Numer. Math. (2012) 120:639–670
DOI 10.1007/s00211-011-0420-1

Numerische Mathematik

Accurate solutions of M-matrix Sylvester equations

Jungong Xue · Shufang Xu · Ren-Cang Li

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

Abstract This paper is concerned with a relative perturbation theory and its entrywise relatively accurate numerical solutions of an M-matrix Sylvester equation $AX + XB = C$ by which we mean both A and B have positive diagonal entries and nonpositive off-diagonal entries and $P = I_m \otimes A + B^{\mathrm{T}} \otimes I_n$ is a nonsingular M-matrix, and C is entrywise nonnegative. It is proved that small relative perturbations to the entries of A, B, and C introduce small relative errors to the entries of the solution X. Thus the smaller entries of X do not suffer bigger relative errors than its larger entries, unlikely the existing perturbation theory for (general) Sylvester equations. We then discuss some minor but crucial implementation changes to three existing numerical methods so that they can be used to compute X as accurately as the input data deserve.

Mathematics Subject Classification (2000) 15A24 · 65F05 · 65F10 · 65G99

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,
P.O. Box 19408, Arlington, TX 76019, USA
e-mail: rcli@uta.edu

1 Introduction

An n-by-n real matrix A is called an *M-matrix* if it can be written as $A = \gamma I_n - E$ such that $\gamma \geq \rho(E)$, where E is n-by-n and entrywise nonnegative, I_n is the $n \times n$ identity matrix, and $\rho(\cdot)$ is the spectral radius of a matrix. It is called a *nonsingular M-matrix* if $\gamma > \rho(E)$ and a *singular M-matrix* if $\gamma = \rho(E)$. Necessarily, an M-matrix A has nonpositive off-diagonal entries and nonnegative diagonal entries. For a nonsingular or irreducible M-matrix A, its diagonal entries are positive.

In this paper, we are concerned with the following Sylvester equation

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

where both $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$ have positive diagonal entries and nonpositive off-diagonal entries and

$$P = I_m \otimes A + B^{\mathrm{T}} \otimes I_n \tag{1.2}$$

is a nonsingular M-matrix, and $C \in \mathbb{R}^{n\times m}$ is entrywise nonnegative. Here and in what follows, $\otimes$ is the usual Kronecker product of matrices (or vectors regarded as matrices). We call this type of Sylvester equation (1.1) an *M-Matrix Sylvester Equation* (MSE).

MSEs appear frequently in iterative methods for M-matrix Algebraic Riccati equations. See [13–16,19,20,23] and the references therein. M-matrix Lyapunov equations, i.e., $B = A^{\mathrm{T}}$, arise in positive systems [27].

An MSE always has a unique solution. Our first goal in this paper is to present an entrywise relative perturbation analysis for MSE (1.1). Specifically, we seek bounds on the entrywise relative errors in the solution caused by small entrywise relative perturbations to the coefficient matrices A, B, and C. 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.

Existing perturbation theory for Sylvester equations [17] is a general theory for all Sylvester equations with nonsingular P and with arbitrary additive perturbations that are tiny, by which we mean, for example, $\|\widetilde{A} - A\|/\|A\|$ is tiny, where $\|\cdot\|$ is some matrix norm and $\widetilde{A}$ is a perturbed A. The conclusion is that, roughly speaking, the change to the solution X measured in the norm is about $\|P^{-1}\|(\|\widetilde{A} - A\| + \|\widetilde{B} - B\| + \|\widetilde{C} - C\|)$, modulo some constant factor. Such a result may not work well in our case because it fails to tell the actual accuracies of the tiny entries in X. Our analysis, taking full advantage of the special structure in an MSE, suggests that all entries of X, regardless of their magnitudes, are determined to comparable relative accuracy as the input data.

Our second goal is to provide algorithms that are able to deliver computed solutions of an MSE as accurate as the data deserve. These include the Smith algorithm (with a suitable shift) [24] and the classical fixed point iteration methods based on the so-called *regular splitting* [26], both however with some minor but crucial implementation changes, and the GTH-like algorithm [1]. This contrasts favorably to other existing methods, including the Bartels–Stewart algorithm [5] and the Golub–

Nash–Van Loan algorithm [11] that use the Schur forms and/or Hessenberg reduction forms of A and B via orthogonal similarity transformations and including ADI methods [6,28] whose shifts may not always be larger or equal to the largest diagonal entries in A and B. Both the Bartels–Stewart algorithm and the Golub–Nash–Van Loan algorithm are backward stable in the normwise sense but cannot produce solutions with deserving entrywise relative accuracy.

The MSE is closely related to the so-called *M-Matrix Algebraic Riccati Equation*[1] (MARE)

$$XDX - AX - XB + C = 0, \tag{1.3}$$

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

$$W = \begin{array}{c c} & \begin{array}{cc} m & n \end{array} \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} B & -D \\ -C & A \end{pmatrix} \end{array}, \tag{1.4}$$

and W is a nonsingular or an irreducible singular M-matrix. Setting $D = 0$ in (1.3) leads to an MSE (1.1). But according to our definition of an MSE, not all MSEs can arise this way, i.e., those MSEs for which one of A and B is not an M-matrix. It is known [14,15] that MARE (1.3) has a unique minimal nonnegative solution. One important application of MARE (1.3) is in stochastic fluid models [22,23] where each entry of the solution has a physical meaning and it is desirable and important to compute even very tiny entries accurately. To guarantee this, it needs to solve MSE equations arising in certain iterative methods such as the Newton method with high entrywise relative accuracy.

It turns out that if W is entrywise accurate, then the minimal nonnegative solution too is determined and can be computed to a comparable entrywise accuracy. All these will be the subject of study in our paper [31], where the perturbation analysis in this article lays the foundation.

Throughout this article, A, B, and C are reserved for the coefficient matrices of MSE (1.1), and their perturbed ones are denoted, respectively, by the same letters with *tilde*s. For example, the perturbed (1.1) is written as

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

Both A and B have positive diagonal entries and nonpositive off-diagonal entries.

The rest of this paper is organized as follows. Section 2 discusses the relative perturbation theory for the inverse of an M-matrix and how to compute the inverse with the guaranteed accuracy suggested by the theory. Using the results in Sect. 2, Sect. 3 establishes a relative perturbation theory for an MSE as well as examples to illustrate the theory including how it compares to the existing (general) theory. Section 4

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

explains that three types of methods—the GTH-like method, classical fixed point iterations, and the Smith algorithm—after small implementation changes can be used to solve an MSE 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 $\mathbb{R}^{n\times m}$ is the set of all $n \times m$ real matrices, $\mathbb{R}^n = \mathbb{R}^{n\times 1}$, and $\mathbb{R} = \mathbb{R}^1$. I_n (or simply I if its dimension is clear from the context) is the $n \times n$ identity matrix and e_j is its jth column. $\mathbf{1}_n \in \mathbb{R}^n$ is the vector of all ones. The superscript "$\cdot^{\mathrm{T}}$" takes the transpose of a matrix or a vector. For $Z \in \mathbb{R}^{n\times m}$,

1. $Z_{(i,j)}$ refers to its (i, j)th entry;
2. $|Z|$ is the matrix with its (i, j)th entry $|Z_{(i,j)}|$;
3. $\mathrm{vec}(Z) \in \mathbb{R}^{nm}$ is obtained by packing Z's 1st column followed by its 2nd column and so on;
4. When $m = n$, $\mathrm{diag}(Z)$ is the diagonal matrix with the same diagonal entries as Z's, and

$$\varrho(Z) = \rho([\mathrm{diag}(Z)]^{-1}[\mathrm{diag}(Z) - Z]).$$

Inequality $X \le Y$ means $X_{(i,j)} \le Y_{(i,j)}$ for all (i, j), and similarly for $X < Y$, $X \ge Y$, and $X > Y$. In particular, $X \ge 0$ means that X is entrywise nonnegative. With the indeterminant $0/0$ regarded as 0, $X \oslash Y$ denotes the matrix (or vector) entrywise division, i.e., $(X \oslash Y)_{(i,j)} = X_{(i,j)}/Y_{(i,j)}$. We use $\mathrm{fl}(\cdot)$ to denote the numerically computed result of an expression, and $\boldsymbol{u}$ the unit machine roundoff.

2 Inverse of an M-matrix

In this section, we present some results on the inverse of a nonsingular M-matrix. These results will be cited frequently later in the sections for our entrywise perturbation analysis and numerical algorithms for MSE (1.1). The following result is well-known [7].

Theorem 2.1 *Let $A \in \mathbb{R}^{n\times n}$ have nonpositive off-diagonal entries. The following statements are equivalent:*

(a) *A is a nonsingular M-matrix;*
(b) *$A^{-1} \ge 0$;*
(c) *$Au > 0$ for some $u > 0$;*
(d) *All eigenvalues of A have positive real parts.*

A matrix $A \in \mathbb{R}^{n\times n}$ is *reducible* if there is a permutation matrix $\Pi \in \mathbb{R}^{n\times n}$ such that

$$\Pi^{\mathrm{T}} A \Pi = \begin{pmatrix} A_{11} & A_{12} \\ & A_{22} \end{pmatrix},$$

where both A_{11} and A_{22} are square matrices; it is *irreducible* if it is not reducible.

Throughout the rest of this section, $A \in \mathbb{R}^{n\times n}$ *is a nonsingular M-matrix which is perturbed to* $\widetilde{A} \in \mathbb{R}^{n\times n}$, and

$$A = D - N, \quad D = \mathrm{diag}(A), \quad \varrho(A) \stackrel{\text{def}}{=} \rho(D^{-1}N) < 1. \tag{2.1}$$

2.1 Perturbation theory

The following theorem is essentially [2, Theorem 2.5], except $u = \mathbf{1}_n$ there, but a minor modification to its proof works for any $u > 0$.

Theorem 2.2 [2] *If there exist $\epsilon \in \mathbb{R}$ and $u \in \mathbb{R}^n$ such that $0 \le \epsilon < 1$, $u > 0$, and*

$$|(\widetilde{A} - A) \oslash A|_{(i,j)} \le \epsilon \quad \textit{for } i \ne j, \quad \textit{and} \quad |Au - \widetilde{A}u| \le \epsilon Au, \tag{2.2}$$

then $\widetilde{A}$ is a nonsingular M-matrix, and

$$\frac{(1-\epsilon)^{n-1}}{(1+\epsilon)^n} A^{-1} \le \widetilde{A}^{-1} \le \frac{(1+\epsilon)^{n-1}}{(1-\epsilon)^n} A^{-1}. \tag{2.3}$$

Remark 2.1 We make a few comments here.

1. Necessarily $Au \ge 0$ in (2.2).
2. The inequalities in (2.3) are sharp. This is evident for the scalar case $n = 1$. In general, consider $A \in \mathbb{R}^{n\times n}$ given by

$$A_{(i,i)} = 1, \quad A_{(i,i+1)} = -1, \quad A_{(n,1)} = -\theta, \quad \text{all other } A_{(i,j)} = 0,$$

where $0 < \theta < 1$, sufficiently small, such that $\zeta \stackrel{\text{def}}{=} \epsilon \frac{1+\theta^{1/n}}{1-\theta^{1/n}} < 1$. First perturb A to $\widetilde{A}$ as

$$\widetilde{A}_{(i,i)} = 1 - \epsilon, \quad \text{and} \quad \widetilde{A}_{(i,j)} = (1+\epsilon)A_{(i,j)} \qquad \text{for } i \ne j.$$

Take $u = (1, \theta^{1/n}, \theta^{2/n}, \ldots, \theta^{(n-1)/n})^{\mathrm{T}}$. Then $Au = (1-\theta^{1/n})u$, $0 < Au - \widetilde{A}u = \zeta\, Au$. Now apply Theorem 2.2 to get

$$\widetilde{A}^{-1} \le \frac{(1+\zeta)^{n-1}}{(1-\zeta)^n} A^{-1}. \tag{2.4}$$

As $\theta \to 0^+$, it can be seen that $\zeta \to \epsilon$, and

$$(A^{-1})_{(1,n)} \to 1, \quad (\widetilde{A}^{-1})_{(1,n)} \to \frac{(1+\epsilon)^{n-1}}{(1-\epsilon)^n};$$

so both sides of (2.4) approach the same value, meaning the right inequality in (2.3) is in general sharp. Similarly perturb A to $\widetilde{A}$ as

$$\widetilde{A}_{(i,i)} = 1 + \epsilon, \quad \text{and} \quad \widetilde{A}_{(i,j)} = (1-\epsilon)A_{(i,j)} \quad \text{for } i \ne j$$

to conclude that the left inequality in (2.3) is in general sharp as well.

3. If $|Au - \widetilde{A}u| \le \epsilon Au$ in (2.2) is replaced by $|u^{\mathrm{T}}A - u^{\mathrm{T}}\widetilde{A}| \le \epsilon\, u^{\mathrm{T}}A$, the conclusions of the theorem are still valid.

