i)}/B_{(j,j)}$ if $\lambda_1 \le \lambda_2$ and a decreasing function of $t = A_{(i,i)}/B_{(j,j)}$ if $\lambda_1 > \lambda_2$. Therefore

$$P_\Phi(y \otimes u) \ge \widetilde{P}_\Phi(y \otimes u) \ge (1 - \chi\,\epsilon)P_\Phi(y \otimes u). \tag{3.16}$$

For $i \ne j$,

$$\left|(\widetilde{A} - \Phi\widetilde{D})_{(i,j)} - (A - \Phi D)_{(i,j)}\right| = \epsilon\,\left|(A - \Phi D)_{(i,j)}\right|, \tag{3.17}$$

$$\left|(\widetilde{B} - \widetilde{D}\Phi)_{(i,j)} - (B - D\Phi)_{(i,j)}\right| = \epsilon\,\left|(B - D\Phi)_{(i,j)}\right| \tag{3.18}$$

which lead to $|(\widetilde{P}_\Phi - P_\Phi) \oslash P_\Phi|_{(i,j)} \le \epsilon$ for $i \ne j$. Now apply [21, Theorem 2.2] to (2.4) and (3.4) to get (3.12) for $i = 1$ and to (2.5) and (3.5) to get (3.12) for $i = 2$.

The case in which $A - \Phi D$ and/or $B - D\Phi$ are reducible can only possibly occur when W is a nonsingular M-matrix because of the conditions of the theorem and Theorem 2.1. Again by Theorem 2.1, $A - \Phi D$ and $B - D\Phi$ are nonsingular. Replace A and B by $A - \xi\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ and $B - \xi\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}$ for sufficiently tiny ξ. Then (3.12) will hold. Letting $\xi \to 0$ yields (3.12) for the case in which $A - \Phi D$ and/or $B - D\Phi$ are reducible.

Finally since $\Phi_2 \le \Phi_1 \le \kappa\Phi$ and $\check{\Phi}_2 \ge \Phi_2$, we have

$$\begin{aligned}0 \le \widetilde{\Phi} - \Phi &= \check{\Phi}_1 - \Phi_1 + \Phi_2 - \check{\Phi}_2 + \Delta \\ &\le \check{\Phi}_1 - \Phi_1 + \Delta \\ &\le \left[2mn\chi\,\kappa\,\epsilon + O(\epsilon^2)\right]\Phi,\end{aligned}$$

as expected. □

Proof of Theorem 2.4 We only prove the case in which $u_1^{\mathrm{T}} v_1 > u_2^{\mathrm{T}} v_2$. By Theorem 2.1, $0 \le \lambda_1 < \lambda_2 = 1$. We still have (2.4), (2.5), (3.4), (3.5), $\widetilde{\Phi} = \check{\Phi}_1 - \check{\Phi}_2 + \Delta$ by (3.7). Instead of (3.11), we will have

$$|\Delta| \le \left[4\gamma^2\epsilon^2 + O(\epsilon^3)\right] \check{\Phi}_2. \tag{3.19}$$

The singularity of W disallows us to make directional perturbations, as in (3.8) and (3.9), while keeping $\widetilde{W}$ an M-matrix, and thus we no longer have $0 \le \Phi_i \le \check{\Phi}_i$. But we still have (3.13), and instead of (3.14)

$$\begin{aligned} \widetilde{P}_{\Phi}(y \otimes u) &\ge [(1-\epsilon) - (1+\epsilon)\lambda_1] y \otimes (D_1 u) \\ &\quad + [(1-\epsilon) - (1+\epsilon)\lambda_2](D_2 y) \otimes u, \qquad (3.20) \\ \widetilde{P}_{\Phi}(y \otimes u) &\le [(1+\epsilon) - (1-\epsilon)\lambda_1] y \otimes (D_1 u) \\ &\quad + [(1+\epsilon) - (1-\epsilon)\lambda_2](D_2 y) \otimes u. \qquad (3.21) \end{aligned}$$

Use a similar argument that led to (3.16) above to get

$$(1 + \tfrac{1}{2}\chi\,\epsilon) P_{\Phi}(y \otimes u) \ge \widetilde{P}_{\Phi}(y \otimes u) \ge (1 - \tfrac{1}{2}\chi\,\epsilon) P_{\Phi}(y \otimes u). \tag{3.22}$$

Also (3.17) and (3.18) with "=" replaced by "$\le$" are valid. By [21, Theorem 2.2], we have for $i = 1, 2$

$$|\check{\Phi}_i - \Phi_i| \le \left[mn\chi\,\epsilon + O(\epsilon^2)\right] \Phi_i.$$

Finally use $|\widetilde{\Phi} - \Phi| \le |\check{\Phi}_1 - \Phi_1| + |\Phi_2 - \check{\Phi}_2| + |\Delta|$ combined with (3.11) to complete the proof. □

Proof of Theorem 2.5 We first consider the case in which $u_1^{\mathrm{T}} \mathbf{1}_m > u_2^{\mathrm{T}} \mathbf{1}_n$. Then,

$$(B - D\Phi)\mathbf{1}_m = (\widetilde{B} - \widetilde{D}\widetilde{\Phi})\mathbf{1}_m = 0,$$

and thus $\lambda_1 < 1$, $\widetilde{\lambda}_1 < 1$ and $\lambda_2 = \widetilde{\lambda}_2 = 1$. By comparing (3.4) to (2.4) and comparing (3.5) to (2.5), it follows from [21, Theorem 3.3] that for $i = 1, 2$

$$|\check{\Phi}_i - \Phi_i| \le \left[mn\chi\,\epsilon + O(\epsilon^2)\right] \Phi_i.$$

Then again use $|\widetilde{\Phi} - \Phi| \le |\check{\Phi}_1 - \Phi_1| + |\Phi_2 - \check{\Phi}_2| + |\Delta|$ combined with (3.19) to complete the proof.

