An Error Estimate for Matrix Equations

Yang Cao and Linda Petzold *

Department of Computer Science, University of California Santa Barbara
CA 93106, USA.

Abstract.

This paper proposes a new method for estimating the error in the solution of matrix equations. The estimate is based on the adjoint method in combination with small sample statistical theory. It can be implemented simply and is inexpensive to compute. Numerical examples are presented which illustrate the power and effectiveness of the new method.

Keywords: condition number, adjoint method, Lyapunov equation, Sylvester equation.

Mathematics Subject Classification: 65F35, 65F30, 15A12.

1 Introduction

Matrix equations play an important role in applied mathematics and control theory. Examples of well-known matrix equations include the Sylvester equation

\[ AX - XB = C, \]

where \( A \in R^{n \times n} \), \( B \in R^{m \times m} \), \( C \in R^{n \times m} \), and \( X \in R^{n \times m} \) is to be determined. The Sylvester equation arises in many applications. For example, the finite difference discretization of a separable elliptic boundary value problem on a rectangular domain can be written in the form of a Sylvester equation [21]. The Sylvester equation arises also in the design of reduced order observers [2], and in many eigenproblems [1], [11]. When \( B = -A^T \), the Sylvester equation reduces to the Lyapunov equation

\[ AX + XA^T = C. \]

Important applications of the Lyapunov equation include model reduction [3], [17], stability analysis of linear systems [19] and in the solution of another important matrix equation: the algebraic Riccati equation

\[ A^T X + AX - XG X = W; \]

*This work was supported by grants: NSF/ITR ACI-0086061, NSF/KDI ATM-9873133 and DOE DE-F603-00ER25430.
which is widely applied in optimal control theory [15], [16].

In this paper we are concerned with estimating the error in the numerical solution of matrix equations. The error in the solution arises from many sources, such as the uncertainty in the data or the round-off error in the numerical solution. The classic method is to consider all these sources as the perturbations to the data. Thus the numerical solution is studied as the exact solution of the perturbed system. The error in the numerical solution is then the effect of the perturbation. The condition number provides a norm bound to quantitatively measure this effect. The actual error is then bounded in norm by the product of the condition number and the corresponding norm bound of the perturbations. In practice we can obtain a bound on the perturbations according to different applications. Thus the condition number is essential for estimating the error in the solution and has been a major topic in perturbation theory.

Before proceeding, we must specify the norm with which to measure the error. In the literature of perturbation theory for matrix equations, the Frobenius norm is most often used because it is the vector 2-norm when the system is formulated as a linear system (see [13] as an example). Perturbation theory based on other norms has also been discussed for some special cases. In particular, a highly effective simple method for estimating the error in the spectral norm for the stable Lyapunov equation and the algebraic Riccati equation has been proposed in [9], [12], [16]. We briefly describe this method for the stable Lyapunov equation. Suppose $A$ is stable. Thus all the eigenvalues of $A$ are in the left-half plane. For the Lyapunov equation (2), the solution $X$ has the form

$$X = \int_0^\infty e^{At}Ce^{AT}dt.$$  

Consider the perturbed Lyapunov equation

$$ \begin{equation}
(A + \Delta A)(X + \Delta X) + (X + \Delta X)(A + \Delta A)^T = C + \Delta C.
\end{equation}
$$

The corresponding error $\Delta X$ also has the form

$$ \begin{equation}
\Delta X = \int_0^\infty e^{At}(\Delta C + \Delta A(X + \Delta X) + (X + \Delta X)\Delta A^T)e^{AT}dt.
\end{equation}
$$

Let $u$ and $v$ denote the left and right singular vectors of unit length of $\Delta X$, such that

$$u^*\Delta XV = \|\Delta X\|_\rho,$$

where $\|\cdot\|_\rho$ denotes the spectral norm. We have

$$\|\Delta X\|_\rho \leq \int_0^\infty u e^{At}(\Delta C + \Delta A(X + \Delta X) + (X + \Delta X)\Delta A^T)e^{AT}v dt.$$

