G. A. Sitton

Direct Technique for Improving a Matrix Inverse

Abstract: A method is shown that transforms the problem of inverting an ill-conditioned matrix to one of inverting a diagonally dominant matrix. An error analysis is outlined and the method is compared in theory and in result with the commonly used iterative methods. This direct method is demonstrated to be the limiting case of an nth-order iterative procedure as n approaches infinity. Examples are given that show the advantages of the direct method even under adverse conditions. The unreliability of the convergence of the iterative technique due to computational errors is also discussed.

Introduction

The problem of obtaining accurate inverses of large ill-conditioned matrices arises in many numerical procedures[1,2]. Application areas include structural analysis, electrical network analysis, and the solution of systems of differential equations with variable coefficients. To realistically evaluate any method, a detailed analysis of computational errors due to finite significance arithmetic is necessary[3,4]. Consideration is given here to *improving* a given approximate inverse matrix that has previously been obtained by arbitrary means. No discussion, however, is directed to the related problem of solving systems of linear equations.

Given a matrix $\mathbf{A}$ to be inverted and an arbitrary matrix $\mathbf{B}$,

$$\mathbf{A}^{-1} = \mathbf{B}(\mathbf{AB})^{-1} \tag{1}$$

if $\mathbf{A}$ and $\mathbf{B}$ are nonsingular. By letting $\mathbf{B}$ be an estimate of $\mathbf{A}^{-1}$ such that

$$\mathbf{AB} = \mathbf{C} = \mathbf{I} + \boldsymbol{\epsilon}, \tag{2}$$

we show that, under certain conditions, $\langle \mathbf{A}^{-1} \rangle$ is a better approximation to $\mathbf{A}^{-1}$ than is $\mathbf{B}$; here

$$\langle \mathbf{A}^{-1} \rangle = \langle \mathbf{B} \langle \langle \mathbf{C} \rangle^{-1} \rangle \rangle. \tag{3}$$

The notation $\langle\rangle$ is used to indicate a computed value subject to roundoff error, and $\mathbf{I}$ is the unit matrix.

Notice that we could have chosen, instead of (1),

$$\mathbf{A}^{-1} = (\mathbf{BA})^{-1}\mathbf{B} \tag{1'}$$

as our basic premise, but the results will be the same using (1) or (1′) in any analysis. However, the error matrix, $\boldsymbol{\epsilon}$, will be different for the pre- or post-multiplicative correction in the direct method. The better of the two can be decided only by carrying out the actual computations.

Direct method

It is more convenient to work with scalar norms than with matrices to measure error. To establish an improvement criterion, a norm function having the following properties is used:

$$\|\mathbf{X} + \mathbf{Y}\| \leq \|\mathbf{X}\| + \|\mathbf{Y}\|, \text{ and} \tag{4}$$

$$\|c\mathbf{X}\| = |c|\|\mathbf{X}\| \geq 0, \tag{5}$$

where c is a scalar constant; and from (4) and (5),

$$\|\mathbf{X} - \mathbf{Y}\| \leq \|\mathbf{X}\| + \|\mathbf{Y}\|, \text{ and} \tag{6}$$

$$\|\mathbf{XY}\| \leq \|\mathbf{X}\|\,\|\mathbf{Y}\|. \tag{7}$$

There are a number of norms that satisfy these relations; the ones more difficult to compute and use analytically tend to give smaller values as a measure of error[5].

It is also necessary to use the fact that if

$$\|\boldsymbol{\epsilon}\| < 1, \tag{8}$$

$(\mathbf{I} + \boldsymbol{\epsilon})^{-1}$ exists and is given by

$$(\mathbf{I} + \boldsymbol{\epsilon})^{-1} = \mathbf{I} - \boldsymbol{\epsilon} + \boldsymbol{\epsilon}^2 - \boldsymbol{\epsilon}^3 + \cdots. \tag{9}$$

To show the error in approximation (3), all computed quantities must be expressed in terms of actual values and error terms. Regarding $\boldsymbol{\epsilon}$ as a measure of the error in a computed inverse matrix, by analogy with Eq. (2) we can write

$$\mathbf{A}\langle \mathbf{A}^{-1} \rangle = \mathbf{I} + \boldsymbol{\sigma} \quad \text{and} \tag{10}$$

$$\langle \mathbf{C} \rangle \langle \langle \mathbf{C} \rangle^{-1} \rangle = \mathbf{I} + \boldsymbol{\delta}. \tag{11}$$