We now consider the case in which $u_1^{\mathrm{T}} \mathbf{1}_m < u_2^{\mathrm{T}} \mathbf{1}_n$. We will reduce this case to the case in which $u_1^{\mathrm{T}} \mathbf{1}_m > u_2^{\mathrm{T}} \mathbf{1}_n$. Let $\widetilde{u}_1, \widetilde{u}_2$ be positive vectors such that

$$[\widetilde{u}_1^{\mathrm{T}}, \widetilde{u}_2^{\mathrm{T}}]\widetilde{W} = 0.$$

Normalize u_1, u_2 and $\widetilde{u}_1, \widetilde{u}_2$ such that $(u_1^{\mathrm{T}}, u_2^{\mathrm{T}})\mathbf{1}_{m+n} = (\widetilde{u}_1^{\mathrm{T}}, \widetilde{u}_2^{\mathrm{T}})\mathbf{1}_{m+n} = 1$. O'Cinneide [17] showed that

$$|(u_1^{\mathrm{T}}, u_2^{\mathrm{T}}) - (\widetilde{u}_1^{\mathrm{T}}, \widetilde{u}_2^{\mathrm{T}})| \leq \left[2(m+n)\epsilon + O(\epsilon^2)\right] (u_1^{\mathrm{T}}, u_2^{\mathrm{T}}). \tag{3.23}$$

$Z = \Phi^{\mathrm{T}}$ is the minimal nonnegative solution of

$$ZD^{\mathrm{T}}Z - ZA^{\mathrm{T}} - B^{\mathrm{T}}Z + C^{\mathrm{T}} = 0.$$

Let $\Theta = U_1^{-1} Z U_2$, where $U_1 = \mathrm{diag}(u_1)$ is the diagonal matrix with $(U_1)_{(i,i)} = (u_1)_{(i)}$ and $U_2 = \mathrm{diag}(u_2)$. Then

$$\Theta(U_2^{-1}D^{\mathrm{T}}U_1)\Theta - \Theta(U_2^{-1}A^{\mathrm{T}}U_2) - (U_1^{-1}B^{\mathrm{T}}U_1)\Theta + U_1^{-1}C^{\mathrm{T}}U_2 = 0.$$

Now define

$$\Omega = \begin{pmatrix} U_2^{-1}A^{\mathrm{T}}U_2 & -U_2^{-1}D^{\mathrm{T}}U_1 \\ -U_1^{-1}C^{\mathrm{T}}U_2 & U_1^{-1}B^{\mathrm{T}}U_1 \end{pmatrix}.$$

Then, $(u_2^{\mathrm{T}}, u_1^{\mathrm{T}})\Omega = 0$, $\Omega\mathbf{1}_{m+n} = 0$. Similarly, we define $\widetilde{U}_1$, $\widetilde{U}_2$, $\widetilde{\Theta}$, and

$$\widetilde{\Omega} = \begin{pmatrix} \widetilde{U}_2^{-1}\widetilde{A}^{\mathrm{T}}\widetilde{U}_2 & -\widetilde{U}_2^{-1}\widetilde{D}^{\mathrm{T}}\widetilde{U}_1 \\ -\widetilde{U}_1^{-1}\widetilde{C}^{\mathrm{T}}\widetilde{U}_2 & \widetilde{U}_1^{-1}\widetilde{B}^{\mathrm{T}}\widetilde{U}_1 \end{pmatrix}.$$

We have $\widetilde{\Omega}\mathbf{1}_{m+n} = 0$. By (3.23), we have

$$|\widetilde{\Omega} - \Omega| \leq \left([4(m+n)+1]\epsilon + O(\epsilon^2)\right) |\Omega|.$$

Applying the result for $u_1^{\mathrm{T}}\mathbf{1}_m > u_2^{\mathrm{T}}\mathbf{1}_n$ we just obtained, we get

$$|\widetilde{\Theta} - \Theta| \leq \left[4mn[4(m+n)+1]\kappa\frac{1+\lambda_2}{1-\lambda_2}\epsilon + O(\epsilon^2)\right]\Theta.$$

Since $\Phi = U_2^{-1}\Theta^{\mathrm{T}}U_1$ and $\widetilde{\Phi} = \widetilde{U}_2^{-1}\widetilde{\Theta}^{\mathrm{T}}\widetilde{U}_1$,

$$|\widetilde{\Phi} - \Phi| \leq \left[4(m+n)\epsilon + 4mn[4(m+n)+1]\kappa\frac{1+\lambda_2}{1-\lambda_2}\epsilon + O(\epsilon^2)\right]\Phi.$$

The proof is completed. □

4 Algorithms for MARE

We shall explain in what follows three types of methods that can deliver computed Φ with entrywise relative accuracies as deserved by the input data. It is assumed throughout this section that W in (1.2) is a nonsingular M-matrix or an irreducible singular M-matrix with $u_1^{\mathrm{T}} v_1 \neq u_2^{\mathrm{T}} v_2$.

4.1 Fixed point iterative methods

Any splittings for A and B:

$$A = M_1 - K_1, \quad B = M_2 - K_2 \tag{4.1}$$

give rise to an iterative method for MARE (1.1): $X_0 = 0$ and for $k \geq 0$

$$M_1 X_{k+1} + X_{k+1} M_2 = X_k D X_k + K_1 X_k + X_k K_2 + C. \tag{4.2}$$

Convenient ones are those from the so-called *regular splittings* (4.1), namely $M_i^{-1} \geq 0$ and $K_i \geq 0$, in such a way that (4.2) is easy to solve and convergent. The following five choices are obvious ones:

$$M_1 = \mathrm{diag}(A), \quad M_2 = \mathrm{diag}(B); \tag{4.3a}$$
$$M_1 = \mathrm{tril}(A), \quad M_2 = \mathrm{triu}(B); \tag{4.3b}$$
$$M_1 = \mathrm{triu}(A), \quad M_2 = \mathrm{tril}(B), \tag{4.3c}$$
$$M_1 = \mathrm{tril}(A), \quad M_2 = \mathrm{tril}(B), \tag{4.3d}$$
$$M_1 = \mathrm{triu}(A), \quad M_2 = \mathrm{triu}(B), \tag{4.3e}$$

and correspondingly $K_1 = M_1 - A$ and $K_2 = M_2 - B$, where $\mathrm{tril}(\cdot)$ and $\mathrm{triu}(\cdot)$ are MATLAB-like notations that take the lower and upper triangular part of a matrix, respectively.