Both inequalities in (2.2) together imply that the diagonal entries of A are determined with comparable entrywise relative accuracy because

$$\mathrm{diag}(A)u = Nu + v \Rightarrow A_{(i,i)} = \frac{v_{(i)} + \sum_{j=1}^{n} N_{(i,j)}u_{(j)}}{u_{(i)}}, \tag{2.5}$$

where $v = Au$. That is to say that (2.2) is stronger than simply requiring $|A - \widetilde{A}| \le \epsilon|A|$ under which we have a weaker result.

Theorem 2.3 *Suppose* $|A - \widetilde{A}| \le \epsilon|A|$. *If* $\delta \stackrel{\mathrm{def}}{=} \frac{1+\varrho(A)}{1-\varrho(A)}\epsilon < 1$, *then* $\widetilde{A}$ *is a nonsingular* M*-matrix, and*

$$\frac{(1-\delta)^{n-1}}{(1+\delta)^n}A^{-1} \le \widetilde{A}^{-1} \le \frac{(1+\delta)^{n-1}}{(1-\delta)^n}A^{-1}. \tag{2.6}$$

Proof Suppose for the moment that $A = D - N$ as in (2.1) is irreducible; so is $D^{-1}N$. Let u be the Perron eigenvector of $D^{-1}N$, i.e., $D^{-1}Nu = \rho(D^{-1}N)u = \varrho(A)u$. We know that $u > 0$ [7, p.27]. It can be seen that $\widetilde{A}_- \le \widetilde{A} \le \widetilde{A}_+$, where $\widetilde{A}_\pm = (1 \pm \epsilon)D - (1 \mp \epsilon)N$. Now

$$\begin{aligned} Au &= [1 - \varrho(A)]Du, \\ \widetilde{A}_\pm u &= [(1 \pm \epsilon) - (1 \mp \epsilon)\varrho(A)]Du \\ &= (1 \pm \delta)Au. \end{aligned}$$

Since $0 \le \delta < 1$, by Theorem 2.2 both $\widetilde{A}_\pm$ are nonsingular M-matrices and so is $\widetilde{A}$, and

$$\frac{(1-\delta)^{n-1}}{(1+\delta)^n}A^{-1} \le \widetilde{A}_+^{-1} \le \widetilde{A}^{-1} \le \widetilde{A}_-^{-1} \le \frac{(1+\delta)^{n-1}}{(1-\delta)^n}A^{-1},$$

as was to be shown.

Now consider the case in which A is reducible. For sufficiently small $\xi > 0$, $A - \xi\, \mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ is an irreducible M-matrix. Apply what we just proved to this modified A and then let $\xi \to 0^+$ to conclude the proof. □

Remark 2.2 The inequality (2.6) implies

$$|\widetilde{A}^{-1} - A^{-1}| \le [(2n-1)\delta + O(\delta^2)]A^{-1} \tag{2.7}$$

for sufficiently small δ. Two comments are in order:

1. The factor $2n - 1$ in the linear term can be improved. In fact, Xue and Jiang [30], using a more complicated argument, gave another version of (2.7) with the linear term $(2n-1)\delta$ replaced by

$$\left(\frac{n}{1-\varrho(A)}+n-1\right)\epsilon=\frac{2n-1-(n-1)\varrho(A)}{1-\varrho(A)}\epsilon\le(2n-1)\delta.$$

2. Modulo the factor $2n-1$, the factor δ in the linear term in (2.7) is asymptotically best possible because

$$\max_{\widetilde{A}}\max_{i,j}|(\widetilde{A}^{-1}-A^{-1})\oslash A^{-1}|_{(i,j)}\ge\frac{\delta}{1-\delta}\quad\text{subject to }|A-\widetilde{A}|\le\epsilon|A|. \tag{2.8}$$

Suppose for the moment that A is irreducible. Let $u>0$ be the Perron eigenvector of $D^{-1}N$. Consider $\widetilde{A}=(1-\epsilon)D-(1+\epsilon)N\le A$ and thus $\widetilde{A}^{-1}\ge A^{-1}$, and

$$b\stackrel{\text{def}}{=}\widetilde{A}u=D[1-\varrho(A)-\epsilon(1+\varrho(A))]u=(1-\delta)Au.$$

Therefore $\widetilde{A}^{-1}b=(1-\delta)^{-1}A^{-1}b$. We have for each i

$$\begin{aligned}\frac{\delta}{1-\delta}&=\frac{\sum_{j=1}^{n}(\widetilde{A}^{-1}-A^{-1})_{(i,j)}b_{(j)}}{\sum_{\ell=1}^{n}(A^{-1})_{(i,\ell)}b_{(\ell)}}\\&=\sum_{j=1}^{n}\frac{(\widetilde{A}^{-1}-A^{-1})_{(i,j)}}{(A^{-1})_{(i,j)}}\frac{(A^{-1})_{(i,j)}b_{(j)}}{\sum_{\ell=1}^{n}(A^{-1})_{(i,\ell)}b_{(\ell)}}\\&\le\max_{1\le j\le n}|(\widetilde{A}^{-1}-A^{-1})\oslash A^{-1}|_{(i,j)},\end{aligned}$$

as was to be shown. Consider now A is reducible. There is a permutation matrix Π such that $\Pi^{\mathrm{T}}A\Pi$ is block upper-triangular with all diagonal blocks square and irreducible. It can be seen that one of the diagonal block, say A_k, has the property that $\varrho(A_k)=\varrho(A)$. Since A_k^{-1} is a submatrix of A^{-1}, we see that, subject to $|A-\widetilde{A}|\le\epsilon|A|$,

$$\max_{\widetilde{A}}\max_{i,j}|(\widetilde{A}^{-1}-A^{-1})\oslash A^{-1}|_{(i,j)}\ge\max_{\widetilde{A}}\max_{i,j}|(\widetilde{A}_k^{-1}-A_k^{-1})\oslash A_k^{-1}|_{(i,j)}\ge\frac{\delta}{1-\delta},$$

as needed.

Remark 2.3 Under the conditions of Theorem 2.3, the commonly used first order error analysis goes as follows. Write $\widetilde{A}=A+(\Delta A)$ and $\widetilde{A}^{-1}=A^{-1}+E$. We have

$$AE=-(\Delta A)A^{-1}-(\Delta A)E$$

by expanding $\widetilde{A}\widetilde{A}^{-1}=[A+(\Delta A)](A^{-1}+E)=I$. Because $A^{-1}\ge 0$,

$$\begin{aligned}|E|&\le|A^{-1}(\Delta A)A^{-1}|+O(\epsilon^2)\\&\le\epsilon\,A^{-1}\,|A|\,A^{-1}+O(\epsilon^2)\end{aligned} \tag{2.9}$$

$$= \epsilon\,(I - D^{-1}N)^{-1}(I + D^{-1}N)\,A^{-1} + O(\epsilon^2)$$
$$= \epsilon\left[A^{-1} + 2\sum_{i=1}^{\infty}(D^{-1}N)^i\,A^{-1}\right] + O(\epsilon^2). \tag{2.10}$$

Since A^{-1} and $\widetilde{A}^{-1}$ have the same zero-nonzero pattern[2], we get

$$|[(A+\Delta A)^{-1} - A^{-1}] \oslash A^{-1}| \le \epsilon\left[\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} + 2\left(\sum_{i=1}^{\infty}(D^{-1}N)^i\,A^{-1}\right) \oslash A^{-1}\right] + O(\epsilon^2). \tag{2.11}$$

The linear terms in (2.10) and (2.11) are sharp. See Proposition 2.1 below. Compared to (2.6),

1. (2.11) is only a first order bound,
2. It gives no indication why and when $|[(A + \Delta A)^{-1} - A^{-1}] \oslash A^{-1}|$ is tiny, unlike (2.6) which says $|[(A + \Delta A)^{-1} - A^{-1}] \oslash A^{-1}|$ is proportional to $[1 - \varrho(A)]^{-1}\epsilon$.

But (2.11) is easily implementable for the practical purpose of error estimation: calculate enough terms in the series $[I_n + 2\sum_i (D^{-1}N)^i]A^{-1}$ to have at least one or more correct decimal digits in every entry. This can be costly sometimes, though, when the series is slowly convergent and some entries are of much tinier magnitudes than others (because convergence to different entries is not uniform in general).

One consequence of (2.7) and (2.9) is the remarkable inequality in the following proposition.

Proposition 2.1 *Let* $\chi = \frac{1+\varrho(A)}{1-\varrho(A)}$. *We have*

$$\limsup_{\epsilon\to 0}\frac{|[(A + \Delta A)^{-1} - A^{-1}] \oslash A^{-1}|}{\epsilon} = A^{-1}\,|A|\,A^{-1} \le (2n-1)\chi\,A^{-1}.$$

Proof The limit equation holds by taking $\Delta A = \epsilon|A|$ or $\Delta A = -\epsilon|A|$ in Remark 2.3. On the other hand, (2.6) implies that the limit is no larger than $(2n-1)\chi\,A^{-1}$. □

Proposition 2.1 implies that the componentwise condition number for the inverse of a nonsingular M-matrix A in the sense of [9] is

$$(A^{-1}\,|A|\,A^{-1}) \oslash A^{-1}. \tag{2.12}$$

2.2 Accurate inverse

Remarks 2.1 and 2.2 say that in general the suggested accuracies by Theorems 2.2 and 2.3 under the specified entrywise perturbations are best possible. In this subsection, we show how to numerically compute A^{-1} with the suggested accuracies.

[2] This is because $A^{-1} = (I - D^{-1}N)^{-1}D^{-1} = \sum_{i=0}^{\infty}(D^{-1}N)^i D^{-1}$ and $\widetilde{A}^{-1} = \sum_{i=0}^{\infty}(\widetilde{D}^{-1}\widetilde{N})^i\,\widetilde{D}^{-1}$, where $\widetilde{D} = \mathrm{diag}(\widetilde{A})$ and $\widetilde{A} = \widetilde{D} - \widetilde{N}$.

According to [1], it is numerically advantageous to represent A by the triplet $\{N, u, v\}$ whenever $0 < u \in \mathbb{R}^n$ is available such that $v = Au \geq 0$. A's diagonal can then be conveniently (and accurately) recovered by (2.5) when needed. In what follows, we will treat indistinguishably an M-matrix and its parameterized triplet representation if available, and write

$$A = \{N, u, v\}, \quad \text{where } u > 0 \text{ and } v = Au \geq 0. \tag{2.13}$$

It is worth pointing out that numerically v is not exactly Au but an entrywise accurate approximation which is all that is needed for the GTH-like algorithm [1], an extension of the GTH algorithm [12] for stochastic matrices, to work.

There are two cases to consider. First, when this triplet representation is known, the GTH-like algorithm [1] which is *entrywise forward stable* computes A^{-1} with entrywise relative accuracy dictated by the working precision [1].

If a triplet representation is not known *a priori*, then all we need is to find a positive vector $u \in \mathbb{R}^n$ such that $\widetilde{v} = \mathrm{fl}(Au) \geq 0$. This can often be achieved by solving $(I - D^{-1}N)u = \mathbf{1}_n$ for u by, e.g., Gaussian elimination (with partial pivoting), because theoretically $u = (I - D^{-1}N)^{-1}\mathbf{1}_n > 0$. *Unless $I - D^{-1}N$ is almost singular*, $u > 0$ and the residual $(I - D^{-1}N)u - \mathbf{1}_n$ for the computed u is tiny, relative to $\mathbf{1}_n$. This means

$$\widetilde{v} = \mathrm{fl}(Au) = \mathrm{fl}(D[I - D^{-1}N]u) \approx D\mathbf{1}_n \geq 0.$$

Even if $I - D^{-1}N$ is almost singular, it is still possible that $u > 0$ and $\widetilde{v} = \mathrm{fl}(Au) \geq 0$. Suppose for the moment this is the case. It is not difficult to show that

$$\widetilde{v} = \mathrm{fl}(Au) = \widetilde{A}u \quad \text{for some } \widetilde{A} \text{ satisfying } |\widetilde{A} - A| \leq [n\boldsymbol{u} + O(\boldsymbol{u}^2)]|A|,$$

where $\boldsymbol{u}$ is the machine unit roundoff. Split $\widetilde{A} = \widetilde{D} - \widetilde{N}$ with $\widetilde{D} = \mathrm{diag}(\widetilde{A})$. We then have found $\widetilde{A} = \{\widetilde{N}, u, \widetilde{v}\}$. Apply Theorem 2.3 to get

$$|\widetilde{A}^{-1} - A^{-1}| \leq [n(2n-1)\delta + O(\delta^2)]A^{-1}, \tag{2.14}$$

where $\delta = \frac{1+\varrho(A)}{1-\varrho(A)}\boldsymbol{u}$. But since $\widetilde{N}$ is unknown and in actual computation, we have to compute the inverse of the M-matrix $\widehat{A} \stackrel{\text{def}}{=} \{N, u, \widetilde{v}\}$ instead by the GTH-like algorithm. Because the algorithm is entrywise forward stable [1], the computed inverse $\widehat{X}$ of $\widehat{A} = \{N, u, \widetilde{v}\}$ by the GTH-like algorithm differs from $\widehat{A}^{-1}$ entrywise by $O(\boldsymbol{u})$, i.e., $|\widehat{A}^{-1} - \widehat{X}| \leq O(\boldsymbol{u})\widehat{A}^{-1}$. Theorem 2.2 says $|\widetilde{A}^{-1} - \widehat{A}^{-1}| \leq [n(2n-1)\boldsymbol{u} + O(\boldsymbol{u}^2)]\widehat{A}^{-1}$. Putting all together to get

$$\begin{aligned} |A^{-1} - \widehat{X}| &\leq |A^{-1} - \widetilde{A}^{-1}| + |\widetilde{A}^{-1} - \widehat{A}^{-1}| + |\widehat{A}^{-1} - \widehat{X}| \\ &\leq [f(n)\delta + O(\delta^2)]A^{-1}, \end{aligned}$$

where f is some low degree polynomial. This implies that the computed $\widehat{X}$ as an approximation to A^{-1} achieves the guaranteed accuracy suggested by Theorem 2.3,

except the low degree polynomial factor $f(n)$ (which commonly appears in most error analysis in Numerical Linear Algebra and usually overestimates the actual).