The matrices ϵ, σ, and δ in Eqs. (2), (10), and (11) are residual matrices. Note also from Eq. (2) that

$$\epsilon = (\mathbf{B} - \mathbf{A}^{-1}). \tag{12}$$

Thus ϵ, σ, and δ may be considered to be relative error matrices. Absolute errors are not usually of interest because they are easily changed by a multiplicative constant; only relative errors are considered in this discussion.

We must also express the error due to a computed matrix multiplication. This is done by defining

$$\langle \mathbf{C} \rangle = (\mathbf{I} + \mu_1)\mathbf{C}, \tag{13}$$

and if $\langle \mathbf{C} \rangle$ is nonsingular,

$$\langle \mathbf{B}\langle \langle \mathbf{C} \rangle^{-1} \rangle \rangle = \mathbf{B}\langle \langle \mathbf{C} \rangle^{-1} \rangle (\mathbf{I} + \mu_2). \tag{14}$$

If we assume that $(\mathbf{I} + \mu_1)^{-1}$ exists, then by substituting Eqs. (2), (10), and (11) in Eq. (3), using the results of Eqs. (13) and (14), and regrouping terms and simplifying, we obtain

$$\sigma = (\mathbf{I} + \mu_1)^{-1} (\mu_2 - \mu_1 + \delta + \delta\mu_2). \tag{15}$$

If $\mu_1 = \mu_2 = 0$, Eq. (15) shows that $\sigma = \delta$. Therefore the error in the final approximation to $\mathbf{A}^{-1}$ is equal to the error introduced in inverting $\langle \mathbf{C} \rangle$, except for the multiplication errors. In these multiplications μ_1 and μ_2 can be kept small by using double-precision arithmetic while accumulating inner products[5].

The norm relation corresponding to (15) is found by applying Eqs. (4) through (7), which yields

$$\|\sigma\| \leq \|(\mathbf{I} + \mu_1)^{-1}\| (\|\mu_2\| + \|\mu_1\| + \|\delta\| + \|\delta\| \|\mu_2\|). \tag{16}$$

Similarly, after expanding as in Eq. (9), we obtain

$$\|(\mathbf{I} + \mu_1)^{-1}\| \leq \|\mathbf{I}\| + \|\mu_1\| + \|\mu_1\|^2 + \cdots, \tag{17}$$

or

$$\|(\mathbf{I} + \mu_1)^{-1}\| \leq \|\mathbf{I}\| + \|\mu_1\|(1 - \|\mu_1\|)^{-1}. \tag{18}$$

Thus the final norm equation is

$$\|\sigma\| \leq [\|\mathbf{I}\| + \|\mu_1\|(1 - \|\mu_1\|)^{-1}] (\|\mu_1\| + \|\mu_2\| + \|\delta\| + \|\delta\| \|\mu_2\|). \tag{19}$$

It remains only to compute or estimate upper bounds on $\|\mu_1\|$, $\|\mu_2\|$, and $\|\delta\|$. It should be clear that $\|\delta\|$ is a function of $\|\epsilon\|$ as well as of $\|\mu_1\|$. Error analyses have been done by Turing[6] for Gauss-Jordan reduction and by Wilkinson[7] for Gaussian elimination. Wilkinson[5] has also shown an error relation for matrix multiplication. A simpler solution might be just to try the method to see if an improvement can be made, i.e., if $\|\sigma\| < \|\epsilon\|$.

If inequality (8) is satisfied and if the multiplicative errors $\|\mu_1\|$ and $\|\mu_2\|$ are negligible, the logic involved in any pivotal inversion scheme is minimal, since only diagonal pivoting is required[8]. The inversion process, called Gauss-Jordan reduction, has low error also because all of the pivot elements are approximately equal to one [see Eq. (2)]. This method for improving an inverse is basically noniterative, but due to rounding errors is not exact, i.e., one could apply Eq. (3) more than once. Empirical evidence, however, has not shown any improvement after a second correction pass was made.