Two pleasant consequences of these splittings in (4.3) are as follows. First, any one of them leads to (4.2) that is easy to solve (see [21, Subsection 4.2]). The second consequence is that any corresponding method (4.2) produces a monotonically convergent sequence to Φ [10, Theorem 2.3]:

$$0 = X_0 \leq X_1 \leq X_2 \leq \cdots, \quad \lim_{i \to \infty} X_i = \Phi. \tag{4.4}$$

Equation (4.2) is an M-Matrix Sylvester equation [21]. A straightforward implementation of (4.2) associated with any of the splittings in (4.3) easily preserves $X_k \geq 0$, but may not numerically preserve the monotonicity in (4.4). There is a better way. From (4.2) for two consecutive steps, we have

$$\begin{aligned} M_1 \Delta_{k+1} + \Delta_{k+1} M_2 &= X_{k+1} D X_{k+1} - X_k D X_k + K_1 \Delta_k + \Delta_k K_2 \\ &= \Delta_k D X_{k+1} + X_k D \Delta_k + K_1 \Delta_k + \Delta_k K_2 \qquad (4.5a) \\ &= X_{k+1} D \Delta_k + \Delta_k D X_k + K_1 \Delta_k + \Delta_k K_2, \qquad (4.5b) \end{aligned}$$

where $\Delta_k = X_{k+1} - X_k$. We therefore suggest to implement (4.2) as follows:

Algorithm 1
Fix Point Iterative Method for MARE $XDX - AX - XB + C = 0$ **with** (4.1).
1 Solve $M_1X_1 + X_1M_2 = C$ for X_1;
2 $\Delta_0 = X_1$;
3 For $k = 0, 1, \ldots$, until convergence
4 Solve either (4.5a) or (4.5b) for Δ_{k+1};
5 $X_{k+2} = X_{k+1} + \Delta_{k+1}$;
6 Enddo.

With each of the splittings in (4.3), Algorithm 4.1 is guaranteed to produce a *linearly* convergent sequence of X_k [10]. Since all involved arithmetic operations are adding two nonnegative numbers, dividing a nonnegative number by a positive number, or multiplying two nonnegative numbers, Algorithm 4.1 is forward stable. Thus at convergence, the converged X_k is entrywise relatively accurate, unless the required number of steps so gargantuan that the accumulated roundoff errors become too great to overcome.

At Line 4, there is a choice of solving (4.5a) or (4.5b). Under what circumstances to favor one over the other is not known at this time. It is probably either one will work just fine.

In our numerical tests, we use if $\max_{i,j} |(X_{k+1} - X_k) \oslash X_{k+1}|_{(i,j)} \leq \epsilon$ to terminate the iteration at Line 3. For a justification, see [21, Item 4 in Remark 4.1].

For the ease of later references, we will use `FPa`, `FPb`, `FPc`, `FPd`, and `FPe` to denote Algorithm 4.1 combined with the respective splittings (4.3a)–(4.3e).

4.2 Structure-preserving doubling algorithm for MARE

The basic idea of doubling algorithms in an iterative scheme is to compute only the 2^kth approximations, instead of every approximation in the process. It traces back to 1970s (see [2] and references therein). Recent resurgence of interests in the idea has led to efficient doubling algorithms for various nonlinear matrix equations. The interested reader is referred to [5] for a more general presentation. The use of a structure-preserving doubling algorithm (SDA) to solve an MARE was first proposed and analyzed by Guo, Lin, and Xu [13]. For MARE (1.1), the method simultaneously computes the minimal nonnegative solutions of (1.1) and its complementary M-Matrix Algebraic Riccati Equation (cMARE)

$$YCY - YA - BY + D = 0. \tag{4.6}$$

The method can be naturally regarded as an extension of the Smith algorithm [19] adopted by [21] for an MSE. In fact, it degenerates to the Smith algorithm when $D = 0$. It starts by choosing $\mu > 0$ to satisfy

$$\mu \geq \mu_{\text{opt}} \stackrel{\text{def}}{=} \max\{\max_i A_{(i,i)}, \max_j B_{(j,j)}\}. \tag{4.7}$$

Let

$$E_0 = V_\mu^{-1}[(B-\mu I) - DA_\mu^{-1}C], \quad F_0 = U_\mu^{-1}[(A-\mu I) - CB_\mu^{-1}D], \tag{4.8a}$$
$$X_0 = 2\mu U_\mu^{-1} C B_\mu^{-1}, \quad Y_0 = 2\mu B_\mu^{-1} D U_\mu^{-1}, \tag{4.8b}$$

where

$$A_\mu = A + \mu I_n, \qquad B_\mu = B + \mu I_m, \tag{4.9a}$$
$$U_\mu = A_\mu - CB_\mu^{-1}D, \quad V_\mu = B_\mu - DA_\mu^{-1}C. \tag{4.9b}$$

SDA then produces the sequences $\{E_k\}$, $\{X_k\}$, $\{Y_k\}$ and $\{F_k\}$ by

$$E_{k+1} = E_k(I_m - Y_k X_k)^{-1} E_k, \tag{4.10a}$$
$$F_{k+1} = F_k(I_n - X_k Y_k)^{-1} F_k, \tag{4.10b}$$
$$X_{k+1} = X_k + F_k(I_n - X_k Y_k)^{-1} X_k E_k, \tag{4.10c}$$
$$Y_{k+1} = Y_k + E_k(I_m - Y_k X_k)^{-1} Y_k F_k. \tag{4.10d}$$

As is, two inverses $(I_n - X_k Y_k)^{-1}$ and $(I_m - Y_k X_k)^{-1}$ need to be computed per step. But we can use the Sherman–Morrison–Woodbury formula [6, p. 95] to get rid of one of them:

$$(I_m - Y_k X_k)^{-1} = I_m + Y_k(I_n - X_k Y_k)^{-1} X_k,$$
$$(I_n - X_k Y_k)^{-1} = I_n + X_k(I_m - Y_k X_k)^{-1} Y_k.$$

So alternatively

$$E_{k+1} = E_k[I_m + Y_k(I_n - X_k Y_k)^{-1} X_k] E_k, \tag{4.10a$'$}$$
$$F_{k+1} = F_k[I_n + X_k(I_m - Y_k X_k)^{-1} Y_k] F_k, \tag{4.10b$'$}$$
$$X_{k+1} = X_k + F_k X_k (I_m - Y_k X_k)^{-1} E_k, \tag{4.10c$'$}$$
$$Y_{k+1} = Y_k + E_k Y_k (I_n - X_k Y_k)^{-1} F_k. \tag{4.10d$'$}$$

This leads to three sets of formulas for advancing the iteration:

1. (4.10a), (4.10b), (4.10c), and (4.10d);
2. (4.10a), (4.10b′), (4.10c′), and (4.10d);
3. (4.10a′), (4.10b), (4.10c), and (4.10d′).