When either $u > 0$ or $\widetilde{v} = \mathrm{fl}(Au) \geq 0$ fails to hold for the computed u through solving $(I - D^{-1}N)u = \mathbf{1}_m$, $I - D^{-1}N$ is almost singular[3]. Solving $(I - D^{-1}N)u = \mathbf{1}_m$ is basically one step of the inverse iteration. Which means u may come out very close to the Perron eigenvector of $D^{-1}N$. There are two subcases:

A is irreducible. So is $D^{-1}N$. Its Perron eigenvector is entrywise positive. Therefore we may continue the inverse iteration for a few more steps: repeat until $u > 0$ and $v = Au \geq 0$,

$$\text{solve } (I - D^{-1}N)u_1 = u \text{ for } u_1; \text{ set } u := u_1/\|u_1\|_\infty;$$

where the ℓ_∞-norm $\|u_1\|_\infty = \max_i |(u_1)_{(i)}|$. We may also use the inverse iteration described in [8,29] which is more expensive per step because the linear system differs from one step to another.

A is reducible. More steps of the inverse iteration may not help because the Perron eigenvector may have entries whose values are 0. For such a case, we first find a permutation matrix Π such that[4]

$$\Pi^{\mathrm{T}} A \Pi = \begin{pmatrix} A_{11} & -A_{12} & \dots & -A_{1q} \\ & A_{22} & \dots & -A_{2q} \\ & & \ddots & \vdots \\ & & & A_{qq} \end{pmatrix},$$

where all A_{ii} are nonsingular irreducible M-matrices, and all $A_{ij} \geq 0$ for $i \neq j$. It suffices to be able to compute all $A_{ii}^{-1} \geq 0$ as accurately as they can be because then $(\Pi^{\mathrm{T}} A \Pi)^{-1}$ is block upper-triangular with diagonal blocks A_{ii}^{-1} and off-diagonal blocks computed without a single subtraction. Each A_{ii}^{-1} can be computed accurately as described above for the irreducible A.

3 Entrywise perturbation analysis

Throughout this section, $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$, and they are split as

$$A = D_1 - N_1, \quad D_1 = \mathrm{diag}(A), \quad \varrho_1 = \varrho(A) \stackrel{\mathrm{def}}{=} \rho(D_1^{-1}N_1), \tag{3.1a}$$

$$B = D_2 - N_2, \quad D_2 = \mathrm{diag}(B), \quad \varrho_2 = \varrho(B) \stackrel{\mathrm{def}}{=} \rho(D_2^{-1}N_2). \tag{3.1b}$$

with all $A_{(i,i)} > 0$, all $B_{(j,j)} > 0$, and all $N_i \geq 0$. Also $0 \leq C \in \mathbb{R}^{n\times m}$. Possibly one of ϱ_i may be bigger or equal to 1. Recall MSEs (1.1) and (1.5). P is defined by (1.2) and $\widetilde{P}$ is defined similarly. Set

[3] Or $I - D^{-1}N$ *is* singular. This situation will come up later and yet we need a triplet representation for it.

[4] This can be done by MATLAB function `dmperm`. For numerical efficiency, this should be done first actually to decouple $AX + XB = C$.

$$\tau_1 = \frac{\min_i A_{(i,i)}}{\max_j B_{(j,j)}}, \quad \tau_2 = \frac{\min_j B_{(j,j)}}{\max_i A_{(i,i)}}. \tag{3.2}$$

3.1 Main results

Our main entrywise relative perturbation results for MSE (1.1) are Theorems 3.1 and 3.2. Theorem 3.1 assumes that both A and B are nonsingular M-matrices, and Theorem 3.2 does not require this but have conditions to imply that $P = I_m \otimes A + B^{\mathrm{T}} \otimes I_n$ is a nonsingular M-matrix. Note that when both A and B are nonsingular M-matrices, P is a nonsingular M-matrix, but not the other way around.

Because MSE (1.1) is equivalent to $P\,\mathrm{vec}(X) = \mathrm{vec}(C)$ and P is a nonsingular M-matrix, entrywise relative perturbation results in Sect. 2.1 on the inverse of a nonsingular M-matrix are naturally applicable to yield entrywise relative perturbation bounds for X. Indeed this is the case for Theorem 3.1 below. But other resulting bounds by such a straightforward application will contain $\varrho(P)$. Because of its huge dimensional size relative to the sizes of A and B, estimates on $\varrho(P)$ are potentially much more difficult to obtain than ϱ_i. For this reason, in what follows we adopt a not-so-straightforward approach to establish bounds involving ϱ_i but not $\varrho(P)$.

For the sake of presentation, we introduce

$$\eta^+_{p;t} = \left(\frac{1+t}{1-t}\right)^p, \quad \eta^-_{p;t} = \left(\frac{1-t}{1+t}\right)^p \quad \text{for} \quad 0 \le t < 1 \text{ and } p \ge 0. \tag{3.3}$$

For sufficiently tiny $t \ge 0$, we have asymptotically $\eta^{\pm}_{p;t} = 1 \pm 2pt + O(t^2)$. Our main results in this section take the form

$$\eta^-_{mn;\chi\epsilon}\, X \le \widetilde{X} \le \eta^+_{mn;\chi\epsilon}\, X, \quad \text{or asymptotically,} \tag{3.4a}$$

$$|\widetilde{X} - X| \le [2mn\,\chi\epsilon + O(\epsilon^2)]X, \tag{3.4b}$$

where χ is to be specified later.

Theorem 3.1 *Suppose that A and B are nonsingular M-matrices and that there exist $0 \le \epsilon < 1, 0 < u \in \mathbb{R}^n$, and $0 < y \in \mathbb{R}^m$ such that*

$$|(\widetilde{A} - A) \oslash A|_{(i,j)} \le \epsilon \;\; \textit{for } i \ne j, \quad |Au - \widetilde{A}u| \le \epsilon\, Au, \tag{3.5a}$$

$$|(\widetilde{B} - B) \oslash B|_{(i,j)} \le \epsilon \;\; \textit{for } i \ne j, \quad |B^{\mathrm{T}}y - \widetilde{B}^{\mathrm{T}}y| \le \epsilon\, B^{\mathrm{T}}y, \tag{3.5b}$$

$$|C - \widetilde{C}| \le \epsilon C. \tag{3.5c}$$

Then $\widetilde{A}$ and $\widetilde{B}$ too are nonsingular M-matrices, and (3.4) holds with $\chi = 1$.

Proof The MSEs can be rewritten equivalently as $P\,\mathrm{vec}(X) = \mathrm{vec}(C)$ and $\widetilde{P}\,\mathrm{vec}(\widetilde{X}) = \mathrm{vec}(\widetilde{C})$. P is a nonsingular M-matrix because A and B are. The inequalities in (3.5a) and (3.5b) guarantee that $\widetilde{A}$ and $\widetilde{B}$ are nonsingular M-matrices, too; so is $\widetilde{P}$. By (3.5a) and (3.5b), we have

$$\begin{aligned}\widetilde{P}(y \otimes u) &= y \otimes \widetilde{A}u + \widetilde{B}^{\mathrm{T}} y \otimes u \\ &\le (1+\epsilon)[y \otimes Au + B^{\mathrm{T}} y \otimes u] \\ &= (1+\epsilon)\, P(y \otimes u),\end{aligned}$$

and similarly $\widetilde{P}(y \otimes u) \ge (1-\epsilon)\, P(y \otimes u)$. Together they imply

$$|P(y \otimes u) - \widetilde{P}(y \otimes u)| \le \epsilon\, P(y \otimes u).$$

On the other hand, the off-diagonal entries at the same positions for P and $\widetilde{P}$ are one of the three possible cases:

$$0 \text{ and } 0, \text{ or } A_{(i,j)} \text{ and } \widetilde{A}_{(i,j)}, \text{ or } B_{(k,\ell)} \text{ and } \widetilde{B}_{(k,\ell)},$$

where $i \ne j$ and $k \ne \ell$. Thus $|(\widetilde{P} - P) \oslash P|_{(i,j)} \le \epsilon$ for $i \ne j$ by (3.5a) and (3.5b). So the conditions of Theorem 2.2 are satisfied for P and $\widetilde{P}$, and thus

$$\begin{aligned}\mathrm{vec}(\widetilde{X}) &= \widetilde{P}^{-1}\, \mathrm{vec}(\widetilde{C}) \\ &\le \frac{(1+\epsilon)^{mn-1}}{(1-\epsilon)^{mn}} P^{-1}\, (1+\epsilon)\, \mathrm{vec}(C) \\ &= \frac{(1+\epsilon)^{mn}}{(1-\epsilon)^{mn}}\, \mathrm{vec}(X), \\ \mathrm{vec}(\widetilde{X}) &\ge \frac{(1-\epsilon)^{mn}}{(1+\epsilon)^{mn}}\, \mathrm{vec}(X),\end{aligned}$$

as was to be shown. □

Remark 3.1 Alternatively if the last inequalities in (3.5a) and (3.5b) are replaced by

$$|u^{\mathrm{T}} A - u^{\mathrm{T}} \widetilde{A}| \le \epsilon\, u^{\mathrm{T}} A, \quad |y^{\mathrm{T}} B^{\mathrm{T}} - y^{\mathrm{T}} \widetilde{B}^{\mathrm{T}}| \le \epsilon\, y^{\mathrm{T}} B^{\mathrm{T}},$$

respectively, the conclusion of this theorem still holds.

It is not necessary to require that both A and B are nonsingular M-matrices in order for $AX + XB = C$ to have a unique solution. What is necessary is that P is nonsingular. The following lemma presents conditions to ensure the nonsingularity of P.

Lemma 3.1 *If either*

$$0 \le \varrho_1 < 1, \quad 0 \le \varrho_1 \le \varrho_2 < 1 + \tau_1(1 - \varrho_1), \tag{3.6}$$

or

$$0 \le \varrho_2 \le \varrho_1 < 1 + \tau_2(1 - \varrho_2), \quad 0 \le \varrho_2 < 1, \tag{3.7}$$

then $P = I_m \otimes A + B^{\mathrm{T}} \otimes I_n$ *is a nonsingular* M*-matrix.*

Proof We first consider the case in which both A and B are irreducible. We'll only consider the case (3.6). Denote by $u > 0$ and $y > 0$ the Perron eigenvectors of $D_1^{-1}N_1$ and $D_2^{-1}N_2^{\mathrm{T}}$, respectively, i.e.,

$$D_1^{-1}N_1u = \varrho_1 u, \quad D_2^{-1}N_2^{\mathrm{T}}y = \varrho_2 y. \tag{3.8}$$

Now if also $\varrho_2 < 1$, then it is evident that P is a nonsingular M-matrix. Assume that $\varrho_2 \geq 1$. We have $P(y \otimes u) = (1-\varrho_1)y \otimes (D_1u) + (1-\varrho_2)(D_2y) \otimes u$. Notice

$$(D_2y) \otimes u \leq \frac{\max_j B_{(j,j)}}{\min_i A_{(i,i)}} y \otimes (D_1u) = \tau_1^{-1} y \otimes (D_1u)$$

to get

$$P(y \otimes u) \geq [(1-\varrho_1) + (1-\varrho_2)\tau_1^{-1}]y \otimes (D_1u) > 0 \tag{3.9}$$

by (3.6). Therefore P is a nonsingular M-matrix by Theorem 2.1.