As an added point of interest, it may be possible to improve an associated determinant, which can usually be computed along with the inverse matrix at little expense. This can be shown by using the determinant relation

$$|\mathbf{XY}| = |\mathbf{X}|\,|\mathbf{Y}| \tag{20}$$

in Eq. (1). The result is

$$|\mathbf{A}| = |\mathbf{B}^{-1}|\,|\mathbf{AB}|. \tag{21}$$

If $\mathbf{B} \approx \mathbf{A}^{-1}$, then $\mathbf{B}^{-1} \approx \mathbf{A}$ and $|\mathbf{B}^{-1}| \approx |\mathbf{A}|$. Thus Eq. (21) may be used with Eq. (3) to improve a determinant along with an inverse matrix. The error analysis procedure for (21) would be similar to that for (3).

Classical method

The most commonly described matrix improvement technique[5, 9–12] is iterative:

$$\mathbf{B}_{i+1} = \mathbf{B}_i(\mathbf{I} - \epsilon_i), \tag{22}$$

where

$$\mathbf{AB}_i = \mathbf{I} + \epsilon_i. \tag{2'}$$

The associated error relation is

$$\epsilon_{i+1} = -\epsilon_i^2. \tag{23}$$

Convergence of (22) is assured if $\|\epsilon_0\| < 1$[9], where $\mathbf{B}_0 = \mathbf{B}$, the first approximation to $\mathbf{A}^{-1}$. Equation (23) must be considered an optimistic relation since, as in the direct method, error is introduced in the process of matrix multiplication. An analysis has been done by Householder[13] which shows Eq. (22) not to be strictly true. This means that the number of iterations required cannot be accurately determined beforehand and convergence must be tested for each iteration. Also, the iterative method offers no means to improve an associated determinant.

It has been shown[1, 13] that a general nth order iterative procedure is

$$\mathbf{B}_{i+1} = \mathbf{B}_i[\mathbf{I} - \epsilon_i + \epsilon_i^2 - \cdots + (-\epsilon_i)^n], \tag{24}$$

where

$$\epsilon_{i+1} = \epsilon_i(-\epsilon_i)^n. \tag{25}$$

As in the classical or first-order method, Eq. (25) is valid only if all computational errors are zero. A similarity

Table 1 Euclidean norms[a] of matrix inversion calculations done in single-precision arithmetic.

Matrix order	Method[b] 1	2	3	4	5	6
2	1.8×10^{-12}	0	0	0	0	0
3	2.2×10^{-11}	4.5×10^{-12}	1.8×10^{-12}	4.5×10^{-12}	4.5×10^{-12}	4.5×10^{-12}
4	4.9×10^{-9}	1.1×10^{-9}	1.6×10^{-9}	9.8×10^{-10}	9.8×10^{-10}	1.1×10^{-9}
5	8.2×10^{-8}	5.5×10^{-8}	7.8×10^{-8}	9.3×10^{-8}	9.3×10^{-8}	6.8×10^{-8}
6	6.7×10^{-6}	5.9×10^{-7}	4.8×10^{-7}	3.8×10^{-7}	3.8×10^{-7}	4.7×10^{-7}
7	9.8×10^{-5}	2.2×10^{-5}	2.8×10^{-5}	7.1×10^{-5}	7.1×10^{-5}	5.4×10^{-5}
8	1.7×10^{-4}	2.1×10^{-4}	2.1×10^{-4}	2.1×10^{-4}	1.7×10^{-4}	1.7×10^{-4}
9	8.1×10^{-1}	3.2×10^{-2}	3.1×10^{-2}	3.5×10^{-2}	4.9×10^{-2}	3.4×10^{-2}
10	2.4×10^{1}	6.0×10^{0}	7.2×10^{0}	3.7×10^{0}	3.0×10^{0}	2.2×10^{0}

[a]Defined by Eqs. (27) and (28) in the text.
[b]Described in the text on page 415, column 2.

may be seen between Eqs. (19) and (24); the direct method is equivalent to the limiting case of the general iterative scheme used once (i.e., $n \to \infty$), when no other errors are considered.