As to under what circumstances to favor one over another, we offer this rough suggestion. Since two additional matrix-matrix multiplications have to be done in order to get rid of one of the inverses, we should prefer the first set if $m \approx n$. But if either $m \ll n$ or $m \gg n$, not computing the inverse of the larger one in size between $I_n - X_k Y_k$ and $I_m - Y_k X_k$ should be preferred and thus to use the second set if $m \ll n$, and the third set if $m \gg n$. Algorithm 2 below picks the first set for the sake of presentation.

It is shown in [13] (for nonsingular W with strict inequality in (4.7)) and in [12] (for irreducible singular W) that for W satisfying (1.4)

$$0 \le X_1 \le X_2 \le \cdots, \quad \lim_{k\to\infty} X_k = \Phi,$$
$$0 \le Y_1 \le Y_2 \le \cdots, \quad \lim_{k\to\infty} Y_k = \Psi,$$

where Ψ is the minimal nonnegative solution of cMARE (4.6), and $I_m - Y_k X_k$ and $I_n - X_k Y_k$ are nonsingular M-matrices for all k, and

$$0 \le \Phi - X_k \le L_\mu^{2^k} \Phi R_\mu^{2^k}, \quad 0 \le \Psi - Y_k \le R_\mu^{2^k} \Psi L_\mu^{2^k},$$

where

$$R_\mu = (B - D\Phi + \mu I_m)^{-1}(B - D\Phi - \mu I_m),$$
$$L_\mu = (A - C\Psi + \mu I_n)^{-1}(A - C\Psi - \mu I_n).$$

So the convergence is quadratic if $u_1^{\mathrm{T}} v_1 \ne u_2^{\mathrm{T}} v_2$ [5,12,13]. It is proved in [12, Theorem 4.4] that both $\rho(L_\mu)$ and $\rho(R_\mu)$ are less than 1 and are nondecreasing functions in μ satisfying (4.7).

The algorithm is now given below. Afterwards we'll comment on its implementation detail.

Algorithm 2
SDA for MARE $XDX - AX - XB + C = 0$ **and,**
as a by-product, for cMARE $YCY - YA - BY + D = 0$.

1 Pick μ such that $\mu \ge \mu_{\mathrm{opt}}$;
2 $A_\mu \stackrel{\mathrm{def}}{=} A + \mu I$, $B_\mu \stackrel{\mathrm{def}}{=} B + \mu I$;
3 Compute A_μ^{-1} and B_μ^{-1};
4 Compute V_μ and U_μ as in (4.9b) and then their inverses;
5 Compute E_0, F_0, X_0, and Y_0 as in (4.8);
6 Compute $(I - X_0 Y_0)^{-1}$;
7 Compute X_1 and Y_1 by (4.10c) and (4.10d);
8 For $k = 1, 2, \ldots$, until convergence
9 Compute E_k and F_k by (4.10a) and (4.10b) (after substituting $k+1$ for k);
10 Compute $(I - X_k Y_k)^{-1}$ and $(I - Y_k X_k)^{-1}$;
11 Compute X_{k+1} and Y_{k+1} by (4.10c) and (4.10d);
12 Enddo

Remark 4.1 When the input W is in the usual matrix format and the algorithm is implemented straightforwardly as is with all inverses (of M-matrices) calculated by, e.g., the Gaussian elimination (with partial pivoting), Algorithm 4.2 is the same as the original version in [13]. This version usually works well, as the numerical examples in [13] showed. But as discovered in [1], when it comes to an M-matrix, it pays to have its triplet representation which, if known to have entrywise accuracy, can produce

its inverse with comparable entrywise accuracy. We now comment on how the relevant lines in the above algorithm can be implemented differently for better numerical accuracy.

1. If input $W \in \mathbb{R}^{(m+n)\times(m+n)}$ (a nonsingular or an irreducible singular M-matrix) is given in its triplet form $W = \left\{N, \begin{pmatrix} y \\ u \end{pmatrix}, \begin{pmatrix} z \\ v \end{pmatrix}\right\}$ with $y, z \in \mathbb{R}^m$, $u, v \in \mathbb{R}^n$, and $y, z, u, v > 0$, then

$$\begin{aligned} A &= \{N_1, u, v + Cy\}, & B &= \{N_2, y, z + Du\}, \\ A_\mu &= \{N_1, u, v + Cy + \mu u\}, & B_\mu &= \{N_2, y, z + Du + \mu y\}. \end{aligned}$$

   Consequently, A_μ^{-1} and B_μ^{-1} at Line 3 can be computed using the GTH-like algorithm [1].
2. Line 1 asks $\mu \geq \mu_{\text{opt}}$. The bigger the μ, the less likely some catastrophic cancelations may occur in computing the diagonal entries of $A - \mu I$ and $B - \mu I$ in (4.8a). But, as a tradeoff, the whole convergence is slowed [12, Theorem 4.4]. As we discussed the same issue in [21] for the Smith method, when all $A_{(i,i)}$ and $B_{(j,j)}$ are known to be exact floating point numbers, taking $\mu = \mu_{\text{opt}}$ is fine; otherwise we may take $\mu = \eta \times \mu_{\text{opt}}$ to avoid any catastrophic cancelation for some $\eta > 1$ but not too close to 1.
3. At Line 3, if the triplet representation for A_μ and B_μ are available, use the GTH-like algorithm [1]. Otherwise refer to [21, Subsection 2.2] to compute their inverses with guaranteed entrywise relative errors about in the orders of

$$\left[1 - \varrho(A)\right]^{-1} \boldsymbol{u}, \quad \left[1 - \varrho(B)\right]^{-1} \boldsymbol{u},$$

   respectively, where $\boldsymbol{u}$ is unit machine roundoff.
4. At Lines 4, 6 and 10, refer to [21, Subsection 2.2] for computing the inverses. The computed inverse Z^{-1} will have entrywise relative error about $O([1-\varrho(Z)]^{-1}\,\boldsymbol{u})$, where Z denotes any of the matrices to be inverted in the lines. Specifically at Line 4, $\varrho(A_\mu) \leq \varrho(A) \leq \lambda_1$ and $\varrho(B_\mu) \leq \varrho(B) \leq \lambda_2$. At Lines 6 and 10, since M-matrices $I - X_k Y_k \leq I - \Phi\Psi$ and $I - Y_k X_k \leq I - \Psi\Phi$ [12,13],

$$\varrho(I - X_k Y_k) \leq \varrho(I - \Phi\Psi), \quad \varrho(I - Y_k X_k) \leq \varrho(I - \Psi\Phi).$$

