CREATIVE MATH.
12 (2003), 15 - 20

On approximating the inverse of a matrix

Ion Păvăloiu

Abstract. In this note we deal with two problems: the first regards the efficiency in approximating the inverse of a matrix by the Shulz-type methods, and the second is the problem of evaluating the errors in the approximation of the inverses of the perturbed matrices.

1. Introduction

In this note we deal with two problems: the first regards the efficiency in approximating the inverse of a matrix by the Shulz-type methods, and the second is the problem of evaluating the errors in the approximation of the inverses of the perturbed matrices.

As it is well known, given a nonsingular matrix $A \in \mathbb{R}^{m\times m}$ and a matrix $D_0 \in \mathbb{R}^{m\times m}$ such that

$$\|I - AD_0\| \leq q < 1 \tag{1}$$

with $q \in \mathbb{R}$ and I the m-th order unit matrix, then, for $k \in \mathbb{N}, k \geq 2$ fixed, the sequence of matrices $(D_n)_{n\geq 0}$ given by

$$\begin{aligned} F_n &= I - AD_n \\ D_{n+1} &= D_n\left(I + F_n + F_n^2 + \cdots + F_n^{k-1}\right), \quad n = 0, 1, ... \end{aligned} \tag{2}$$

is convergent and $\lim_{n\to\infty} D_n = A^{-1}$. Moreover, $(F_n)_{n\geq 0}$ verifies

$$F_{n+1} = F_n^k, \quad n = 0, 1, ... \tag{3}$$

The methods of type (2) represent generalizations of the well known Shulz method. Relation (3) shows that the convergence order of sequence $(D_n)_{n\geq 0}$ is k, $k \geq 2$.

We introduce the notion of efficiency index of method (2). We notice that at each iteration step, the number of the matrix sums required is equal to the number of matrix products which appear in (2). Moreover, for computing the sum $I+F_n+\cdots+F_n^k$ we may use a method similar to the Horner scheme, i.e.

$$I + F_n + F_n^2 + \cdots + F_n^{k-1} = \left\{\left\{\left[\left(F_n + I\right)F_n + I\right]F_n + I\right\} + \cdots\right\}. \tag{4}$$

Received: 28.02.2003; In revised form: 30.10.2003
Key words and phrases. *Inverse of a matrix, perturbed matrix, efficiency in approximating the inverse of a matrix.*

In this way the matrix sums required reduce to sums in which one term is the identity matrix. This remark is also valid for the term $F_n = I - AD_n$. The operation consisting of one matrix product and one matrix sum (regardless of their order) we call it computing unit.

Definition 1. *The efficiency index of method (2) is given by*

$$E_k = k^{1/s}, \tag{5}$$

where $s \in \mathbb{N}$ represents the number of computing units required at each iteration step of method (2).

This definition is given by analogy to the efficiency index introduced by A.M. Ostrowski in [2]. The definition may also be motivated by the following reasoning.

From (3) it follows

$$\|F_{n+1}\| \leq \|F_n\|^k, \quad n = 0, 1, ... \tag{6}$$

whence

$$\|F_{n+1}\| \leq \|F_0\|^{k^{n+1}}, \quad n = 0, 1, ... \tag{7}$$

The above inequalities lead to the following error bounds:

$$\left\|A^{-1} - D_n\right\| \leq \left\|A^{-1}\right\| \|F_0\|^{k^n}, \quad n = 0, 1, ... \tag{8}$$

Consider now two methods of type (2), having the convergence orders k_1 and k_2 respectively. Assume that, for achieving the same precision, these methods require n_1 respectively n_2 iteration steps. Then (8) implies

$$k_1^{n_1} = k_2^{n_2}. \tag{9}$$

The total number of computing units is $n_1 s_1$ in the first case and $n_2 s_2$ in the second case.

It is clear now that the method with convergence order k_1 is more efficient than the other if

$$n_1 s_1 < n_2 s_2. \tag{10}$$

Relations (9) and (10) lead us to

$$k_1^{1/s_1} > k_2^{1/s_2}. \tag{11}$$

Taking into account Definition 1, it follows that among the methods of type (2) for different values of k, the most efficient is given by the one with high efficiency index.

We shall determine in the following section the optimal method, i.e., having the high efficiency index, when $k \in \mathbb{N}$, $k \geq 2$.