Then the spectral norm condition number is given by [12]

$$ \begin{equation}
K = 2\|A\|_\rho\|H\|_\rho,
\end{equation}
$$

where $H$ solves the equation $AH + HA^T = -I$, and $\int_0^\infty \|e^{At}u\|^2 dt \leq \|H\|$ for any unit length $u$. The relative error is bounded by

$$ \begin{equation}
\frac{\|\Delta X\|_\rho}{\|X + \Delta X\|_\rho} \leq K \left[ \frac{\|\Delta A\|_\rho}{\|A + \Delta A\|_\rho} + \frac{\|\Delta C\|_\rho}{\|C + \Delta C\|_\rho} \right].
\end{equation}$$

2
The relation (7) is also used in [9] to estimate the error in the Frobenius norm although no proof is given. This error bound is specific to the stable Lyapunov equation, because only with stability do we have the formula (5). For the unstable Lyapunov equation, a counterexample can be found in [9]. There is a formula similar to (5) for the stable Sylvester equation. However, a similar generalization based on that formula requires the solution of the matrix equations \(AH_1 + H_1A^T = -I\) and \(BH_2 + H_2B^T = -I\), which is much more costly than the case of the stable Lyapunov equation.

In this paper we focus on the Frobenius norm for the general matrix equation. In the following, the norm refers to the Frobenius norm for matrices or the 2-norm for vectors if not otherwise specified. To simplify the presentation, we will take the Sylvester equation as our example. Numerical results will be given for both the Sylvester and Lyapunov equations. For the Sylvester equation, it is well known that when \(A\) and \(B\) have no eigenvalues in common, the Sylvester equation (1) is nonsingular. Numerical algorithms for the solution of the Sylvester equation can be found in [10], [11], [13]. Perturbation theory for the Sylvester equation has been studied as a generalization of perturbation theory for linear systems [8], [13]. For this purpose, (1) is written in the form

\[
(I_m \otimes A - B^T \otimes I_n)\text{vec}(X) = \text{vec}(C),
\]

where \(A \otimes B = (a_{ij}B)\) is the Kronecker product, and the \(\text{vec}\) operator stacks the columns of a matrix into one long vector. An error bound is then formed using the condition number of the linear system (8). Here we have to be very careful because the transformation from (1) to (8) may cause an ill-conditioned system to be misunderstood to be a well-conditioned one. For example, in [8] the condition number when \(n = m\) is defined as

\[
k(A, B) = \frac{\max_i \sigma_i(I \otimes A - B^T \otimes I)}{\min_i \sigma_i(I \otimes A - B^T \otimes I)}.\]

This is the 2-norm condition number of the linear system (8), but it may not always yield an accurate measurement of the condition of the Sylvester equation (1). Later in Example 1 we will see that (9) gives a small condition number for an ill-conditioned system.

The proper definition of the condition of a matrix equation is discussed in [13]. Consider the perturbed Sylvester equation

\[
(A + \Delta A)(X + \Delta X) - (X + \Delta X)(B + \Delta B) = C + \Delta C,
\]

where \(\|\Delta A\| \leq \epsilon \alpha, \|\Delta B\| \leq \epsilon \beta,\) and \(\|\Delta C\| \leq \epsilon \gamma.\) Let \(P = I_m \otimes A - B^T \otimes I_n.\) After dropping the second-order terms, we obtain

\[
A\Delta X - \Delta XB = \Delta C - \Delta AX + X\Delta B.
\]

Writing this system in the form

\[
P\text{vec}(\Delta X) = \begin{bmatrix} X^T \otimes I_n & -I_m \otimes X & I_{mn} \end{bmatrix} \begin{bmatrix} \text{vec}(\Delta A) \\ \text{vec}(\Delta B) \\ \text{vec}(\Delta C) \end{bmatrix},
\]

the condition number (for Frobenius norm) is given by ([13], p. 318)

\[
\Phi = \|P^{-1}[\alpha (X^T \otimes I_n) - \beta (I_m \otimes X) - \gamma I_{nm}]\|_2/\|X\|_F.
\]
This can be simplified and weakened to the more often quoted condition number (see [12], for example)

\[
\Psi = \|P^{-1}\|_2 \frac{(\alpha + \beta)\|X\|_F + \gamma}{\|X\|_F}.
\]

The error bound is then given by

\[
\frac{\|\Delta X\|_F}{\|X\|_F} \leq \sqrt{3\Phi \epsilon},
\]

or

\[
\frac{\|\Delta X\|_F}{\|X\|_F} \leq \sqrt{3\Psi \epsilon}.
\]

It is well known[13] that (11) generates a sharper condition number than (12) but the computation of (11) is more expensive.