   Thus the entrywise relative errors in the final computed Φ and Ψ can be bounded by something in the order of

$$\left(\kappa\chi + [1 - \varrho(I - \Phi\Psi)]^{-1} + [1 - \varrho(I - \Psi\Phi)]^{-1}\right)\boldsymbol{u}.$$

5. At Line 10, we need triplet representations for $I - X_k Y_k$ and $I - Y_k X_k$ in order to use the GTH-like algorithm to accurately compute their inverses. We can certainly use the idea in [21, Subsection 2.2] for this purpose. But as convergence begins to occur, we may be able to achieve some saving by utilizing the triplet representations from the previous step. For example, suppose that we have $I - X_k Y_k =$

$\{N, u, v\}$ and that X_k and Y_k already have converged to Φ and Ψ to certain degree. Then u likely is suitable as part of a triplet representation for $I - X_{k+1}Y_{k+1}$, too. Its suitability can be easily verified by testing if $(I - X_{k+1}Y_{k+1})u \geq 0$.

6. At Line 8, the same stopping criterion for the modified Smith algorithm as discussed in [21, Remark 4.1] can be used here, too.

4.3 Newton method

The Newton method played an essential role in [9–11] for solving (1.1) and establishing its many properties. It goes as follows: $X_0 = 0$ and for $k \geq 1$

$$(A - X_k D)X_{k+1} + X_{k+1}(B - DX_k) = C - X_k D X_k. \tag{4.11}$$

If W in (1.2) is a nonsingular M-matrix or an irreducible singular M-matrix, it is shown in [9] that

$$0 = X_0 \leq X_1 \leq X_2 \leq \cdots \leq \Phi, \quad \lim_{k\to\infty} X_k = \Phi. \tag{4.12}$$

As far as its implementation is concerned, solving (4.11) may fail this monotonic property. A better formula can be gotten from subtracting (4.11) from it for the next step to get

$$(A - X_{k+1}D)\Delta_{k+1} + \Delta_{k+1}(B - DX_{k+1}) = \Delta_k D \Delta_k, \tag{4.13}$$

where $\Delta_k = X_{k+1} - X_k$. The monotonic property in (4.12) is also an obvious consequence of (4.13), too, after noticing that

$$P_{X_k} = I_n \otimes (A - X_k D) + (B - DX_k)^{\mathrm{T}} \otimes I_m$$

is a nonsingular M-matrix [9]. We also have

$$(A - X_k D)(\Phi - X_{k+1}) + (\Phi - X_{k+1})(B - DX_k) = (\Phi - X_k)D(\Phi - X_k) \tag{4.14}$$

which suggests that the Newton method is eventually quadratically convergent.

Equation (4.13) is an MSE. It must be solved carefully in order to make sure $\Delta_{k+1} \geq 0$ and thus preserve the monotonic property in (4.12). For example, the commonly used methods in [3,8] may fail in that aspect. However, any of the methods discussed in [21, Section 4] will work. Since the direct method in [21] there is too expensive even for modest m and n, and the iterative methods in [21] may take as many steps in order to solve (4.13) accurately[6] as the methods in subsections 4.1 and 4.2 in this paper, it is usually cheaper to solve an MARE directly by them.

[6] The correctness of (4.13) relies upon X_k being the kth Newton approximation (to the working precision). If (4.13) is not solved accurately enough, then everything, including (4.12), proved for the Newton method may no longer be valid.

5 Numerical examples

In this section, we shall present two numerical examples to test our entrywise perturbation bounds as well the ability of the numerical methods in Sect. 4 to deliver entrywise relative accurate numerical solutions as claimed. We will use two error measures to gauge accuracy in computed solution $\widehat{\Phi}$: the Normalized Residual (NRes)

$$\texttt{NRes} = \frac{\|\widehat{\Phi}D\widehat{\Phi} - A\widehat{\Phi} - \widehat{\Phi}B + C\|_1}{\|\widehat{\Phi}\|_1(\|\widehat{\Phi}\|_1\|D\|_1 + \|A\|_1 + \|B\|_1) + \|C\|_1},$$

a commonly used measure in general because it can be easily computed, and the entrywise relative error (ERErr),

$$\texttt{ERErr} = \max_{i,j} |(\widehat{\Phi} - \Phi) \oslash \Phi|_{(i,j)}$$

which is not available in actual computations but is made available here for our testing purpose. Both errors are 0 for the exact solution, but numerically they can only be made as small as $O(\boldsymbol{u})$. As we will see, to achieve $\widehat{\Phi}$ with deserved entrywise relative accuracy, tiny NRes (as tiny as $O(\boldsymbol{u})$) is not sufficient.

Example 5.1 ([13, Example 6.2]) $A, B, C, D \in \mathbb{R}^{n\times n}$ are given by

$$A = \begin{pmatrix} 3 & -1 & & \\ & 3 & \ddots & \\ & & \ddots & -1 \\ -1 & & & 3 \end{pmatrix}, \quad B = A, \quad C = I_n, \quad D = \xi I_n,$$

where ξ is a parameter that modulates the quadratic term in MARE (1.1). If $\xi = 0$, it becomes the MSE example in [21]. The numerical results below indicate close resemblance between the MSE and this MARE in their solution behavior. It can be seen that

$$A\mathbf{1}_n = 2\mathbf{1}_n, \quad \mathbf{1}_n^{\mathrm{T}}B = 2\mathbf{1}_n^{\mathrm{T}}.$$