The first inequality in (3.9) leads to

$$P(y \otimes u) \geq \{[(1-\varrho_1) + (1-\varrho_2)\tau_1^{-1}] \min_i A_{(i,i)}\}\, y \otimes u,$$
$$P^{-1}(y \otimes u) \leq \{[(1-\varrho_1) + (1-\varrho_2)\tau_1^{-1}] \min_i A_{(i,i)}\}^{-1}\, y \otimes u.$$

By [7, p.28]

$$\rho(P^{-1}) \leq \{[(1-\varrho_1) + (1-\varrho_2)\tau_1^{-1}] \min_i A_{(i,i)}\}^{-1}.$$

Let $\alpha(P)$ denote the eigenvalue of P with the smallest real part. Then [18, Problem 19 on p.129]

$$\alpha(P) = \frac{1}{\rho(P^{-1})} \geq ((1-\varrho_1) + (1-\varrho_2)\tau_1^{-1}) \min_i A_{(i,i)} > 0. \tag{3.10}$$

Now for reducible A or B, let $\widehat{A} = A - \xi \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}}$ and $\widehat{B} = B - \xi \mathbf{1}_m \mathbf{1}_m^{\mathrm{T}}$ for sufficiently small ξ, and let $\widehat{\varrho}_1$, $\widehat{\varrho}_2$ and $\widehat{P}$ be defined similarly to their counterparts ϱ_1, ϱ_2 and P. We have

$$\alpha(\widehat{P}) \geq [(1-\widehat{\varrho}_1) + (1-\widehat{\varrho}_2)\widehat{\tau}_1^{-1}] \min_i \widehat{A}_{(i,i)}.$$

As $\xi \to 0^+$, by the continuity of eigenvalues, $\alpha(\widehat{P}) \to \alpha(P)$, $\widehat{\varrho}_1 \to \varrho_1$, and $\widehat{\varrho}_2 \to \varrho_2$. Thus (3.10) holds for the reducible case, too. Hence P is a nonsingular M-matrix. □

Theorem 3.2 *Under the conditions of* Lemma 3.1*, suppose*

$$|A - \widetilde{A}| \leq \epsilon |A|, \quad |B - \widetilde{B}| \leq \epsilon |B|, \quad |C - \widetilde{C}| \leq \epsilon C, \tag{3.11}$$

and let

$$\chi = \begin{cases} \dfrac{1+\varrho_1+(1+\varrho_2)\tau_1^{-1}}{1-\varrho_1+(1-\varrho_2)\tau_1^{-1}}, & \textit{if } (3.6), \\ \dfrac{1+\varrho_2+(1+\varrho_1)\tau_2^{-1}}{1-\varrho_2+(1-\varrho_1)\tau_2^{-1}}, & \textit{if } (3.7). \end{cases} \tag{3.12}$$

If $\chi\,\epsilon < 1$, *then* (3.4) *holds.*

Proof We shall prove the claim under (3.6). Lemma 3.1 says that P is a nonsingular M-matrix. Suppose for the moment that both A and B are irreducible, and let $u > 0$ and $y > 0$ be the Perron eigenvectors of $D_1^{-1}N_1$ and $D_2^{-1}N_2^{\mathrm{T}}$, respectively. So (3.8) holds. We have

$$P(y\otimes u) = (1-\varrho_1)y\otimes D_1u + (1-\varrho_2)D_2y\otimes u, \tag{3.13}$$

$$\widetilde{P}(y\otimes u) \geq [(1-\epsilon)-(1+\epsilon)\varrho_1]y\otimes(D_1u) + [(1-\epsilon)-(1+\epsilon)\varrho_2](D_2y)\otimes u, \tag{3.14}$$

$$\widetilde{P}(y\otimes u) \leq [(1+\epsilon)-(1-\epsilon)\varrho_1]y\otimes(D_1u) + [(1+\epsilon)-(1-\epsilon)\varrho_2](D_2y)\otimes u. \tag{3.15}$$

The conditions of this theorem guarantee that the right-hand sides of these inequalities are positive. Now we look at the entrywise ratio $[\widetilde{P}(y\otimes u)]\oslash[P(y\otimes u)]$ whose typical kth entry satisfies, by (3.13) and (3.14),

$$\frac{[\widetilde{P}(y\otimes u)]_{(k)}}{[P(y\otimes u)]_{(k)}} \geq \frac{[(1-\epsilon)-(1+\epsilon)\varrho_1]A_{(i,i)}+[(1-\epsilon)-(1+\epsilon)\varrho_2]B_{(j,j)}}{(1-\varrho_1)A_{(i,i)}+(1-\varrho_2)B_{(j,j)}} \tag{3.16}$$

$$\geq \frac{[(1-\epsilon)-(1+\epsilon)\varrho_1]\tau_1+[(1-\epsilon)-(1+\epsilon)\varrho_2]}{(1-\varrho_1)\tau_1+(1-\varrho_2)} \tag{3.17}$$

for some i and j, where the inequality sign in (3.17) is due to the fact that the right-hand side of (3.16) is an increasing function of $t = A_{(i,i)}/B_{(j,j)}$ because its derivative with respect to t is

$$\frac{2\epsilon(\varrho_2-\varrho_1)}{[(1-\varrho_1)t+(1-\varrho_2)]^2} \geq 0,$$

by (3.6). Therefore $\widetilde{P}(y\otimes u) \geq (1-\delta)P(y\otimes u)$, where $\delta = \chi\,\epsilon$. Similarly, we can show that $\widetilde{P}(y\otimes u) \leq (1+\delta)P(y\otimes u)$. It can be seen that $|(\widetilde{P}-P)\oslash P|_{(i,j)} \leq \epsilon$ for $i \neq j$. Apply Theorem 2.2 to complete the proof.

Now for possibly reducible A and B, we apply the result just proved for the irreducible case to MSEs with $A - \xi\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}$ and $B - \xi\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ and with $\widetilde{A} - \xi\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}$ and $\widetilde{B} - \xi\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$, and then let $\xi \to 0^+$ to conclude the proof. □

Theorem 3.2 is applicable to the following important cases:

1. Both A and B are nonsingular M-matrices, i.e., $\varrho_{\max} = \max\{\rho_1, \rho_2\} < 1$. This application is done in Corollary 3.1 below.
2. $0 < \varrho_1 < \varrho_2 = 1$. The application is straightforward—χ becomes

$$\chi = \frac{1 + \varrho_1 + 2\tau_1^{-1}}{1 - \varrho_1}.$$

But if also $B^{\mathrm{T}} y = \widetilde{B}^{\mathrm{T}} y = 0$ for the Perron eigenvector y of $D_2^{-1} N_2^{\mathrm{T}}$, a sharper bound can be gotten. See Theorem 3.3 below. This case will become important later in our perturbation analysis for the Wiener–Hopf factorization in [31].

Corollary 3.1 *Suppose that A and B are nonsingular M-matrices, and set $\varrho_{\max} = \max\{\varrho_1, \varrho_2\}$ and $\chi = \frac{1+\varrho_{\max}}{1-\varrho_{\max}}$. If (3.11) holds and $\chi\,\epsilon < 1$, then $\widetilde{A}$ and $\widetilde{B}$ are nonsingular M-matrices, and (3.4) holds.*

Proof That $\widetilde{A}$ and $\widetilde{B}$ are nonsingular M-matrices is because of Theorem 2.3. It can be verified that

$$\frac{1 + \varrho_i + (1 + \varrho_j)\tau_i^{-1}}{1 - \varrho_i + (1 - \varrho_j)\tau_i^{-1}} \le \frac{1 + \varrho_{\max}}{1 - \varrho_{\max}}.$$

Apply Theorem 3.2 to conclude the proof. □

Theorem 3.3 *Assume $\varrho_1 < \varrho_2 = 1$, B is irreducible, and $\widetilde{B}^{\mathrm{T}} y = 0$, where $y > 0$ is the Perron vector of $D_2^{-1} N_2^{\mathrm{T}}$. Let $\chi = \frac{1+\varrho_1}{1-\varrho_1}$. If (3.11) holds and $\chi\,\epsilon < 1$, then (3.4) holds.*

Proof Assume that A is irreducible and then use the limiting argument for the reducible case. We have $y \otimes u > 0$. Let $\delta = \chi\,\epsilon$. Use $B^{\mathrm{T}} y = \widetilde{B}^{\mathrm{T}} y = 0$ to get

$$\begin{aligned}
P(y \otimes u) &= (1 - \varrho_1) y \otimes D_1 u, \\
\widetilde{P}(y \otimes u) &\ge ((1 - \epsilon) - (1 + \epsilon)\varrho_1) y \otimes D_1 u \\
&\ge (1 - \delta) P(y \otimes u), \\
\widetilde{P}(y \otimes u) &\le (1 + \delta) P(y \otimes u).
\end{aligned}$$

Apply Theorem 2.2 to complete the proof. □

Remark 3.2 1. For a more general splitting $A = D_1' - N_1'$ and $B = D_2' - N_2'$, where D_1' and D_2' are positive diagonal and N_1' and N_2' are nonnegative. Define $\varrho_i' = \rho((D_i')^{-1} N_i')$ for $i = 1, 2$, then all the results above still hold with ϱ_i replaced by ϱ_i', τ_i by τ_i', where

$$\tau_1' = \frac{\min_i (D_1')_{(i,i)}}{\max_j (D_2')_{(j,j)}}, \quad \tau_2' = \frac{\min_j (D_2')_{(j,j)}}{\max_i (D_1')_{(i,i)}}.$$

2. Similarly to what is in Remark 2.3, the commonly used first order error analysis can be performed, too, under the assumption that P is a nonsingular M-matrix and (3.11) hold. Write $\widetilde{Z} = Z + (\Delta Z)$ for $Z = A, B, C$, and X, where ΔZ is the perturbation to Z. Substitute them into $\widetilde{A}\widetilde{X} + \widetilde{X}\widetilde{B} = \widetilde{C}$ and use $AX + XB = C$ to get

$$A(\Delta X) + (\Delta X)B = (\Delta C) - (\Delta A)X - X(\Delta B) - (\Delta A)(\Delta X) - (\Delta X)(\Delta B).$$

Define linear operator $\mathcal{L} : Z \to AZ + ZB$ whose matrix representation is P for which $P^{-1} \geq 0$. Therefore,

$$\begin{aligned} |\Delta X| &\leq \mathcal{L}^{-1}(|(\Delta C) - (\Delta A)X - X(\Delta B)|) + O(\epsilon^2) \\ &\leq \epsilon\, \mathcal{L}^{-1}(C + |A|X + X|B|) + O(\epsilon^2) \\ &= 2\epsilon\, \mathcal{L}^{-1}(D_1 X + X D_2) + O(\epsilon^2). \end{aligned} \tag{3.18}$$

Since X and $\widetilde{X}$ have the same zero-nonzero pattern, we have

$$\begin{aligned} |(\Delta X) \oslash X| &\leq 2\epsilon \Upsilon \oslash X + O(\epsilon^2) \\ &\leq 2\gamma\epsilon \mathbf{1}_n \mathbf{1}_m^{\mathrm{T}} + O(\epsilon^2) \end{aligned} \tag{3.19}$$

where $\Upsilon \in \mathbb{R}^{n\times m}$ and γ are defined by

$$A\Upsilon + \Upsilon B = D_1 X + X D_2, \quad \gamma = \max_{i,j}(\Upsilon \oslash X)_{(i,j)}. \tag{3.20}$$

Two immediate comments are

1. $\Upsilon = \mathcal{L}^{-1}(C) + \mathcal{L}^{-1}\left(|\mathcal{L}|(\mathcal{L}^{-1}(C))\right) \geq \mathcal{L}^{-1}\left(|\mathcal{L}|(\mathcal{L}^{-1}(C))\right)$ since $X = \mathcal{L}^{-1}(C)$.
2. The linear terms in (3.18) and (3.19) are sharp. See Proposition 3.1 below.

Compared to (3.4a) of the theorems and corollaries, (3.19) is a first order bound (with its linear term sharp) and it also does not reveal the informative insight into the sensitivity of X as (3.4a) does, e.g., its proportionality to $(1 - \varrho_{\max})^{-1}$ in Corollary 3.1. But (3.19) is easily implementable for the practical purpose of error estimation. The following iterative scheme: $\Upsilon_0 = X$ and for $k \geq 0$

$$D_1 \Upsilon_{k+1} + \Upsilon_{k+1} D_2 = N_1 \Upsilon_k + \Upsilon_k N_2 + D_1 X + X D_2 \tag{3.21}$$

produces a sequence $\{\Upsilon_i\}$ that monotonically convergent to Υ because P is a nonsingular M-matrix. Iterate (3.21) enough steps until Υ_k has one or more correct decimal digits in each of its entries. This can be costly, though, when it is slowly convergent and Υ has entries of widely varying magnitudes.