We note that an important difference between (9) and (11) or (12) is that the condition numbers in (11) and (12) are defined directly from the Sylvester equation, whereas (9) includes the effect of the transformation. Consider the following example

**Example 1** Let

\[
A = \begin{bmatrix} 1 + \epsilon & 2\epsilon \\ \epsilon & 1 - \epsilon \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} \epsilon & 2\epsilon \\ \epsilon & -\epsilon \end{bmatrix},
\]

where \(\epsilon\) is very small. Definition (9) yields a condition number of 1. Both (11) and (12) yield large condition numbers. Actually when \(\epsilon\) is small, the eigenvalues of \(A\) are very close to the eigenvalues of \(B\), which makes this system close to a singular one. For example, taking \(\epsilon = 10^{-14}\), (11) yields a condition number of \(1.4 \times 10^{14}\). This is quite large. When solving this system in Matlab, the numerical solution is \(X = \begin{bmatrix} 1.0003 & -0.0005 \\ 0.0003 & 1.0003 \end{bmatrix}\) while the exact solution is \(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\). Thus the relative error is \(5 \times 10^{-4}\), which is much larger than the estimate (9) would indicate. Another difference between (9) and (11) or (12) is that (11) and (12) are related to the right-hand side. But this is not the reason that the error estimate based on (9) is an underestimate in this example. We can weaken (12) further to \(2\|P^{-1}\|_2(\alpha + \beta)\) to make it unrelated to the right-hand side. It still generates a large condition number while (9) does not. The reason (9) gives a misleading condition number is that during the transformation from (1) to (8), it includes the step \((1 + \epsilon) - 1 = \epsilon\). When \(\epsilon\) is small, this transformation is ill-conditioned.

Although both (11) and (12) can give a proper condition number of the system, the computational cost to determine \(\|P^{-1}\|_2\) is high. Iterative methods are generally used to estimate this term. Moreover, both (11) and (12) are based on norm-wise analysis. It is assumed that the norm of the perturbations is bounded by \(\epsilon\) times the norm of the data. This assumption cannot take into account the structure of the perturbation. Thus for structured problems, norm-wise estimates yield an overestimate. For example, consider the following well-conditioned system
Example 2
\[
\begin{bmatrix}
2 & 0 \\
0 & \delta
\end{bmatrix} X - X \begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix} = \begin{bmatrix}
1 & 0 \\
0 & \delta
\end{bmatrix},
\]
which has the unique solution \( X = \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix} \). The norm-wise condition number gives \( \Phi = O(\frac{1}{\delta}) \), \( \Psi = O(\frac{\delta}{1}) \). When \( \delta \) is small, the norm-wise condition number is large. But because the system has a diagonal structure, this system is actually well-conditioned. When solving this system in Matlab, even when we take \( \delta \) to be \( 10^{-14} \), the numerical solution is still accurate although the condition number given by (11) is \( 10^{14} \). In this case, we need a condition number which can reflect the structure. Component-wise analysis, which has been proposed for linear systems [5], [13], [20], can be used to derive a proper condition number in this case. It scales the perturbation with the data and usually gives a sharper error estimation. But it has not yet been applied to the perturbation theory of matrix equations, to the best of our knowledge.

Yet another motivation for an improved error estimation method is the possibility of providing a subspace condition estimate. For some problems, we may be concerned with a subspace of the solution rather than the full space. In this case, even if the system is ill-conditioned, we may still be able to safely obtain a solution in a well-conditioned subspace. Such examples for linear systems can be found in [5], [6].