2. Optimal efficiency index

Assume that we use (4) at each iteration step in (2). It can be easily seen that for the sum in (4) there are needed $k-2$ matrix products. Relation (2) shows that 2 more matrix products are required at each iteration step, so in total we need k matrix products.

Taking into account (5), it follows that the *efficiency index* of method (2) is given by

$$\overline{E}_k = k^{1/k}. \tag{12}$$

Considering the function $f : (0, +\infty) \to \mathbb{R}$, $f(x) = x^{\frac{1}{x}}$, it can be easily seen that this function attains a maximum value at $x = e$. Since f is increasing on $(0, e)$ and decreasing on $(e, +\infty)$, it follows that $\overline{E}_k$ is the largest for $k = 3$.

We have proved the following result.

Theorem 2. *Among the methods (2) for $k = \mathbb{N}$, $k \geq 2$, the method with highest efficiency index is given by:*

$$\begin{cases} F_n = I - AD_n \\ D_{n+1} = D_n \left(I + F_n + F_n^2 \right), \ n = 0, 1, ... \end{cases} \tag{13}$$

with D_0 verifying $\|I - AD_0\| \leq q < 1$.

By (4), the above method may be written as

$$\begin{aligned} F_n &= I - AD_n \\ D_{n+1} &= D_n \left[(F_n + I) F_n + I \right], \quad n = 0, 1, ... \end{aligned} \tag{14}$$

In this case, (7) becomes

$$\|F_{n+1}\| \leq \|F_0\|^{3^{n+1}}, \quad n = 0, 1, ... \tag{15}$$

and for the error bounds one has

$$\left\|A^{-1} - D_n\right\| \leq \left\|A^{-1}\right\| \|F_n\| \leq \left\|A^{-1}\right\| \|F_0\|^{3^n}, \ n = 0, 1, ... \tag{16}$$

It can be easily seen that under (1), one has the inequality

$$\left\|A^{-1}\right\| \leq \frac{\|D_0\|}{1 - \|F_0\|} \tag{17}$$

whence

$$\left\|A^{-1} - D_n\right\| \leq \|D_0\| \frac{\|F_0\|^{3^n}}{1 - \|F_0\|}, n = 0, 1, ... \tag{18}$$

Analogously, for any method of type (2) one may deduce the evaluation

$$\left\|A^{-1}-D_n\right\| \leq \|D_0\| \frac{\|F_0\|^{k^n}}{1-\|F_0\|}, \quad n=0,1,\ldots . \tag{19}$$

3. Error bounds in case of perturbed matrices

In practice, the elements of the matrix A are usually obtained as results of certain experiments, measurements, approximations etc. Therefore their values are altered by errors. Consequently we replace A by the approximation $\widetilde{A}$. For a rigorous interpretation of the results, it is necessary to know an error bound $\varepsilon > 0$ for which

$$\left\|A-\widetilde{A}\right\| \leq \varepsilon. \tag{20}$$

Instead of sequence $(D_n)_{n\geq 0}$ we consider $\left(\widetilde{D}_n\right)_{n\geq 0}$, generated by

$$\begin{aligned} \widetilde{F}_n &= I-\widetilde{A}\widetilde{D}_n; \\ \widetilde{D}_{n+1} &= \widetilde{D}_n\left(I+\widetilde{F}_n+\widetilde{F}_n^2+\cdots+\widetilde{F}_n^{k-1}\right), \quad n=0,1,\ldots . \end{aligned} \tag{21}$$

We assume that the matrices $\widetilde{A}$ and $\widetilde{D}_0$ above obey

$$\left\|I-\widetilde{A}\widetilde{D}_0\right\| \leq \overline{q} < 1. \tag{22}$$

It follows that $\widetilde{A}$ is invertible: $\exists \widetilde{A}^{-1}$ and by (18) we get

$$\left\|\widetilde{A}^{-1}-\widetilde{D}_n\right\| \leq \left\|\widetilde{D}_0\right\| \frac{\left\|\widetilde{F}_0\right\|^{k^n}}{1-\left\|\widetilde{F}_0\right\|}, \quad n=0,1,\ldots . \tag{23}$$