Proposition 3.1 *Under the conditions of* Theorem 3.2*, we have*

$$\limsup_{\epsilon\to 0} \frac{|\Delta X|}{\epsilon} = 2\mathcal{L}^{-1}(C) + 2\mathcal{L}^{-1}\left(|\mathcal{L}|(\mathcal{L}^{-1}(C))\right), \tag{3.22}$$

$$\mathcal{L}^{-1}\left(|\mathcal{L}|(\mathcal{L}^{-1}(C))\right) \leq (mn\chi - 1)\, \mathcal{L}^{-1}(C), \tag{3.23}$$

where linear operator $\mathcal{L}(Z) = AZ + ZB$ *and* Υ *is defined by the first equation in* (3.20)*, and* χ *in* Theorem 3.2.

Proof In Item 2 of Remark 3.2, take $\Delta C = \pm\epsilon\, C$, $\Delta A = \mp\epsilon\, |A|$, and $\Delta B = \mp\epsilon\, |B|$ to see (3.22). On the other hand, Theorem 3.2 says that this limit is no larger than $2mn\chi\, \mathcal{L}^{-1}(C)$. □

Proposition 3.1 implies that the componentwise condition number of MSE (1.1) in the sense of [9] is

$$[2\mathcal{L}^{-1}(C) + 2\mathcal{L}^{-1}\left(|\mathcal{L}|(\mathcal{L}^{-1}(C))\right)] \oslash X. \tag{3.24}$$

3.2 Discussions

The standard perturbation results [17, p. 313] suggest that the error $\|\widetilde{X} - X\|$ could be as big as inversely proportional to the smallest singular value of P or equivalently the separation between A and $-B$. Our entrywise relative perturbation bound by Theorem 3.1 is always tiny and those by Theorem 3.2, Corollary 3.1 and Theorem 3.3 are inversely proportional to

$$1 - \varrho_i + (1 - \varrho_j)\tau_i^{-1}, \quad 1 - \varrho_{\max}, \quad 1 - \varrho_1,$$

respectively. Since usually the smallest singular value of P is smaller and can be made arbitrarily smaller than any of them (even for M-matrices A and B as indicated by Example 3.1 below), our bounds are often (much) sharper in such a situation.

Example 3.1 Let the entries of $U_m \in \mathbb{R}^{m\times m}$ be 0 except its $(i, i+1)$th entries which are 1, and let $A = \frac{1}{2}I_n - \omega U_n$ and $B = \frac{1}{2}I_m - \omega U_m^{\mathrm{T}}$. Both A and B are nonsingular M-matrices, $\varrho_{\max} = \varrho_1 = \varrho_2 = 0$, and thus $1 - \varrho_{\max} = 1$. Consequently $\chi = 1$ in Theorem 3.2 and therefore small entrywise relative perturbations as in (3.11) will only introduce small entrywise relative changes to the solution X of MSE $AX + XB = C$. On the other hand, $\|P^{-1}\|_2 = O(\omega^{m+n-2})$ for $\omega > 1$, where $\|\cdot\|_2$ denotes the spectral norm of a matrix. Consequently the standard perturbation results [17, p. 313] can produce error bounds on $\|\widetilde{X} - X\|/\|X\|$ that are too larger than the normwise relative perturbations to A, B, and C to be useful.

In order to see $\|P^{-1}\|_2 = O(\omega^{m+n-2})$, we let $\Omega_k = \mathrm{diag}(1, \omega, \ldots, \omega^{k-1})$. Then we have

$$\begin{aligned}
A &= \Omega_n^{-1}(\tfrac{1}{2}I_n - U_n)\Omega_n,\\
B^{\mathrm{T}} &= \Omega_m^{-1}(\tfrac{1}{2}I_m - U_m)\Omega_m,\\
P &= (\Omega_m^{-1} \otimes \Omega_n^{-1})[I_m \otimes (\tfrac{1}{2}I_n - U_n) + (\tfrac{1}{2}I_m - U_m) \otimes I_n](\Omega_m \otimes \Omega_n)\\
&= (\Omega_m^{-1} \otimes \Omega_n^{-1}) \begin{pmatrix} T & -I_n & & \\ & \ddots & \ddots & \\ & & \ddots & -I_n \\ & & & T \end{pmatrix}_{m\times m} (\Omega_m \otimes \Omega_n),
\end{aligned}$$

where $T = I_n - U_n$. Therefore

$$P^{-1} = \begin{pmatrix} \Omega_n^{-1}T^{-1}\Omega_n & \omega\Omega_n^{-1}T^{-2}\Omega_n & \cdots & \omega^{m-1}\Omega_n^{-1}T^{-m}\Omega_n \\ & \Omega_n^{-1}T^{-1}\Omega_n & \ldots & \omega^{m-2}\Omega_n^{-1}T^{-m+1}\Omega_n \\ & & \ddots & \vdots \\ & & & \Omega_n^{-1}T^{-1}\Omega_n \end{pmatrix}.$$

Now notice $T^{-k} = \sum_{i=0}^{n-1} \frac{k(k+1)\cdots(k+i-1)}{i!} U_n^i$ to see that as $\omega \to \infty$

$$\frac{\omega^{k-1}\Omega_n^{-1}T^{-k}\Omega_n}{\omega^{m+n-2}} \to \begin{cases} 0, & \text{for } k < m, \\ \frac{m(m+1)\cdots(m+n-2)}{(n-1)!} e_1 e_n^{\mathrm{T}}, & \text{for } k = m. \end{cases}$$

Therefore as $\omega \to \infty$

$$\|P^{-1}\|_2 = \frac{(m+n-2)!}{(m-1)!\,(n-1)!}\,\omega^{m+n-2} + O(\omega^{m+n-3}),$$

as expected. This expression exposes two factors that contribute to the rapid growth of $\|P^{-1}\|_2$: the factor involving the factorials and the factor ω^{m+n-2}. Both grow prohibitively fast.

We suspect that the factor $2mn$ in (3.4b) probably overestimate entrywise relative changes. Our suspicion comes from the following—usually extreme—example.

Example 3.2 In Example 3.1 above: $A = \frac{1}{2}I_n - \omega U_n$ and $B = \frac{1}{2}I_m - \omega U_m^{\mathrm{T}}$, we perturb both $\frac{1}{2}$ to $\frac{1}{2}(1-\epsilon)$ and both ω to $\omega(1+\epsilon)$, where $-1 < \epsilon < 1$. Following the line of arguments there, we see

$$\widetilde{P}_{(1,mn)}^{-1} = \frac{(1+\epsilon)^{m+n-2}}{(1-\epsilon)^{m+n-1}}\,P_{(1,mn)}^{-1}.$$

So for $C = e_n e_m^{\mathrm{T}}$ perturbed to $\widetilde{C} = (1+\epsilon)e_n e_m^{\mathrm{T}}$, we have

$$\widetilde{X}_{(1,1)} = \frac{(1+\epsilon)^{m+n-1}}{(1-\epsilon)^{m+n-1}}\,X_{(1,1)}.$$

Since usually such A and B are the worst of all, we conjecture that $2mn$ might be replaceable by $2(m+n-1)$.

The linear term in (3.4b) reflects the correct order of entrywise relative changes in X for all covered cases by the theorems and corollaries above. This is explained by the following examples.

Example 3.3 For Theorem 3.1, we take A and B as $\frac{1}{2}I_n - \omega U_n$ and $\frac{1}{2}I_m - \omega U_m^{\mathrm{T}}$, respectively, except introducing a tiny number θ to A's $(n,1)$th position and to B's $(1,m)$th position. For $Z = A$ and B, perturb Z to $\widetilde{Z}$ as

$$\widetilde{Z}_{(i,i)} = (1-\epsilon)Z_{(i,i)}, \quad \text{and} \quad \widetilde{Z}_{(i,j)} = (1+\epsilon)Z_{i,j)} \quad \text{for } i \neq j.$$

Following the argument in Item 2 of Remark 2.1, we see the conditions of Theorem 3.1 with $C = e_n e_m^{\mathrm{T}}$ and $\widetilde{C} = (1+\epsilon)e_n e_m^{\mathrm{T}}$ are satisfied. As $\theta \to 0$,

$$\widetilde{X}_{(1,1)} \to \frac{(1+\epsilon)^{m+n-1}}{(1-\epsilon)^{m+n-1}} X_{(1,1)} = (2[m+n-1]\epsilon + O(\epsilon^2))\, X_{(1,1)}.$$

So the linear term in (3.4b) is at least of $O(\epsilon)$ in general.

Example 3.4 For Theorem 3.2, the linear term in (3.4b) contains the factor

$$[1 - \varrho_i + (1-\varrho_j)\tau_i^{-1}]^{-1}\,\epsilon$$

that we shall argue is asymptotically best possible. Consider $A = \alpha I_n - N_1$ and $B = \beta I_m - N_2$, both irreducible. Then $\tau_1 = \alpha/\beta = \tau_2^{-1}$. Consider $\widetilde{A} = (1-\epsilon)\alpha I_n - (1+\epsilon)N_1$ and $\widetilde{B} = (1-\epsilon)\beta I_m - (1+\epsilon)N_2$. It can be verified that

$$\widetilde{P}(y \otimes u) = (1-\delta)P(y \otimes u),$$

where $\delta = \chi\epsilon$, y, and u are as defined in the proof of Theorem 3.2. Let $C \in \mathbb{R}^{n\times m}$ be such that $\mathrm{vec}(C) = P(y \otimes u)$. Then $AX + XB = C$ has the solution $X = uy^{\mathrm{T}}$ and $\widetilde{A}\widetilde{X} + \widetilde{X}\widetilde{B} = C$ has solution $\widetilde{X} = \frac{1}{1-\delta}uy^{\mathrm{T}}$. Thus for all (i, j)

$$|(\widetilde{X} - X) \oslash X|_{(i,j)} = \frac{\delta}{1-\delta},$$

indicating asymptotically the factor $[1 - \varrho_i + (1-\varrho_j 2)\tau_i^{-1}]^{-1}\,\epsilon$ cannot be improved. A similar argument to Item 2 of Remark 2.2 can be used to get

$$\max_{\widetilde{A},\,\widetilde{B}} \max_{i,j} |(\widetilde{P} - P) \oslash P|_{(i,j)} \geq \frac{\delta}{1-\delta} \quad \text{subject to (3.11).}$$

Example 3.5 In principle, Example 3.4 covers Corollary 3.1. But we can have examples without requiring D_i being the multiples of the identity matrices. In fact, we may pick any nonsingular and irreducible M-matrices A and B with $\varrho_1 = \varrho_2$. Perturb A and B to

$$\widetilde{A} = (1-\epsilon)D_1 - (1+\epsilon)N_1, \quad \widetilde{B} = (1-\epsilon)D_2 - (1+\epsilon)N_2.$$

Let $u > 0$ and $y > 0$ be the Perron eigenvectors of $D_1^{-1}N_1$ and $D_2^{-1}N_2^{\mathrm{T}}$, respectively. We have

$$\begin{aligned} P(y \otimes u) &= (1-\varrho_{\max})(y \otimes D_1 u + D_2 y \otimes u), \\ \widetilde{P}(y \otimes u) &= [(1-\epsilon) - (1+\epsilon)\varrho_{\max}][y \otimes D_1 u + D_2 y \otimes u] \\ &= (1-\chi\epsilon)P(y \otimes u). \end{aligned}$$

The rest of construction is the same as Example 3.4.

Similarly for Theorem 3.3, we can have examples without requiring D_i being the multiples of the identity matrices as well. Take any nonsingular and irreducible

M-matrix $A = D_1 - N_1$ and a singular and irreducible M-matrix B. Let $\widetilde{A} = (1-\epsilon)D_1 - (1+\epsilon)N_1$ and $\widetilde{B} = B$. The rest of construction is similar and thus omitted.

4 Algorithms for Sylvester equations

The perturbation theorems for MSE

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

in Sect. 3 cover two cases:

1. A and B are nonsingular M-matrices, and $C \geq 0$;
2. One of A and B is a nonsingular M-matrix (while the other may not) and

$$P \stackrel{\text{def}}{=} I_m \otimes A + B^{\mathrm{T}} \otimes I_n \quad \text{is a nonsingular } M\text{-matrix},$$

and again $C \geq 0$.