Thus we seek an error estimation method which is general, efficient and can utilize the structure of the system. Some progress has already been made towards this end. The small sample statistical estimation (SSSE) method [14] uses finite differencing combined with small sample Monte Carlo theory to derive an efficient method to estimate the error. The SSSE method is promising, but the finite difference method leads to concern about the perturbation stepsize selection. Recently a new method for error estimation of linear systems was proposed in [5]. It yields the component-wise error estimate. Thus it can utilize the structure of the system and scale the perturbation with the data. It can also be used to obtain a subspace error estimate. The computational cost is low, and with high probability the result is accurate to within a factor of 10, which is the usual requirement for error estimation. The purpose of this paper is to derive this method in a more general way and apply it to matrix equations. We derive our method by combining the adjoint sensitivity method, instead of finite difference, with the SSSE method. Numerical experiments show that this method gives a sharper error estimate than (14) for the general Sylvester equation. It is much cheaper than (13), while the accuracy is similar. It yields a sharper error estimate for matrix equations with some special structures. For the stable Lyapunov equation, our method yields an error estimate which is sharper than (7), but at a higher computational cost. This paper is organized as follows. In Section 2 we introduce the method of adjoint sensitivity analysis for matrix equations. In Section 3 we present the error estimation technique, which is based on adjoint sensitivity analysis and the small sample statistical method[14]. Numerical results are given in Section 4.

2 Sensitivity Analysis for Matrix Equations

Consider the original system (1) and the perturbed system (10). Following [18], we treat the perturbations as parameters to the original system. Each piece of data in \( A, B \) and \( C \) is treated as a parameter. The error in the solution is estimated by the product of the norm bound of
the perturbation and the sensitivity of the solution to the perturbations. Thus the problem of estimating the error in the solution can be considered as a problem of computing the magnitude of the sensitivity of the solution with respect to all of these parameters. The parameter space has a large dimension. For the Sylvester equation (1), the parameter space has dimension $n^2 + m^2 + mn$. Computing the sensitivity $\frac{dx}{dp}$ of each solution component with respect to so many parameters is computationally intractable for a large system. Fortunately we do not need to exactly compute the sensitivities with respect to each parameter. $\|\frac{dx}{dp}\|$, the magnitude of the sensitivity over all the parameters can give us enough information about the condition of the system. The idea of our error estimation is based on this observation.

To estimate the magnitude of the sensitivity, in this section we present an algorithm for computing the sensitivity of a scalar function $g(x)$ (called a derived function) with respect to potentially many parameters. The concept of derived function comes from the literature of sensitivity analysis of differential equations. It is a user-defined function. Different functions $g$ yield different sensitivities. The function $g$ provides flexibility for different types of measurements of the error. Although the computational cost for computing the sensitivity of the solution $X$ with respect to so many parameters is high, the sensitivity of the scalar function $g(X)$ with respect to those parameters can be computed very efficiently by the adjoint method. In this paper we will present one possible choice of $g$. Other choices of $g$ are possible; for example some other choices for linear systems are discussed in [5]. The derived function can be naturally combined with the the SSSE method to estimate the magnitude of the sensitivity $\|\frac{dx}{dp}\|$. This magnitude is defined as the condition of the system [18] and is used to derive an error estimate.

We derive the method for the sensitivity analysis of a scalar derived function with respect to many parameters from a general nonlinear matrix equation with one parameter $p$.

(16) \[ F(X, p) = 0, \]

where $X \in \mathbb{R}^{m \times n}$, $F : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}$, and $p$ is a parameter which represents the perturbation from one direction. (16) can take the form of the Sylvester equation, Lyapunov equation, Algebraic Riccati equation, or other matrix equations. The existence and uniqueness of the solution to (16) is assumed. Suppose the scalar derived function $g(X)$ is given. The basic sensitivity problem is to compute $\frac{dg}{dp}$ for the parameter $p$. Applying the chain rule, we obtain

(17) \[ \frac{dg}{dp} = \sum_{i,j} \frac{dg}{dx_{ij}} \frac{dx_{ij}}{dp}. \]