We consider[7] $\xi = 0.2$ and $m = n = 100$. For testing purpose, we computed an "exact" solution Φ and Ψ by the computerized algebra system *Maple* with 100 decimal digits. This "exact" solution Φ's entries range from 10^{-43} to 0.17 and Ψ's entries range from 2.2×10^{-44} to 0.03. SDA with Kahan's stopping criterion works extremely well: in just 7 iterations, it produces $\widehat{\Phi}$ with an entrywise relative error 1.9×10^{-14} and $\widehat{\Psi}$ with 3.8×10^{-15}. But it takes rather long for the fixed point iterations, except for FPe which is based on splitting A and B into their upper triangular part and strictly lower

[7] We tried several other values of ξ. Note by Lemma 3.1 that as ξ increases $\min_{i,j} \Phi_{(i,j)}$ increases, and qualitative behaviors reported for $\xi = 0.2$ in this example seem to hold for other $\xi > 0$ as well.

triangular part $e_n e_1^{\mathrm{T}}$ and fairly fast because of that. The Newton method combined with `lyap` on (4.13) gives

$$\texttt{NRes} = 8.6 \times 10^{-15}, \quad \text{but} \quad \texttt{ERErr} = 2.1 \times 10^{+27}$$

indicating no relative accuracy at all for some of the entries in the computed $\widehat{\Phi}$ because `lyap` cannot solve (4.13) with high relative accuracy for tiny solution entries.

Figure 1 displays the convergence history for SDA and the fixed point iterations for Φ. The curves for `NRes` look very nice—noticeable drops every steps; the curves for entrywise relative errors, however, show very little improvements for the first many iterations, especially so for `FPa` and `FPb`. Notice that `NRes` reach to about $O(10^{-16})$ before all entries of Φ are converged with their deserved accuracy—especially so for the fixed point iterations, for example at Iteration 25, `FPb` has `NRes` about $O(10^{-16})$ but some entries in the computed $\widehat{\Phi}$ do not even have one decimal digit correct!

Next we relatively perturb each entries of A, B, C, and D to illustrate the effectiveness of our perturbation bounds. We still take $n = 100$ for which we have the "exact" solution to compare to. In MATLAB, each nonzero entry in A, B, C, and D is multiplied by

$$1 + (\texttt{rand} - .5) * \Gamma * \texttt{eps}, \tag{5.1}$$

where Γ is an adjustable parameter. We then compute the solution $\widetilde{\Phi}$ of the perturbed MARE by SDA with Kahan's stopping criterion and use this solution as the "true" solution of the perturbed MARE. Let ϵ be the smallest one to satisfy (2.2). Figure 2 plots the entrywise relative errors in $\widetilde{\Phi}$ compared to Φ as ϵ varies. To explain this figure, we have computed

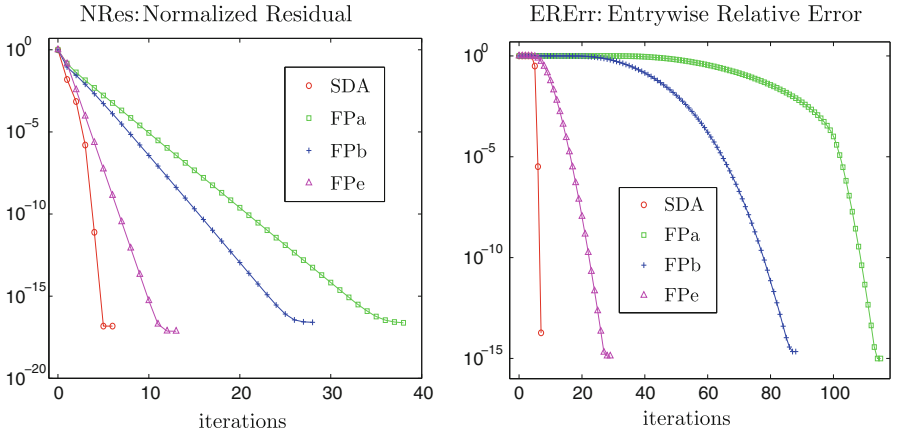

Fig. 1 Example 5.1, $n = 100$. Convergence history for SDA and the fixed point iterations. *Left* `NRes`; *Right* entrywise relative errors. Curves for `FPc` and `FPd` are indistinguishable from those for `FPb` and `FPa`, respectively, and thus not plotted

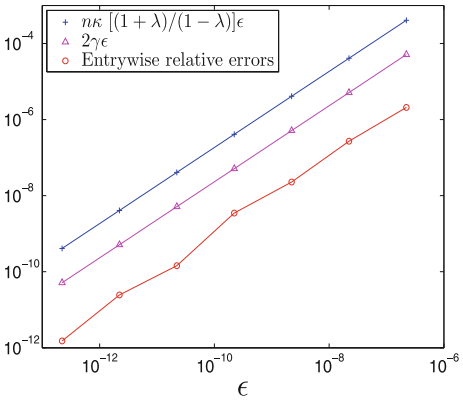

Fig. 2 Example 5.1, $n = 100$. Entrywise relative errors in $\widetilde{\Phi}$ as ϵ varies. It is estimated via a least squares fit that true entrywise relative errors (marked by *circles*) behave like 9.5ϵ

λ_1	λ_2	γ	κ	$\varrho(I - \Phi\Psi)$	$\varrho(I - \Psi\Phi)$
0.3502	0.3502	115.8	8.892	7.245×10^{-3}	7.245×10^{-3}

that have appeared in our error analysis, except $\varrho(I-\Phi\Psi)$ and $\varrho(I-\Psi\Phi)$ which control the relative entrywise accuracies in computed $(I - X_kY_k)^{-1}$ and $(I - Y_kX_k)^{-1}$ in SDA. That both $\varrho(I - \Phi\Psi)$ and $\varrho(I - \Psi\Phi)$ are so tiny suggests all inverses $(I - X_kY_k)^{-1}$ and $(I - Y_kX_k)^{-1}$ should have been computed very accurately. Since $\lambda_1 \approx \lambda_2$, the error bounds by Theorems 2.2 and 2.3 are, up to the 1st order term,

$$2\gamma\,\epsilon, \quad 2n^2\kappa\,\frac{1+\lambda}{1-\lambda}\,\epsilon.$$

They are plotted in Fig. 2 after the second expression is reduced by a factor $2n$.