Together, they cover all possible cases in our definition of an MSE: *both $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$ have positive diagonal entries and nonpositive off-diagonal entries and $P = I_m \otimes A + B^{\mathrm{T}} \otimes I_n$ is a nonsingular M-matrix, and $C \geq 0$* because this definition implies at least one of A and B is a nonsingular M-matrix. To see this, we note the eigenvalues of P are all possible sums $\mu + \nu$, where μ and ν are eigenvalues of A and B, respectively. P being a nonsingular M-matrix implies by Theorem 2.1 that all $\Re(\mu + \nu) = \Re(\mu) + \Re(\nu) > 0$, where $\Re(\cdot)$ takes the real part of a complex number. So either $\min_{\mu\in\mathrm{eig}(A)} \Re(\mu) > 0$ or $\min_{\nu\in\mathrm{eig}(B)} \Re(\nu) > 0$, where $\mathrm{eig}(\cdot)$ is the set of the eigenvalues of a matrix. Thus one of A and B must be a nonsingular M-matrix again by Theorem 2.1. Furthermore with

$$\tau = \frac{1}{2}\left[\min_{\mu\in\mathrm{eig}(A)} \Re(\mu) - \min_{\nu\in\mathrm{eig}(B)} \Re(\nu)\right] \tag{4.1}$$

all eigenvalues of both $A - \tau I_n$ and $B + \tau I_m$ have positive real parts and therefore both $A - \tau I_n$ and $B + \tau I_m$ are nonsingular M-matrices.

Without loss of generality, we may assume that both A and B are irreducible. Otherwise the Eq. (1.1) can be decomposed into a sequence of smaller Sylvester equations that can be solved sequentially and each smaller equation has the form of (1.1) with irreducible A and B. In fact, let $\Pi_1 \in \mathbb{R}^{n\times n}$ and $\Pi_2 \in \mathbb{R}^{m\times m}$ be two permutation matrices such that

$$\Pi_1^{\mathrm{T}} A \Pi_1 = \begin{pmatrix} A_{11} & -A_{12} & \dots & -A_{1q} \\ & A_{22} & \dots & -A_{2q} \\ & & \ddots & \vdots \\ & & & A_{qq} \end{pmatrix}, \quad \Pi_2^{\mathrm{T}} B \Pi_2 = \begin{pmatrix} B_{11} & & & \\ -B_{21} & B_{22} & & \\ \vdots & \vdots & \ddots & \\ -B_{p1} & -B_{p2} & \dots & B_{pp} \end{pmatrix},$$

where $A_{ij} \in \mathbb{R}^{n_i\times n_j}$, $B_{ij} \in \mathbb{R}^{m_i\times m_j}$, all A_{ii} and B_{jj} are irreducible, all $I_{m_j} \otimes A_{ii} + B_{jj}^{\mathrm{T}} \otimes I_{n_i}$ are nonsingular M-matrices, and all $A_{ij} \geq 0$ and $B_{ij} \geq 0$ for $i \neq j$.

Partition $\Pi_1^{\mathrm{T}} X \Pi_2 = (X_{ij})$ and $\Pi_1^{\mathrm{T}} C \Pi_2 = (C_{ij})$ into $q \times p$ block matrices with $X_{ij},\ C_{ij} \in \mathbb{R}^{n_i \times m_j}$. Then the (i, j)th block from both sides of $(\Pi_1^{\mathrm{T}} A \Pi_1)(\Pi_1^{\mathrm{T}} X \Pi_2) + (\Pi_1^{\mathrm{T}} X \Pi_2)(\Pi_2^{\mathrm{T}} B \Pi_2) = \Pi_1^{\mathrm{T}} C \Pi_2$ gives

$$A_{ii} X_{ij} + X_{ij} B_{jj} = C_{ij} + \sum_{\ell=i+1}^{q} A_{i\ell} X_{\ell j} + \sum_{\ell=j+1}^{p} X_{i\ell} B_{\ell j}$$

which suggests that the block columns of $\Pi_1^{\mathrm{T}} X \Pi_2$ can be computed sequentially from the last to the first and within each block column from the last block at the bottom upwards to the first block at the top. Each of these smaller equations has the form of (1.1) is an MSE with irreducible A and B and nonnegative C.

In what follows, we may assume A and B are irreducible if necessary. But most parts below do not need the irreducibility of A and/or B to work, however. So unless we explicitly state that A and B are irreducible, they may be reducible.

When A and B are (singular or nonsingular) M-matrices, we argue that it suffices to consider the case when

We have vectors $u, v \in \mathbb{R}^m$ and $y, z \in \mathbb{R}^n$ such that

$$u > 0,\ Au = v \geq 0, \quad \text{and } y > 0,\ B^{\mathrm{T}} y = z \geq 0, \tag{4.2}$$

i.e., $A = \{N_1, u, v\}$ and $B^{\mathrm{T}} = \{N_2^{\mathrm{T}}, y, z\}$.

If the vectors u, v, y, and z are given to begin with, then no argument is needed. If, however, M-matrices A and/or B are given in their usual matrix format, approximate triplet representations for A and B^{T} can be computed as described in Sect. 2.2.

4.1 Direct method

It is based on Gaussian elimination to solve, equivalently, $P\,\mathrm{vec}(X) = \mathrm{vec}(C)$. It can be seen from (4.2) that

$$P = \{I_m \otimes N_1 + N_2^{\mathrm{T}} \otimes I_n,\ y \otimes u,\ y \otimes v + z \otimes y\}. \tag{4.3}$$

Thus the GTH-like algorithm [1] applies.

When one of A and B is not an M-matrix but P is, we do not have (4.2). A triplet representation for P can still be found. Suppose that A is a nonsingular M-matrix and $A = \{N_1, u, v\}$ is either known or computed as usual. We use the idea in Sect. 2.2 to find the Perron vector $y > 0$ of $D_2^{-1} N_2^{\mathrm{T}}$, assuming B is irreducible. Since B is not an M-matrix, it is possible that some of the entries of $z = B^{\mathrm{T}} y$ are negative. Even so, we still have (4.3) if $y \otimes v + z \otimes y \geq 0$. If, however, $y \otimes v + z \otimes y \ngeq 0$, then we recompute a triplet representation for A in which $u > 0$ is the Perron vector of $D_1^{-1} N_1$ (assuming A is irreducible) and $v = Au$. If the conditions of Theorem 3.2 are satisfied, then $y \otimes v + z \otimes y \geq 0$ now. We again have (4.3).

This direct method costs $O(m^3n^3)$ and thus becomes very expensive even for modest m and n. But for small m and n, it is an ideal method for computing "exact" solutions to be used for any testing purpose.

4.2 Fixed point iterative methods

Any pair of splittings for A and B

$$A = M_1 - K_1, \quad B = M_2 - K_2 \tag{4.4}$$

gives rise to a splitting for P

$$P = \mathcal{M} - \mathcal{K}, \quad \mathcal{M} = I_m \otimes M_1 + M_2^{\mathrm{T}} \otimes I_n, \quad \mathcal{K} = I_m \otimes K_1 + K_2^{\mathrm{T}} \otimes I_n, \tag{4.5}$$

and correspondingly an iterative method for MSE (1.1): $X_0 = 0$ and for $k \geq 0$

$$M_1 X_{k+1} + X_{k+1} M_2 = K_1 X_k + X_k K_2 + C. \tag{4.6}$$

Convenient ones are those from the so-called *regular splittings* (4.5), namely $\mathcal{M}^{-1} \geq 0$ and $\mathcal{K} \geq 0$, in such a way that (4.6) is easy to solve. The following five choices are obvious ones:

$$M_1 = \mathrm{diag}(A), \quad M_2 = \mathrm{diag}(B), \tag{4.7a}$$
$$M_1 = \mathrm{tril}(A), \quad M_2 = \mathrm{triu}(B), \tag{4.7b}$$
$$M_1 = \mathrm{triu}(A), \quad M_2 = \mathrm{tril}(B), \tag{4.7c}$$
$$M_1 = \mathrm{tril}(A), \quad M_2 = \mathrm{tril}(B), \tag{4.7d}$$
$$M_1 = \mathrm{triu}(A), \quad M_2 = \mathrm{triu}(B), \tag{4.7e}$$

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

Since P is a nonsingular M-matrix, the corresponding (4.5) are all regular. Therefore [26] $\rho(\mathcal{M}^{-1}\mathcal{K}) < 1$ and the iterative method (4.6) converges and moreover

$$0 = X_0 \leq X_1 \leq X_2 \leq \cdots, \quad \lim_{k\to\infty} X_k = X. \tag{4.8}$$

Implementing (4.6), although a Sylvester equation itself, is easy for each of (4.7). This is evident for (4.7a). For (4.7b)–(4.7e), both M_1 and M_2 are triangular and (4.6) can be decomposed into a sequence of triangular linear systems[5]. A straightforward implementation of (4.6) as is always gives $X_k \geq 0$ for all k but may not preserve the

[5] For *upper* triangular M_1 and upper (lower) triangular M_2, the columns of X_{k+1} can be computed sequentially one column at a time from the first to the last (from the last to the first) by solving m *upper* triangular linear systems of order n. Similarly for *lower* triangular M_1 and upper (lower) triangular M_2, the columns of X_{k+1} can be computed sequentially one column at a time from the first to the last (from the last to the first) by solving m *lower* triangular linear systems of order n.

monotonicity in (4.8). There is a better way. From (4.6) for two consecutive steps, we have

$$M_1(X_{k+2} - X_{k+1}) + (X_{k+2} - X_{k+1})M_2 = K_1(X_{k+1} - X_k) + (X_{k+1} - X_k)K_2.$$

Set $\Delta_k = X_{k+1} - X_k$. We therefore suggest to implement (4.6) as follows:

Algorithm 4.1
Fixed Point Iterative Method for MSE $AX + XB = C$ **with** (4.4).
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 $M_1\Delta_{k+1} + \Delta_{k+1}M_2 = K_1\Delta_k + \Delta_k K_2$ for Δ_{k+1};
5 $X_{k+2} = X_{k+1} + \Delta_{k+1}$;
6 Enddo.

With each of the splittings in (4.7), Algorithm 4.1 is guaranteed to produce a *linearly* convergent sequence of X_k satisfying (4.8) with the rate of convergence $\rho(\mathcal{M}^{-1}\mathcal{K})$. 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 $\rho(\mathcal{M}^{-1}\mathcal{K})$ is very close to 1 (this usually happens when P is nearly singular) to require the number of steps so gargantuan that accumulated roundoff errors become too great to overcome.

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

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 (4.7a)–(4.7e).

4.3 Smith method

This iterative method to solve (1.1) is taken from Smith [24] and works for A and B with (4.2) and at least one of them is a nonsingular M-matrix while the other is a (singular or nonsingular) M-matrix. For any scalar μ, we have

$$(A + \mu I)X(B + \mu I) - (A - \mu I)X(B - \mu I) = 2\mu C.$$

If $\mu > 0$, then both $A + \mu I$ and $B + \mu I$ are nonsingular and thus

$$X = X_0 + F_0 X E_0,$$

where $X_0 = 2\mu(A + \mu I)^{-1}C(B + \mu I)^{-1}$, and

$$F_0 = (A + \mu I)^{-1}(A - \mu I), \quad E_0 = (B - \mu I)(B + \mu I)^{-1}.$$

Furthermore for $\mu > 0$, $\rho(F_0)\rho(E_0) < 1$ because all eigenvalues of a nonsingular M-matrix have positive real parts. So the solution of (1.1) admits the following series expansion

$$X = \sum_{i=0}^{\infty} F_0^i X_0 E_0^i$$

which can be quickly approximated by [24]

$$X_{k+1} = X_k + F_0^{2^k} X_k E_0^{2^k} \quad \text{for } k \geq 0.$$

In fact $X_k = \sum_{i=0}^{2^k-1} F_0^i X_0 E_0^i$.

For our purpose, we shall pick a μ such that

$$\mu \geq \mu_{\text{opt}} \stackrel{\text{def}}{=} \max\{\max_i A_{(i,i)}, \max_j B_{(j,j)}\}, \tag{4.9}$$

and that all $A_{(i,i)} - \mu$ and $B_{(j,j)} - \mu$ are calculated with high relative accuracy. Inequality (4.9) which can be easily satisfied ensures

$$\begin{aligned} &F_0 \leq 0, \quad E_0 \leq 0, \quad \text{and} \quad \rho(F_0)\rho(E_0) < 1, \\ &0 \leq X_k \leq X_{k+1}, \quad \text{and} \quad 0 \leq X - X_k = F_0^{2^k} X E_0^{2^k} \to 0 \quad \text{as } k \to \infty. \end{aligned}$$

The convergence is quadratic, and asymptotically

$$\left[\frac{\|X - X_k\|}{\|X\|}\right]^{1/2^k} \sim \rho(E_0)\rho(F_0). \tag{4.10}$$

Since $\rho(E_0)$ and $\rho(F_0)$ are decreasing functions of μ subject to (4.9), we should pick μ as small as possible. The requirement that all $A_{(i,i)} - \mu$ and $B_{(j,j)} - \mu$ are calculated with high relative accuracy, together with (4.9), ensure that E_0 and F_0 are computed entrywise with high relative accuracy. This is because $A + \mu I \geq A$ and $B + \mu I \geq B$ are M-matrices and thus $(A + \mu I)^{-1} \geq 0$ and $(B + \mu I)^{-1}$ can be computed with entrywise relative errors no worse than about $(1 - \varrho_1)^{-1}\boldsymbol{u}$ and $(1 - \varrho_2)^{-1}\boldsymbol{u}$, respectively (see Sect. 2.2) and because $\mu I - A \geq 0$ and $\mu I - B \geq 0$. In order to make sure that all $A_{(i,i)} - \mu$ and $B_{(j,j)} - \mu$ are calculated with high relative accuracy, we consider two cases:

1. If all $A_{(i,i)}$ and $B_{(j,j)}$ are known to be exact floating point numbers[6], we take $\mu = \mu_{\text{opt}}$.
2. If all $A_{(i,i)}$ and $B_{(j,j)}$ are contaminated with tiny relative errors to begin with or due to decimal-to-binary conversions, we may simply take, e.g., $\mu = \eta \cdot \mu_{\text{opt}}$ to avoid possible catastrophic cancelations for some $\eta > 1$ but not too close to 1, and at the same time not to degrade too much the rate of convergence as given by (4.10).