$\frac{dx_{ij}}{dp}$ can be formulated as a matrix, denoted by $\frac{dx}{dX}$. For example, if $g(X) = \|X\|$, we have $\frac{dx}{dX} = \frac{X}{\|X\|}$. Using the inner product of two matrices, which is defined by

(18) \[ \langle X, Y \rangle = \text{tr}(X^T Y) = \sum_{i,j} x_{ij} y_{ij}, \]

we can express the sensitivity of $g(X)$ with respect to $p$ as

(19) \[ \frac{dg}{dp} = \langle \frac{dg}{dX}, X_p \rangle, \]
where $X_p = \frac{dX}{dp}$. On the other hand, differentiating (16), we have

$$F_p + \frac{dF}{dX}X_p = 0,$$

where $F_p = \frac{dF}{dp}$, and $\frac{dF}{dX}$ is the directional derivative operator defined as

$$\frac{dF}{dX}Y = \lim_{h \to 0} \frac{F(X + hY) - F(X)}{h}.$$

Taking the inner product of both sides of (20) with a matrix $\Lambda$ yields

$$< \Lambda, F_p + \frac{dF}{dX}X_p > = 0.$$

Thus

$$< \Lambda, \frac{dF}{dX}X_p > = - < \Lambda, F_p > .$$

The idea of the adjoint sensitivity method is to choose $\Lambda$ so that

$$< \Lambda, \frac{dF}{dX}X_p > = < \frac{dg}{dX}, X_p > .$$

Substituting the above and (22) into (19), we obtain another formula for the sensitivity

$$\frac{dg}{dp} = - < \Lambda, F_p > .$$

To solve for $\Lambda$ from (23), let the adjoint operator $(\frac{dF}{dX})^*$ of $\frac{dF}{dX}$ be defined by

$$< Z, \frac{dF}{dX}Y > = < (\frac{dF}{dX})^*Z, Y > .$$

With the adjoint operator, (23) becomes

$$< (\frac{dF}{dX})^*\Lambda, X_p > = < \frac{dg}{dX}, X_p > .$$

Now solve the adjoint equation

$$(\frac{dF}{dX})^*\Lambda = \frac{dg}{dX}$$

for the matrix $\Lambda$. The sensitivity is then given by (24).

(27) can take different forms for different choices of $F$. For the Sylvester equation, we have $F(X) = AX - XB - C$. The differential operator is given by

$$\frac{dF}{dX}Y = AY - YB.$$

Thus

$$< Z, \frac{dF}{dX}Y >= < Z, AY - YB >= < A^T Z - Z B^T, Y > .$$
The adjoint operator is given by
\[
\left( \frac{dF}{dX} \right)^* Z = A^T Z - Z B^T,
\]
and the adjoint equation is
\[
A^T \Lambda - \Lambda B^T = \frac{dg}{dX}.
\]

The sensitivity \( \frac{dg}{dp} \) is then given by \( \langle \Lambda, A_p X - X B_p - C_p \rangle \). Here \( p \) denotes a parameter, and \( A_p, B_p \) and \( C_p \) denote the corresponding derivatives of \( A, B \) and \( C \) with respect to \( p \).

3 Error Estimation

The perturbations to the data are considered as parameters of the system. The error estimation is based directly on the magnitude of the sensitivity to these perturbations. Note that we do not need to compute the sensitivity very accurately. The objective is to estimate the magnitude of the error to within a factor of 10.

The perturbations arise from data error or from round-off error. Either relative or absolute error may be of the most concern, depending on the situation. Relative error is usually of the greatest concern in numerical solution. Thus in the following discussion we will focus on the relative error. The absolute error estimate is easy to obtain in a similar manner.

First let us consider the change in a scalar function \( g \) with respect to the perturbations. The first order approximation yields
\[
|g(X(0)) - g(X(p))| \approx \sum_p |\frac{dg}{dp}| |p|
\]
when the perturbations \( |p| \) are all small. Suppose \( |p| < \epsilon \).

\[
|g(X(0)) - g(X(p))| \leq \sum_p |\frac{dg}{dp}| \epsilon,
\]
where \( \frac{dg}{dp} \) can be computed by (24). Note that \( \Lambda \) computed from (27) is independent of \( p \), and

\[
\sum_p |\frac{dg}{dp}| = \sum_p |\langle \Lambda, F_p \rangle|.
\]

(29) and (30) give the error estimation for the scalar function \( g \).