Example 5.2 This example is for the singular irreducible case constructed as follows. Let $\widehat{A} = \widehat{B} \in \mathbb{R}^{n\times n}$ be the A in Example 5.1, and let $a, b \in \mathbb{R}^n$ be two positive vectors with random entries as obtained by MATLAB's `round(10^5*rand)`. These a, b are then saved in order to repeat the test. Finally set

$$W = \begin{pmatrix} \mathrm{diag}(b) & \\ & \mathrm{diag}(a) \end{pmatrix} \begin{pmatrix} \widehat{B} & -2I_n \\ -2I_n & \widehat{A} \end{pmatrix}.$$

We see $W\mathbf{1}_{2n} = 0$. The random positive integer vectors so constructed serve two purposes here:

1. The resulting MARE can be moved, without any errors, to *Maple* for computing an "exact" Φ for testing purpose.
2. W is a singular irreducible M-matrix with

$$u_1^{\mathrm{T}}v_1 = 1.0872 \times 10^{-2} > u_2^{\mathrm{T}}v_2 = 9.1993 \times 10^{-3}.$$

Finally the coefficient matrices of MARE (1.1) can be read off from W. Our results below are for $m = n = 100$. We compute its "exact" Φ by *Maple*. We find Φ's entries

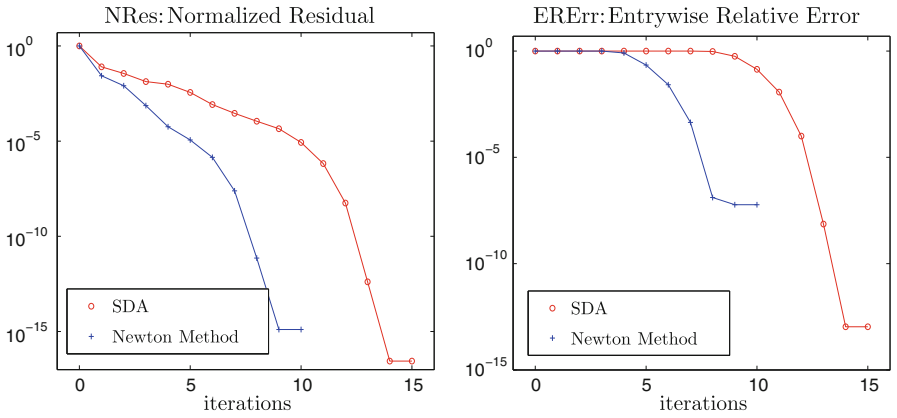

Fig. 3 Example 5.2, $n = 100$. Convergence history for SDA and the Newton method with MATLAB's `lyap` for involved MSEs. *Left* `NRes`; *Right* entrywise relative errors

range from 8.6×10^{-9} to 7.0×10^{-1}. So unlike Example 5.1, we expect the Newton method combined with `lyap` be able to compute Φ with an entrywise relative accuracy about 10^{-7}. Indeed this is what we got.

Figure 3 displays the convergence history for SDA and the Newton method with `lyap` on (4.13). It shows that the Newton method actually converges faster but returns less accurate results at the end as expected: 1.3×10^{-15} for `NRes` and 6.0×10^{-8} for `ERErr`. On the other hand, SDA computes a solution with 2.8×10^{-17} for `NRes` and 1.1×10^{-13} for the entrywise relative error.

To understand this example's numerical behavior, we have computed

λ_1	λ_2	γ	κ	$\varrho(I - \Phi\Psi)$	$\varrho(I - \Psi\Phi)$
0.9486	1.000	810.6	270.7	0.8369	0.8378

These numbers tell us that

1. $A - \Phi D$ is a nonsingular M-matrix and $B - D\Phi$ is a singular M-matrix as they should be.
2. All the inverses $(I - X_k Y_k)^{-1}$ and $(I - Y_k X_k)^{-1}$ within SDA are computed with the entrywise relative error no bigger than $O([1 - 0.84]^{-1}\boldsymbol{u})$.

Now suppose some of the last bits in the nonzero entries of W is perturbed. This gives[8] $\epsilon = 2\boldsymbol{u}$. Thus Theorems 2.2 and 2.5 say that such perturbations will introduce entrywise relative changes to Φ, up to the first order, by no more than

$$2\gamma\, 2\boldsymbol{u} = 3.60 \times 10^{-13}, \quad 4n^2\kappa \frac{1 + \lambda_1}{1 - \lambda_1} 2\boldsymbol{u} = 9.11 \times 10^{-8}.$$

Some comments are in order. First, the accuracy in the computed $\widehat{\Phi}$ by SDA is simply remarkable; Second, the factor $4n^2$, as we pointed out a couple times before, is again

[8] MATLAB's `eps` is actually $2\boldsymbol{u}$ for the IEEE double precision.

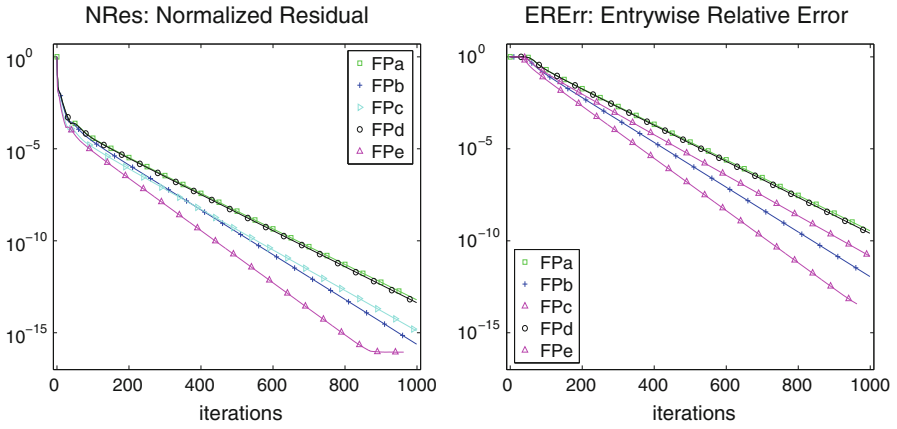

Fig. 4 Example 5.2, $n = 100$. Convergence history for the fixed point iterations. *Left* NRes; *Right* entrywise relative errors

showing its probable overestimate as an artifact of our proofs. After dropping a factor of $4n$, the second first order error bound above gives 2028×10^{-10}, still two magnitudes larger than the first one. This, as well as Example 5.1, echo the comments we made at the beginning of section 2 that the standard first order error analysis produces sharp error estimations.