Comparison of accuracy

In practice $\|\epsilon_0\|$ may be greater than one when a more accurate inverse matrix is desired. To compare the effectiveness of the methods discussed, an inherently ill-conditioned test matrix was chosen. The test matrix is a segment of the Hilbert matrix, which is defined as

$$A_{i,j} = (i + j - 1)^{-1}. \tag{26}$$

Several groups of tests were run to compare these methods using two different computers. On both computers single-precision arithmetic was used except for two tests. In those tests double-precision arithmetic was used only in the accumulation of inner products. The machines used were the Rice University R1 computer using 13 decimal places and an IBM System/360 Model 67 using seven decimal places (both single-precision accuracy). The basic error norm used was the Euclidean norm, given by

$$\|\mathbf{X}\|_E = \left(\sum_{i,j} X_{i,j}^{\,2}\right)^{1/2}. \tag{27}$$

This norm was chosen because it satisfies Eqs. (4) through (7) and is easily computed.

On the R1 computer Eq. (26) was used to define the test matrices through and including the 10×10 case, the maximum Hilbert Matrix usable in single-precision arithmetic. The error in these tests is measured by $\|\epsilon_f\|_E$, where

$$\epsilon_f = \mathbf{A}\,\mathbf{A}_f^{-1} - \mathbf{I} \tag{28}$$

and $\mathbf{A}_f^{-1}$ is the final improved inverse of $\mathbf{A}$ obtained by various methods (see below) for comparison. In each case the initial value $\mathbf{A}_0^{-1}$ was computed using in-place Gauss-Jordan reduction with column pivoting. Both single- and double-precision arithmetic were used for accumulating inner products in two different tests. The results of the first test are shown in Tables 1 and 2 for single- and double-precision arithmetic, respectively.

As numbered in the Tables, the methods used to obtain the matrix inverses for comparison are 1) initial value; 2) single-pass classical, Eq. (22); 3) two-pass classical; 4) single-pass 2nd-order, Eq. (24); 5) single-pass 3rd-order; and 6) direct.

The Euclidean norm values shown in Table 2 are generally smaller than those in Table 1 as a result of using double-precision arithmetic for the inner products in the matrix multiplications. These results are also more representative of the true error in each method since numerical errors are minimized. The zeros in row one of Table 1 are spurious and were caused by mantissa underflow during single-precision inner-product calculations, and it can be assumed that no real improvement was obtained by any method for the 2×2 matrix. Using double-precision arithmetic, however, the higher-order iterative corrections did improve the 2×2 matrix calculation. One can observe from Table 2 that the direct method is about as accurate as all other methods for the test matrices larger than 2×2.

The second series of tests was made to confirm this result and to check that the actual inversion errors being studied are free of unrelated numerical errors. The test matrices here were scaled by their least common factors in order to present the matrices in integer form. This was easily done for the 2×2 through 8×8 matrices, which were investigated using the Model 67. The scaled elements of the Hilbert matrix are thus not truncated as were certain elements of the unscaled form given by Eq. (26). The "exact" inverses of the scaled test matrices

Table 2 Euclidean norms[a] of matrix inversion calculations using double-precision arithmetic for matrix inner products.

Matrix order	Method[b] 1	2	3	4	5	6
2	2.6×10^{-12}	2.0×10^{-12}	9.1×10^{-13}	2.8×10^{-14}	2.8×10^{-14}	2.6×10^{-12}
3	2.0×10^{-11}	1.3×10^{-12}	1.3×10^{-12}	2.2×10^{-12}	2.2×10^{-12}	2.9×10^{-12}
4	4.2×10^{-9}	3.9×10^{-10}	5.0×10^{-10}	1.9×10^{-10}	1.9×10^{-10}	2.4×10^{-10}
5	8.2×10^{-8}	2.7×10^{-8}	3.1×10^{-8}	6.1×10^{-8}	6.1×10^{-8}	5.0×10^{-8}
6	6.7×10^{-6}	9.5×10^{-8}	9.0×10^{-8}	9.9×10^{-8}	9.9×10^{-8}	9.6×10^{-8}
7	9.9×10^{-5}	2.6×10^{-5}	2.2×10^{-5}	2.2×10^{-5}	2.2×10^{-5}	2.1×10^{-5}
8	1.6×10^{-4}	2.3×10^{-5}	1.6×10^{-5}	1.6×10^{-5}	1.7×10^{-5}	3.4×10^{-5}
9	8.2×10^{-1}	4.8×10^{-3}	3.6×10^{-3}	5.1×10^{-3}	3.4×10^{-3}	5.8×10^{-3}
10	2.4×10^{1}	4.7×10^{0}	1.9×10^{0}	6.9×10^{-1}	6.0×10^{-1}	7.2×10^{-1}

[a]Defined by Eqs. (27) and (28) in the text.
[b]Described in the text on page 415, column 2.