Our objective is to estimate the magnitude of the error in the solution of a matrix equation. The error is a multi-dimensional function of the solution. However, to use a multi-dimensional function to estimate the error would be too expensive. Thus a strategy for making use of a scalar function to estimate the error of the vector (matrix) is needed. The small sample statistical estimation (SSSE) method provides such a tool [14].

Consider \( \Delta X \) as a vector. One way to estimate \( \| \Delta X \| \) is to take its inner product with a vector uniformly randomly selected from the unit ball \( S_{k-1} \), where \( k \) is the dimension of \( \Delta X \). For
the Sylvester equation, $k = nm$. Let $R$ be the random matrix uniformly selected from the unit sphere $S_{k-1}$. According to [14], we have

$$E(| < R, \Delta X > |) = E_k \| \Delta X \|,$$

where $E$ denotes the expectation, and $E_1 = 1$, $E_2 = \frac{2}{\pi}$, and for $k > 2$,

$$E_k = \frac{1 \cdot 3 \cdot 5 \cdots (k-2)}{2 \cdot 4 \cdot 6 \cdots (k-1)} \quad \text{for k odd,}$$

$$E_k = \frac{2}{\pi} \cdot \frac{2 \cdot 4 \cdot 6 \cdots (k-2)}{1 \cdot 3 \cdot 5 \cdots (k-1)} \quad \text{for k even.}$$

$E_k$ can be estimated by $\sqrt{\frac{2}{\pi(k-1)}}$. $\| \Delta X \|_F$ can be estimated, using one random matrix, by

$$\| \Delta X \| \approx \frac{| < R, \Delta X > |}{E_k}. \tag{31}$$

Thus we can define the scalar function $g(X) = < R, X >$. Then $\Delta g = < R, \Delta X >$. Since $g$ is a scalar function, the adjoint sensitivity method (29) and (30) can be applied.

The corresponding probability for accuracy to within a factor of $w$ of the error estimation using one random matrix is given by

$$Pr \left( \frac{1}{w} \| \Delta X \|_F \leq \frac{| < R, \Delta X > |}{E_k} \leq w \| \Delta X \|_F \right) \approx 1 - \frac{2}{\pi w}.$$

We require an error estimate which is accurate to within a factor of 10. We call such an error estimate a ‘good estimate’. For one random matrix, the probability of a good estimate is about 93%. If we need a higher probability or a more accurate estimate, more random matrices can be used. Usually we use 2 or 3 random matrices. Let $R_i$ be orthogonal random matrices uniformly selected from $S_{k-1}$. $< R_i, R_j >= 0$ if $i \neq j$. Define $v_i = | < R_i, \Delta X > |$. The error estimates are given by

$$e_2 = \frac{E_2 \sqrt{v_1^2 + v_2^2}}{E_k}$$

with 2 random matrices and

$$e_3 = \frac{E_3 \sqrt{v_1^2 + v_2^2 + v_3^2}}{E_k}$$

with 3 random matrices. The corresponding probabilities are

$$Pr \left( \frac{1}{w} \| \Delta X \|_F \leq e_2 \leq w \| \Delta X \|_F \right) \approx 1 - \frac{\pi}{4w^2}$$

and

$$Pr \left( \frac{1}{w} \| \Delta X \|_F \leq e_3 \leq w \| \Delta X \|_F \right) \approx 1 - \frac{32}{3\pi^2 w^3}.$$

Letting $w = 10$, using 2 random matrices gives a probability of 99.21%, while using 3 random matrices gives a probability of 99.89%. $v = | < R, \Delta X > |$ is estimated by solving the adjoint equation
(27) and applying (29) and (30). Let $\Lambda$ be the solution of (27) and $u = \sum_p | < \Lambda, \frac{dF}{dp} > |$. Then $|v| \leq |u| \epsilon$. Note that the computational cost is not high because we can use the matrix decomposition which was formed when solving the original system. For example, the Schur decomposition used in the solution of the Sylvester equation can be used for solving the corresponding adjoint equation. Below we describe the algorithm using 2 random matrices.