[6] Today most floating point number systems are binary and conform to the IEEE floating point standards [3,4]. No catastrophic cancelation can occur in computing $\alpha - \beta$ for two floating point numbers because the IEEE standards ensure either $\text{fl}(\alpha - \beta) = \alpha - \beta$ exactly when $\alpha - \beta$ is a floating number or suffers a rounding error no more than half unit in its last place [10].

We now formulate the algorithm as Algorithm 4.2, and then comment on its implementation detail afterwards.

Algorithm 4.2
Smith Algorithm for $AX + XB = C$ **with** (4.2).

1 Pick $\mu \geq \max\left\{\max_i A_{(i,i)}, \max_j B_{(j,j)}\right\}$ in such a way that
no catastrophic cancelations in calculating all $A_{(i,i)} - \mu$ and $B_{(j,j)} - \mu$;
2 $A_\mu \stackrel{\text{def}}{=} A + \mu I = \{N_1, u, v + \mu u\}$, $B_\mu^{\mathrm{T}} \stackrel{\text{def}}{=} B^{\mathrm{T}} + \mu I = \{N_2^{\mathrm{T}}, y, z + \mu y\}$;
3 Compute, by the GTH-like algorithm,
$E_0 = B_\mu^{-1}(B - \mu I)$; $F_0 = A_\mu^{-1}(A - \mu I)$; $X_0 = 2\mu A_\mu^{-1} C B_\mu^{-1}$;
4 $X_1 = X_0 + F_0 X_0 E_0$;
5 For $k = 1, 2, \ldots$, until convergence
6 $\quad E_k = E_{k-1}^2$, $F_k = F_{k-1}^2$;
7 $\quad X_{k+1} = X_k + F_k X_k E_k$;
8 Enddo.

Remark 4.1 We have already commented on how to choose μ. We now go down the lines in the rest of the algorithm.

1. Use the triplet representations for A_μ and B_μ^{T} to make the GTH-like algorithm applicable to ensure that X_0, E_0, and F_0 are computed with deserved entrywise relative accuracy. Note, unlike solving (1.1) through $P\,\mathrm{vec}(X) = \mathrm{vec}(C)$, that there is no need to insist having a triplet representation for B^{T}. The algorithm works equally well with the availability of a triplet representation for B.
2. From Line 4 and forward, there is no single substraction involved. Thus all computations are entrywise forward stable.
3. It remains to explain when to stop the iteration to make sure the last X_k has desired entrywise relative accuracy as an approximation to the solution X. To this end, we borrow an idea from Prof. W. Kahan (University of California at Berkeley) who taught it to the third author in the mid-1990s [21].
Consider a nonnegative, monotonically increasing, and convergent sequence $\{\alpha_i\}$, i.e.,

$$0 < \alpha_i \leq \alpha_{i+1}, \quad \lim_{i\to\infty} \alpha_i = \alpha. \tag{4.11}$$

Let $\Delta_i = \alpha_i - \alpha_{i-1}$. *If* Δ_{j+1}/Δ_j *is decreasing for* $j \geq k$, *and if* $\Delta_{k+1}/\Delta_k < 1$, *then*

$$\alpha - \alpha_{k+1} < \frac{(\alpha_{k+1} - \alpha_k)^2}{(\alpha_k - \alpha_{k-1}) - (\alpha_{k+1} - \alpha_k)}. \tag{4.12}$$

Note the right-hand side of (4.12) coincides with the correction of Aitken's Δ^2-method for speeding up a linearly convergent sequence [25, p. 312]. To see (4.12), we let $\delta = \Delta_{k+1}/\Delta_k$. For $i \geq k$, we have

$$0 \leq \Delta_{i+1} = \prod_{j=k}^{i} \frac{\Delta_{j+1}}{\Delta_j}\,\Delta_k$$

which gives $0 \le \Delta_{i+1} \le \delta^{i-k+1}\Delta_k$. Therefore

$$\alpha - \alpha_{k+1} = \sum_{i=k+1}^{\infty} \Delta_{i+1} \le \Delta_k \sum_{i=k+1}^{\infty} \delta^{i-k+1} = \frac{\Delta_{k+1}^2}{\Delta_k - \Delta_{k+1}}$$

which gives (4.12). So to compute α with relative accuracy ϵ, it suffices to stop the iteration as soon as

$$\frac{(\alpha_{k+1} - \alpha_k)^2}{(\alpha_k - \alpha_{k-1}) - (\alpha_{k+1} - \alpha_k)} \le \epsilon\, \alpha_{k+1}. \tag{4.13}$$

A stopping criterion can then be easily drawn from this for the loop in Lines 5–8.

4. In our numerous tests (not all reported later in the next section), such a stopping criterion with ϵ about the machine unit roundoff $\boldsymbol{u}$ works very well for the Smith algorithm which converges quadratically but not so for Algorithm 4.1 which converges linearly. This is because for a linearly convergent sequence $\{\alpha_i\}$ satisfying (4.11) there is no guarantee that Δ_{j+1}/Δ_j is eventually decreasing, and consequently (4.12) is no longer valid. We argue that for a linearly convergent sequence $\{\alpha_i\}$, a simple test whether $|\alpha_{k+1} - \alpha_k| \le \epsilon|\alpha_{k+1}|$ would work just fine. Let r be the rate of convergence. Near convergence, $\alpha_{k+1} - \alpha_k = \alpha_{k+1} - \alpha - (\alpha_k - \alpha) \approx (1 - 1/r)(\alpha_{k+1} - \alpha)$, and $|\alpha_{k+1}| \le |\alpha_{k+1} - \alpha| + \alpha$. So $|\alpha_{k+1} - \alpha_k| \le \epsilon|\alpha_{k+1}|$ implies approximately

$$\alpha - \alpha_{k+1} \lessapprox \frac{r}{1 - r(1+\epsilon)}\,\epsilon\alpha.$$

That is usually sufficient, unless r is too close to 1.

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 relatively accurate numerical solutions. In what follows, We will use two error measures to gauge accuracy in computed solution $\widehat{X}$: the Normalized Residual (`NRes`)

$$\texttt{NRes} = \frac{\|A\widehat{X} + \widehat{X}B - C\|_1}{\|\widehat{X}\|_1(\|A\|_1 + \|B\|_1) + \|C\|_1}, \tag{5.1}$$

a commonly used measure for gauging $\widehat{X}$'s accuracy because it can be easily computed, and the entrywise relative error (`ERErr`),

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

which is not available in actual computations but is made available here for our testing purpose. Ideally both errors are 0, but numerically they can only be made about as tiny as $O(\boldsymbol{u})$. As we will see, for our purpose of getting $\widehat{X}$ with deserved entrywise relative accuracy, tiny `NRes` (as tiny as $O(\boldsymbol{u})$) are not sufficient.

Example 5.1 $A, B, C \in \mathbb{R}^{n\times n}$ are given by

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

This is a very well-conditioned case with

$$A\mathbf{1}_n = 2\mathbf{1}_n, \quad \mathbf{1}_n^{\mathrm{T}} B = 2\mathbf{1}_n^{\mathrm{T}}, \quad \varrho_1 = \varrho_2 = 1/3.$$

As n gets (modestly) big, X's entries show wide variations in magnitude. For testing purpose, we computed an "exact" solution X for $n = 100$ by the computerized algebra system *Maple*[7]. This "exact" solution X's entries range from 9.7×10^{-49} to 0.17. MATLAB's `lyap` which is based on Bartels and Stewart [5] fails to compute $\widehat{X}$ with all entries nonnegative, giving

$$\texttt{ERErr} = 2.3 \times 10^{+32}, \quad \texttt{NRes} = 8.8 \times 10^{-15},$$

as expected. Further looking into the solution by `lyap` reveals that some of X's entries of order $O(10^{-16})$ are computed to negative numbers but still in the order of $O(10^{-16})$. The Smith algorithm with Kahan's stopping criterion works extremely well: in just seven iterations, it produces a computed $\widehat{X}$ with entrywise relative errors smaller than 3×10^{-15}. The convergence to X's entries of different magnitudes by the five fixed point iterations in Sect. 4 is very much uneven, even though their `NRes` are reduced at predictable rates, except for `FPe` which is atypical for the example.

Figure 1 displays the convergence history in terms of `NRes` and entrywise relative errors for the Smith algorithm, `FPa`, `FPb`, and `FPe`. The history curves for `FPc` and `FPd` are nearly indistinguishable from those for `FPb` and `FPa`, respectively. One should not be surprised by the superior performance of `FPe`. This is in large part due to the structure of A (and B) and how the splitting (4.7e) is done. Except for `FPe` which is remarkably fast, the curves for `NRes` look very nice—decreasing substantially every steps; the curves for entrywise relative errors, however, show very little improvements for the first many iterations. This is due to the fact that X has (many) tiny entries.

It is worth pointing out that despite that the fixed point iterations `FPa`, `FPb`, `FPc`, and `FPd` take many more iterations than the Smith algorithm, it does not imply that they are more expensive *for this example* because each step of these fixed point iterations takes $O(n^2)$ flops while each step of the Smith algorithm takes $O(n^3)$ flops! In

[7] http://www.maplesoft.com/.

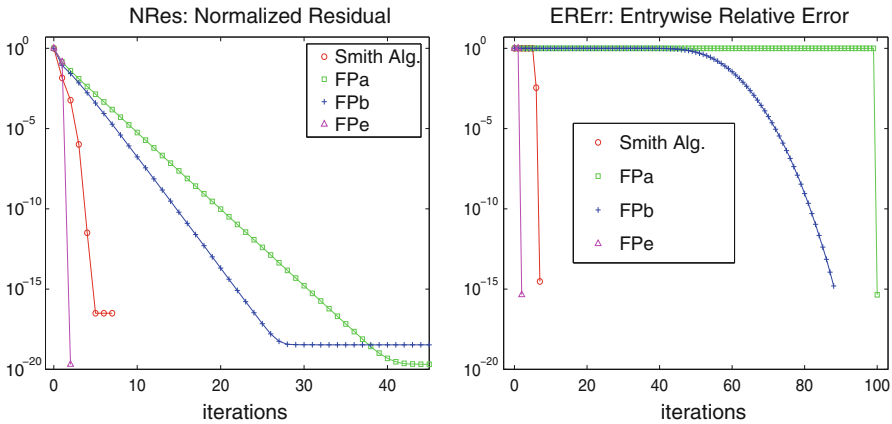

Fig. 1 Example 5.1, $n = 100$. Convergence history for the Smith algorithm and the three fixed point iterations. *Left* `NRes`; *Right* entrywise relative errors. Curves for `FPb` and `FPc` are indistinguishable, and the same can be said for the curves for `FPa` and `FPd`. Curves for `FPe` are atypical

general the flop counts for the fixed point iterations can be up to $O(n^3)$ per step but can be much less if A and B are very sparse. The sparsity of A and B does not affect the flop counts for the Smith algorithm, however.

Next we relatively perturb each entries of A, B, and C 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, and C is multiplied by

$$1 + (\texttt{rand} - .5) * \Gamma * \texttt{eps}, \tag{5.2}$$

where Γ is an adjustable parameter. We then compute the solution of the perturbed MSE (1.5) by the Smith algorithm with Kahan's stopping criterion. Let ϵ be the smallest one to satisfy (3.11).

Figure 2 plots entrywise relative errors in the solution of (1.5) against ϵ (caused by letting $\Gamma = 200 \cdot 10^i$ for $1 \le i \le 7$), very much as predicted by Theorem 3.1 and Corollary 3.1, except the factor $2n^2$ which seems to overestimate the errors.