Unlike in Example 5.1, all fixed point iterations are slowly convergent. With the maximum number of iterations caped at 1000, Figure 4 displays their convergence history. It suggests that the fixed point iterations are too slow to be competitive for this example. □

6 Concluding remarks

We have presented an entrywise perturbation analysis for M-Matrix algebraic Riccati equations (1.1). It is proved small relative perturbations to the entries of A, B, C, and D will only cause small relative changes to each entry of the solution Φ, regardless of its magnitude.

The linear term in our bound from the standard first order error analysis is sharp as shown by Proposition 2.2. The linear terms in our bounds from our second approach have constant factors that are larger than necessary due to the artifact of our proofs. Specifically these constant factors contain a dimensionally dependent factor mn which is the product of our earlier entrywise perturbation analysis for M-matrix Sylvester equation (1.3) in [21] where we conjectured that it could be replaced by something like $m+n$. In spite of being less sharp, we argue that the value of our second approach lies in its ability to demonstrate the effect of the spectral radii λ_i on Φ's sensitivity.

We demonstrated that the fixed point iterations [10] and the Newton method [9,11] with some minor but crucial implementation changes can deliver computed solutions with predicted entrywise relative accuracy according to our analysis. The SDA [13] worked very well in all our numerical tests. We argued in Item 4 of Remark 4.1 that

the entrywise accuracy of computed Φ by SDA depends on $[1-\varrho(I-\Phi\Psi)]^{-1}$ and $[1-\varrho(I-\Psi\Phi)]^{-1}$ being not too big. It would be interesting to develop some bounds on them in terms of λ_i.

Since $XDX - AX - XB + C = 0$ has the same solution(s) as

$$XDX - (A - \tau I)X - X(B + \tau I) + C = 0$$

for any scalar τ. Most results in this paper can be modified to cover the case in which there exists $\tau \in \mathbb{R}$ such that

$$W_\tau \stackrel{\text{def}}{=} \begin{pmatrix} B + \tau I & -D \\ -C & A - \tau I \end{pmatrix}$$

is a nonsingular M-matrix or an irreducible singular M-matrix. Whether such a case would occur in any practical application is unknown to us.

Acknowledgments Xue is supported in part by the National Science Foundation of China Grant 10971036 and Laboratory of Mathematics for Nonlinear Science, Fudan University. Xu is supported in part by the National Science Foundation of China Grant 10731060. Li is supported in part by the National Science Foundation Grant DMS-0810506. The authors wish to thank the anonymous referees for their many helpful comments and suggestions.

References

1. Alfa, A.S., Xue, J., Ye, Q.: Accurate computation of the smallest eigenvalue of a diagonally dominant M-matrix. Math. Comp. **71**, 217–236 (2002)
2. Anderson, B.D.O.: Second-order convergent algorithms for the steady-state Riccati equation. Int. J. Control **28**(2), 295–306 (1978)
3. Bartels, R.H., Stewart, G.W.: Algorithm 432: the solution of the matrix equation $AX - BX = C$. Commun. ACM **8**, 820–826 (1972)
4. Benner, P., Li, R.-C., Truhar, N.: On ADI method for Sylvester equations. J. Comput. Appl. Math. **233**(4), 1035–1045 (2009)
5. Chiang, C.-Y., King-Wah Chu, E., Guo, C.-H., Huang, T.-M., Lin, W.-W., Xu, S.-F.: Convergence analysis of the doubling algorithm for several nonlinear matrix equations in the critical case. SIAM J. Matrix Anal. Appl. **31**(2), 227–247 (2009)
6. Demmel, J.: Applied Numerical Linear Algebra. SIAM, Philadelphia (1997)
7. Gohberg, I., Koltracht, I.: Mixed, componentwise, and structured condition numbers. SIAM J. Matrix Anal. Appl. **14**(3), 688–704 (1993)
8. Golub, G.H., Nash, S., Van Loan, C.F.: Hessenberg–Schur method for the problem $AX + XB = C$. IEEE Trans. Autom. Control AC **24**, 909–913 (1979)
9. Guo, C., Higham, N.: Iterative solution of a nonsymmetric algebraic Riccati equation. SIAM J. Matrix Anal. Appl. **29**, 396–412 (2007)
10. Guo, C.-H.: Nonsymmetric algebraic Riccati equations and Wiener-Hopf factorization for M-matrices. SIAM J. Matrix Anal. Appl. **23**, 225–242 (2001)
11. Guo, C.-H., Laub, A.J.: On the iterative solution of a class of nonsymmetric algebraic Riccati equations. SIAM J. Matrix Anal. Appl. **22**, 376–391 (2000)
12. Guo, C.-H., Iannazzo, B., Meini, B.: On the doubling algorithm for a (shifted) nonsymmetric algebraic Riccati equation. SIAM J. Matrix Anal. Appl. **29**(4), 1083–1100 (2007)
13. Guo, X., Lin, W., Xu, S.: A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation. Numer. Math. **103**, 393–412 (2006)
14. Juang, J.: Existence of algebraic matrix Riccati equations arising in transport theory. Linear Algebra Appl. **230**, 89–100 (1995)

15. Juang, J., Lin, W.-W.: Nonsymmetric algebraic Riccati equations and Hamiltonian-like matrices. SIAM J. Matrix Anal. Appl. **20**(1), 228–243 (1998)
16. Lin, Y., Wei, Y.: Normwise, mixed and componentwise condition numbers of nonsymmetric algebraic Riccati equations. J. Appl. Math. Comput. **27**, 137–147 (2008)
17. O'Cinneide, C.A.: Entrywise perturbation theory and error analysis for Markov chains. Numer. Math. **65**, 109–120 (1993)
18. Rogers, L.: Fluid models in queueing theory and Wiener–Hopf factorization of Markov chains. Ann. Appl. Probab. **4**, 390–413 (1994)
19. Smith, R.A.: Matrix equation $XA + BX = C$. SIAM J. Appl. Math. **16**(1), 198–201 (1968)
20. Varga, R.S.: Matrix Iterative Analysis. Prentice-Hall, Englewood Cliffs (1962)
21. Xue, J., Xu, S., Li, R.-C.: Accurate solutions of M-matrix Sylvester equations. Numer. Math. (2011). doi:10.1007/s00211-011-0420-1