**Error estimate algorithm with 2 random matrices** Suppose we have solved the original matrix equation (1) and the corresponding matrix decomposition is available.

Step 1: Uniformly and randomly choose two orthogonal matrices $R_1, R_2$ from the unit sphere $S_{k-1}$. Solve (27) for the corresponding $\Lambda_1, \Lambda_2$.

Step 2: Compute

$$u_i = \sum_p | < \Lambda_i, \frac{dF}{dp} > |.$$ 

Then the absolute error estimate for $\|\Delta X\|$ is given by

$$e = \frac{E_2 \sqrt{(u_1^2 + u_2^2)}}{E_k} \epsilon.$$ 

The relative error estimate for $\frac{\|\Delta X\|}{\|X\|}$ is given by $\frac{e}{\|X\|}$.

By choosing the scalar function in an appropriate manner, we can generalize the above algorithm to produce a subspace error estimate. Suppose we want to estimate the error of the submatrix $Y = PXQ$, where $P \in R^{p \times n}$, $Q \in R^{m \times q}$ with $p \ll n$, $q \ll m$. We first select a random matrix $R \in R^{p \times q}$ uniformly from the unit sphere $S_{k-1}$, where $k = p \times q$. Then, defining $g(X) = < R, PXQ >$, the adjoint method and small sample statistical method can be applied as before. Note that in this case $\frac{dg}{dX} = P^T R Q^T$. When $p = q = 1$, this subspace error estimate reduces to an error estimate for the individual component.

**Sylvester equation** For the Sylvester equation, the sensitivity with respect to perturbations in $C$ is easy to compute. It is bounded by $< |\Lambda|, |C| >$. For perturbations in $A$, considering the relative error, let $\Delta a_{ij} = \epsilon_{ij} a_{ij}$. Then

$$\frac{\partial F}{\partial \epsilon_{ij}} = A_{ij} X = a_{ji} \epsilon_{ij} X,$$

and

$$< \Lambda, \frac{\partial F}{\partial \epsilon_{ij}} >= \sum_k |a_{ji} \lambda_{ik} x_{jk}|.$$ 

The sensitivity with respect to perturbations in $B$ can be obtained similarly. Thus

$$\sum_p | \frac{dg}{dp} | = < |\Lambda|, |W| > + \sum_{ij} | \sum_k a_{ji} \lambda_{ik} x_{jk} | + \sum_{ij} | \sum_k b_{ij} \lambda_{kj} x_{ki} |$$

$$\leq < |\Lambda|, |A||X| + |X||B| + |C| > .$$

10
Remark 1: We should comment on the difference between this method and the original small sample statistical estimation for matrix equations as in [14]. The method in [14] generates the random vectors from the perturbation space, while our method generates the random vectors from the solution space. For example, for a linear system $Ax = b$, [14] generates random vectors $Z_A$, $Z_b$ with unit length (dimension $n^2 + n$), then multiplies them by a small number $\delta$ and solves the system $(A + \delta Z_A)(x + \Delta x) = b + \delta z_b$. Thus $\frac{||\Delta x||}{\delta}$ is the condition number estimated. The choice of $\delta$ is critical and there is no good rule for it. Our method generates random vectors $z$ with unit length (dimension $n$) from the solution space, constructs a derived function $g(x) = z^T x$, and then solves the adjoint equation. Thus we avoid the critical choice of $\delta$, and we generate lower dimensional random vectors.

Remark 2: Our method for condition and error estimation leads to an a posteriori error estimate. It must solve the matrix equation first, then it can estimate the error in the solution. In contrast, the condition number (9) leads to an a priori error estimate. We note that (11) also leads to an a posteriori error estimate. Our method estimates the error as opposed to bounding it. Although in practice we seldom find that this method leads to an estimate that is smaller than the actual error, we can only claim that our estimate has a very high probability to be no less than $1/10$ times the Frobenius norm of the actual error.