Example 5.2 This is from modifying Example 3.1. We take $m = n$, $C = I_n$, and

$$A = B = I_n - \omega U_n - \theta e_n e_1^{\mathrm{T}}.$$

This is an M-matrix if $1 - \theta\omega^{n-1} > 0$, and in fact

$$\begin{aligned} Au &= v \quad \text{for } u = (\omega^{n-1}, \ldots, \omega, 1)^{\mathrm{T}} \text{ and } v = (0, \ldots, 0, 1 - \theta\omega^{n-1})^{\mathrm{T}}, \\ B^{\mathrm{T}} y &= z \quad \text{for } y = (1, \omega, \ldots, \omega^{n-1})^{\mathrm{T}} \text{ and } z = (1 - \theta\omega^{n-1}, 0, \ldots, 0)^{\mathrm{T}}. \end{aligned}$$

While we have done tests for many values of ω and θ and observed similarly behaviors, what we will report below is for $\omega = 3$ and $\theta\omega^{n-1} = 1/3$. Since we are sure that

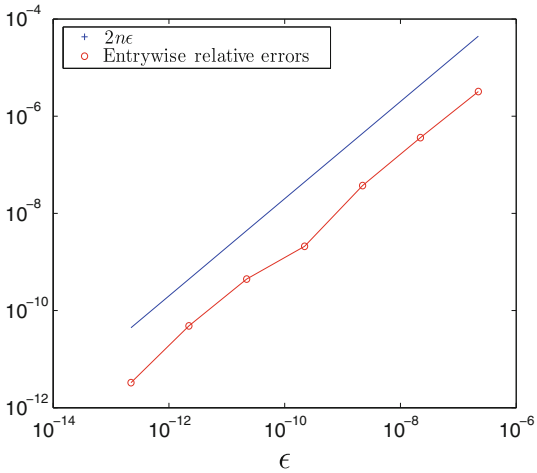

Fig. 2 Example 5.1, $n = 100$: Entrywise relative errors in $\widetilde{X}$ as ϵ varies. It is estimated via a least squares fit that these entrywise relative errors behave like 14.5ϵ. Also plotted is $n\frac{1+\varrho_{\max}}{1-\varrho_{\max}}\epsilon = 2n\epsilon$. The error bound up to the first order by Corollary 3.1 is $2mn\frac{1+\varrho_{\max}}{1-\varrho_{\max}}\epsilon = 4n^2\epsilon$ which may overestimate relative errors too much. Accidentally, $\gamma = 100$ in the first order error bound (3.19); so $2\gamma\epsilon = 2n\epsilon$ here

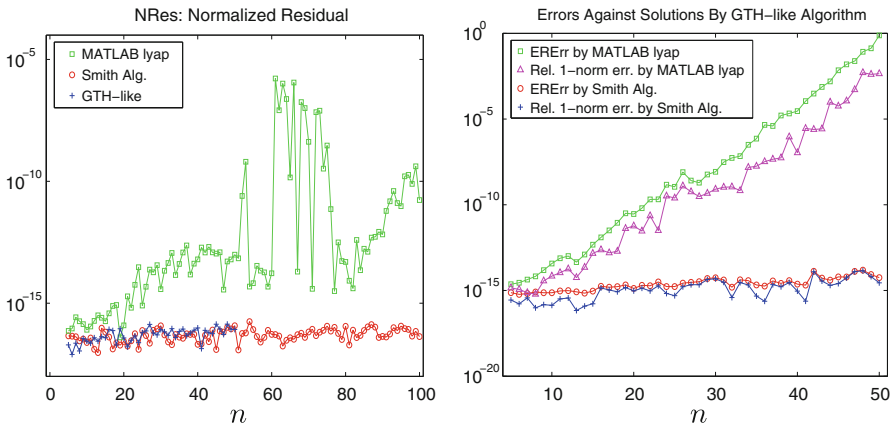

Fig. 3 Example 5.2. *Left* NRes for the computed solutions by MATLAB's `lyap`, the Smith algorithm, and the GTH-like algorithm; *Right* Entrywise relative errors and relative norm errors against the "exact" solutions computed by the GTH-like algorithm for $n \leq 50$

the direct method based on the GTH-like algorithm will deliver entrywise relatively accurate solutions (to almost the full working precision), we regard those solutions as the "exact" ones to check the errors in $\widehat{X}$ computed by the Smith algorithm, the fixed point iterations, and MATLAB's `lyap`. Accuracy results for the fixed point iterations are not reported here as they are similar to those for the Smith algorithm. The high computational complexity $O(n^6)$ of the GTH-like algorithm limits us to cap n at 50 for running the algorithm in MATLAB on a PC. Figure 3 shows the numerical results as n varies. Its left plot displays NRes as defined by (5.1). It shows the overall tendency of NRes getting worse as n increases for MATLAB's `lyap` while NRes for the Smith algorithm and the direct method seem to behave independently of n. Also it shows the unpredictable behavior of NRes for MATLAB's `lyap` for n beyond about 50: NRes can vary between nearly 10^{-15} to almost 10^{-5} with little regard about how big n may be.

In the right plot of Fig. 3, we show entrywise relative errors and

$$\text{Relative 1-norm errors: } \|\widehat{X} - X\|_1 / \|X\|_1$$

of $\widehat{X}$ computed by the Smith algorithm and MATLAB's `lyap` against the "exact" X computed by the GTH-like algorithm. The curves for MATLAB's `lyap` move up (rather quickly) while those for the Smith algorithm stay almost flat. There is a good explanation as to why the error curves for MATLAB's `lyap` behave this way. Since the method is backward stable, we have for MATLAB's `lyap`

$$\|\widehat{X} - X\|_1 / \|X\|_1 \le O(\|P\|_1 \|P^{-1}\|_1 \boldsymbol{u}), \tag{5.3}$$

and as a consequence

$$\begin{aligned} \max_{i,j} \frac{|\widehat{X}_{(i,j)} - X_{(i,j)}|}{X_{(i,j)}} &\le \frac{\|\widehat{X} - X\|_1}{\|X\|_1} \frac{\|X\|_1}{\min_{i,j} X_{(i,j)}} \\ &\le O(\|P\|_1 \|P^{-1}\|_1 \boldsymbol{u}) \times \frac{\max_{i,j} X_{(i,j)}}{\min_{i,j} X_{(i,j)}}. \end{aligned} \tag{5.4}$$

For this example $\|P\|_1 \|P^{-1}\|_1$ increases fast as n increases. So both bounds in (5.3) and (5.4) increase fast, too, as n increases.

6 Concluding remarks

It seems to be more *natural* to call (1.1) an MSE if both A and B are M-matrices, $P = I_m \otimes A + B^{\mathrm{T}} \otimes I_n$ is nonsingular, and $C \ge 0$. Our definition of an MSE seemingly is more broad, but actually equivalent to this *natural* one because $AX + XB = C$ is the same as $(A - \tau I)X + X(B + \tau I) = C$ for any scalar τ and with τ given by (4.1) both $A - \tau I$ and $B + \tau I$ are nonsingular M-matrices.

We have presented an entrywise perturbation analysis for MSE (1.1). It is proved small relative perturbations to the entries of A, B, and C will only cause small relative changes to each entry of the solution X, regardless of its magnitude. We argued that the linear terms in our bounds are asymptotically best possible, except their dimensionally dependent factor $2mn$ which we conjectured could be replaced by something like $2(m + n)$.

The commonly used first order error analysis can be easily performed, too, as we outlined in Remark 3.2, to yield a sharp and easily implementable first order error bound. But our new analysis leads to more insightful bounds in that the effect of $\varrho(A)$ and $\varrho(B)$ in the solution's sensitivity is exposed.

We showed that the GTH-like algorithm [1], the Smith algorithm [24], and the classical fixed point iterations [26] with some minor but crucial implementation changes can deliver computed solutions with predicted entrywise relative accuracy according to our analysis. Our numerical results confirm our analysis and our accuracy claims about the algorithms. We point out in passing that the condition $C \ge 0$ is not necessary

to ensure convergence of X_k to X by any of the iterative methods in Sect. 4; it is just that without $C \geq 0$ convergence is no longer monotonic.

This is the first paper of ours in a sequel of two on accurate solutions of MSE and M-matrix algebraic Riccati equation (1.3) which is more difficult to analyze than MSE because of its nonlinearity in X. The latter will be the subject of our investigation in [31].

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. Comput. **71**, 217–236 (2002)
2. Alfa, A.S., Xue, J., Ye, Q.: Entrywise perturbation theory for diagonally dominant M-matrices with applications. Numer. Math. **90**(3), 401–414 (2002)
3. American National Standards Institute and Institute of Electrical and Electronic Engineers: IEEE standard for binary floating-point arithmetic. ANSI/IEEE Standard, Std 754-1985, New York (1985)
4. American National Standards Institute and Institute of Electrical and Electronic Engineers: IEEE standard for radix independent floating-point arithmetic. ANSI/IEEE Standard, Std 854-1987, New York (1987)
5. Bartels, R.H., Stewart, G.W.: Algorithm 432: The solution of the matrix equation $AX - BX = C$. Commun. ACM **8**, 820–826 (1972)
6. Benner, P., Li, R.-C., Truhar, N.: On ADI method for Sylvester equations. J. Comput. Appl. Math. **233**(4), 1035–1045 (2009)
7. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences. SIAM, Philadelphia, 1994. This SIAM edition is a corrected reproduction of the work first published in 1979 by Academic Press, San Diego, CA
8. Elsner, L., Koltracht, I., Neumann, M., Xiao, D.: On accurate computations of the Perron root. SIAM J. Matrix Anal. Appl. **14**(2), 456–467 (1993)
9. Gohberg, I., Koltracht, I.: Mixed, componentwise, and structured condition numbers. SIAM J. Matrix Anal. Appl. **14**(3), 688–704 (1993)
10. Goldberg, D.: What every computer scientist should know about floating-point arithmetic. ACM Comput. Surv. **23**(1), 5–47 (1991)
11. 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)
12. Grassmann, W.K., Taksar, M.J., Heyman, D.P.: Regenerative analysis and steady-state distributions for Markov chains. Oper. Res. **33**, 1107–1116 (1985)
13. Guo, C., Higham, N.: Iterative solution of a nonsymmetric algebraic Riccati equation. SIAM J. Matrix Anal. Appl. **29**, 396–412 (2007)
14. Guo, C.-H.: Nonsymmetric algebraic Riccati equations and Wiener–Hopf factorization for M-matrices. SIAM J. Matrix Anal. Appl. **23**, 225–242 (2001)
15. 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)
16. Guo, X., Lin, W., Xu, S.: A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation. Numer. Math. **103**, 393–412 (2006)
17. Higham, N.J.: Accuracy and Stability of Numerical Algorithms. 2nd edn. SIAM, Philadephia (2002)
18. Horn, R.A., Johnson, C.R.: Topics in Matrix Analysis. Cambridge University Press, Cambridge (1991)
19. Juang, J.: Existence of algebraic matrix Riccati equations arising in transport theory. Linear Algebra Appl. **230**, 89–100 (1995)

20. Juang, J., Lin, W.-W.: Nonsymmetric algebraic Riccati equations and Hamiltonian-like matrices. SIAM J. Matrix Anal. Appl. **20**(1), 228–243 (1998)
21. Li, R.-C.: Solving secular equations stably and efficiently. Technical Report UCB//CSD-94-851, Computer Science Division, Department of EECS, University of California at Berkeley (1993)
22. Ramaswami, V.: Matrix analytic methods for stochastic fluid flows. In: Key, P., Smith, D. (eds.) Teletraffic Engineering in a Competitive World, vol. **3a** of Teletraffic Science and Engineering. Elsevier Science, Amsterdam, pp. 1019–1030 (1999)
23. Rogers, L.: Fluid models in queueing theory and Wiener–Hopf factorization of Markov chains. Ann. Appl. Probab. **4**, 390–413 (1994)
24. Smith, R.A.: Matrix equation $XA + BX = C$. SIAM J. Appl. Math. **16**(1), 198–201 (1968)
25. Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis. 2nd edn. Springer-Verlag, Berlin (1992)
26. Varga, R.S.: Matrix Iterative Analysis. Englewood Cliffs, NJ (1962)
27. Virnik, E.: Analysis of positive descriptor systems. PhD thesis, Technischen Universität Berlin, Berlin (2008)
28. Wachspress, E.L.: The ADI Model Problem. Windsor, CA (1995) (self-published) (www.netlib.org/na-digest-html/96/v96n36.html)
29. Xue, J.: Computing the smallest eigenvalue of an M-matrix. SIAM J. Matrix Anal. Appl. **17**(4), 748–762 (1996)
30. Xue, J., Jiang, E.: Entrywise relative perturbation theory for nonsingular M-matrices and applications. BIT **35**(3), 417–427 (1995)
31. Xue, J., Xu, S., Li, R.-C.: Accurate solutions of M-matrix algebraic Riccati equations. Numer. Math. doi:10.1007/s00211-011-0421-0