We are interested in conditions which ensure that $\widetilde{A}$ is nonsingular. We consider the identity

$$I-\widetilde{A}^{-1}A = \widetilde{A}^{-1}\left(\widetilde{A}-A\right)$$

which implies

$$\left\|I-\widetilde{A}^{-1}A\right\| \leq \left\|\widetilde{A}^{-1}\right\|\varepsilon,$$

whence, by (17) we get

$$\left\|I-\widetilde{A}^{-1}A\right\| \leq \frac{\varepsilon\left\|\widetilde{D}_0\right\|}{1-\left\|\widetilde{F}_0\right\|} \tag{24}$$

This relation shows that for the existence of the inverse for $\widetilde{A}^{-1}A$ it suffices that

$$r = \frac{\varepsilon \left\| \widetilde{D}_0 \right\|}{1 - \left\| \widetilde{F}_0 \right\|} < 1 \tag{25}$$

whence for ε we get the condition

$$\varepsilon < \frac{1 - \|F_0\|}{\left\| \widetilde{D}_0 \right\|}. \tag{26}$$

Further,

$$A^{-1} = \left(\widetilde{A}^{-1}A \right)^{-1} \widetilde{A}^{-1}$$

whence

$$\left\| A^{-1} \right\| \leq \left\| \widetilde{A}^{-1} \right\| \left\| \widetilde{A}^{-1}A \right\| \leq \frac{\left\| \widetilde{D}_0 \right\|}{1 - \left\| \widetilde{F}_0 \right\| - \varepsilon \left\| \widetilde{D}_0 \right\|}$$

and (26) attracts $1 - \left\| \widetilde{F}_0 \right\| - \varepsilon \left\| \widetilde{D}_0 \right\| > 0$.

The following inequality can be easily proved

$$\left\| A^{-1} - \widetilde{A}^{-1} \right\| \leq \frac{\left\| \widetilde{D}_0 \right\|^2 \varepsilon}{\left(1 - \left\| \widetilde{F}_0 \right\| \right) \left(1 - \left\| \widetilde{F}_0 \right\| - \varepsilon \left\| \widetilde{D}_0 \right\| \right)} \tag{27}$$

which, together with (23) leads to

$$\left\| A^{-1} - \widetilde{D}_n \right\| \leq \frac{\left\| \widetilde{D}_0 \right\|}{1 - \left\| \widetilde{F}_0 \right\|} \left[\frac{\left\| \widetilde{D}_0 \right\| \varepsilon}{1 - \left\| \widetilde{F}_0 \right\| - \varepsilon \left\| \widetilde{D}_0 \right\|} + \|F_0\|^{t^n} \right], \ n = 0, 1, ... \tag{28}$$

This inequality provides a priori evaluations for the error. If we want to stop the iterations at a certain step $\overline{n}$ such that $\left\| \widetilde{F}_{\overline{n}} \right\| \leq \varepsilon_1$, $\varepsilon_1 > 0$ given, then by (7) and (17) it follows

$$\left\| \widetilde{A}^{-1} - \widetilde{D_{\overline{n}}} \right\| \leq \frac{\left\| \widetilde{D}_0 \right\|}{1 - \left\| \widetilde{F}_0 \right\|} \varepsilon_1,$$

which, together with (27) lead to

$$\left\| A^{-1} - \widetilde{D}_{\overline{n}} \right\| \leq \frac{\left\| \widetilde{D}_0 \right\|}{1 - \left\| \widetilde{F}_0 \right\|} \left[\varepsilon_1 + \frac{\varepsilon \left\| \widetilde{D}_0 \right\|}{1 - \left\| \widetilde{F}_0 \right\| - \varepsilon \left\| \widetilde{D}_0 \right\|} \right]$$

which is an a posteriori error bound.

References

[1] Herzberger J., *Explizite Shulz Verfahren höherer ordrnung zur approximation der reversen matrix*, Z. Angew Math. und Mech. 1988, Bd. 68, No. 5, pp. 494-496
[2] Ostrowski M.A., Solution of equations in euclidian and Banach spaces, Academic Press. New York and London (1975)
[3] Stickel E., *On a class of high order methods for investing matrices*, Z. Angew Math. und Mech. 1987, Bd. 67, No. 7, pp. 334-336

"T. Popoviciu" Institute of Numerical Analysis
Str. Fântânele, nr.57, Bloc B7, sc.II, etaj 5, ap. 67-68
Cluj-Napoca, Romania
E-mail address: pavaloiu@ictp.acad.ro