4 Numerical Experiments

The numerical experiments were performed in Matlab on a Linux computer. We chopped the data so that the round-off error is $10^{-8}$, in order to avoid any possibility of differences in the conclusions that could be caused by different machine precisions.

The numerical experiments are based on randomly generated data based on a uniform distribution. We compare our error estimate with the condition error estimate (13), (14), and, when applicable, with the method (7) proposed in [12] for the stable Lyapunov equation. We first generate the random matrices $A$, $B$ and $X$. Then $C$ is determined by $C = AX - XB$. We chop the data of $A$, $B$ and $C$ to a relative error of $10^{-8}$. Then we solve $A\tilde{X} - \tilde{X}B = C$. We compare the error estimates and the actual relative error $\frac{||X - \tilde{X}||}{||X||}$. For the error estimates (13) and (14), $\|P^{-1}\|_2$ is computed accurately without approximation.

When we consider the error estimate, every possible perturbation is taken into account. This usually yields an overestimate of the relative error. In all of our experiments, seldom is an underestimate generated. Thus we only compare the overestimate ratio $\frac{\text{estimate}}{\text{actual relative error}}$.

1. Dense Sylvester Equation 1000 random dense matrices $A$, $B$ and $X$ of dimension 10 were generated. Here we do not assume any structure such as symmetric or positive definite. In Figure 1 we compare the overestimate ratio of different estimates. The mean value of the overestimate ratio of our method is 30.73, while the mean value of the overestimate ratio of (13) is 43.12 and the mean value of the overestimate ratio of (14) is 98.28. In the extreme case, (13) gives an overestimate ratio as high as 4,486, and (14) gives an overestimate ratio as high as 10,200, while the highest overestimate ratio given by our method is 1,308. We can see that generally our method gives a sharper estimate. As we have mentioned, the computational cost for our method is also lower.
2. Diagonal Matrices If the Sylvester equation has some special structure, our method can give even better results. In this experiment we randomly generated diagonal matrices $A$ and $B$. Because of the special structure, the actual relative error is small. The overestimate ratio is plotted in Figure 2. The mean value of the overestimate ratio of our method is 11.58, while the mean value of the overestimate ratio of (13) is 17.96, and the mean value of the overestimate ratio of (14) is 60.37. In the extreme case, (14) gives an overestimate ratio as high as 2,467, and (13) gives an overestimate ratio as high as 722, while the highest overestimate ratio given by our method is 19.44.

3. Stable Lyapunov Equation For the stable Lyapunov equation, [12] gives a simple error estimate (7). In this experiment, we randomly generated $1,000$ stable matrices $A$ of dimension $10$. The overestimate ratio of (7) is compared with our method and the condition estimate method (13) (since (13) is sharper than (14)). The result is plotted in Figure 3. The mean value of the overestimate ratio of our method is 22.90, while the mean value of the overestimate ratio of (13) is 12.31, and the mean value of the overestimate ratio of (7) is 53.00. We can see that our method gives a sharper estimate than (7) but we note that the computational cost of (7) is lower unless we use only one random matrix to generate our error estimate (but then it is less reliable). We note again that the computational cost of our method is much lower than that of (13).
Figure 3: Overestimate ratio of our method (left), condition estimate (13) (middle) and the method (7) of [12] (right) for 1,000 Lyapunov equations with randomly generated stable matrices $A$, and dense matrices $X$ of dimension 10.

5 Conclusion

In this paper we have proposed a new method of error estimation for matrix equations, based on the adjoint sensitivity method and small-sample statistical theory. Our method has low computational cost, and probability of 99% for the accuracy of the error estimate to be within a factor of 10. Numerical experiments show that our method yields sharper estimates for randomly generated dense Sylvester equations than the standard error estimate. For stable Lyapunov equations, our method’s computational cost is higher than that of the error estimate proposed in [12], but the estimate is sharper. Our method can be generalized in a straightforward manner to other matrix equations, for example the discrete-time Lyapunov equation and the algebraic Riccati equation.

Acknowledgments The authors would like to thank Karl Åström and Charles Kenney for their insightful comments in discussions related to this work.

References