Table 3 Relative error measures[a] for matrix inversion calculations done in double-precision arithmetic.

Matrix order	Initial value		Classical (IR) method			Direct method	
	$\|\boldsymbol{\epsilon}'_0\|_r$	$\mathscr{E}(\boldsymbol{\epsilon}'_0)$	f	$\|\boldsymbol{\epsilon}'_f\|_r$	$\mathscr{E}(\boldsymbol{\epsilon}'_f)$	$\|\boldsymbol{\epsilon}'_1\|_r$	$\mathscr{E}(\boldsymbol{\epsilon}'_1)$
2	9.1×10^{-7}	9.5×10^{-7}	1	3.8×10^{-7}	4.8×10^{-7}	3.8×10^{-7}	4.8×10^{-7}
3	6.4×10^{-7}	9.5×10^{-7}	1	2.7×10^{-7}	3.2×10^{-7}	8.3×10^{-7}	9.5×10^{-7}
4	3.0×10^{-7}	4.1×10^{-6}	1	9.6×10^{-8}	1.5×10^{-7}	2.6×10^{-7}	5.8×10^{-7}
5	2.8×10^{-4}	3.7×10^{-4}	2	3.5×10^{-7}	8.7×10^{-7}	6.2×10^{-7}	9.2×10^{-7}
6	8.0×10^{-3}	9.6×10^{-3}	3	6.7×10^{-5}	2.5×10^{-7}	3.2×10^{-7}	2.0×10^{-6}
7	2.1×10^{-2}	2.5×10^{-2}	1	4.5×10^{-4}	5.4×10^{-4}	1.6×10^{-6}	2.4×10^{-4}
8	1.1×10^{-1}	1.2×10^{-1}	1	1.2×10^{-2}	1.4×10^{-2}	1.6×10^{-3}	3.3×10^{-3}

[a]Defined by Eqs. (29) and (32) in the text.

were tabulated in double-precision arithmetic (16 digits) for comparison with the computed inverses.

We compared our direct method with the classical (first-order) iteration in the form of the iterative refinement (IR) method given in Ref. 5. The IR method is an astute combination of Gaussian elimination and classical iteration techniques. If it converges, the IR method can be shown to reduce the maximum relative error column-wise to less than the truncation error of the computer (about 10^{-6} for single-precision arithmetic in the Model 67). Gauss-Jordan reduction was again used to invert the asymmetric correction matrix (2) for the direct method, and double-precision arithmetic was used to accumulate inner products in *all* matrix multiplications. The error measure is given by

$$\mathscr{E}(\boldsymbol{\epsilon}') = \left\| \frac{\|(\boldsymbol{\epsilon}')_j\|_\infty}{\|(\mathbf{A}^{-1})_j\|_\infty} \right\|_\infty, \tag{29}$$

where from Eq. (12)

$$\boldsymbol{\epsilon}' = \mathbf{B} - \mathbf{A}^{-1} = \mathbf{A}^{-1}\boldsymbol{\epsilon}. \tag{30}$$

The subscript j indexes the columns and $\|\mathbf{X}\|_\infty$ indicates the infinity or max norm for a vector $\mathbf{X}$ given by

$$\|\mathbf{X}\|_\infty = \max_j (|X_j|). \tag{31}$$

For further comparison, a relative Euclidean norm was computed that is defined as

$$\|\boldsymbol{\epsilon}'\|_r = \|\boldsymbol{\epsilon}'\|_E / \|\mathbf{A}^{-1}\|_E. \tag{32}$$

Table 3 shows the results for the seven matrices tested using the IR and the direct methods. The value of f, or number of iterations, is given for the IR method; f is always one for the direct method and zero for the initial value. Each iteration was stopped at the lowest value of f for which the following condition was satisfied:

$$\mathscr{E}(\boldsymbol{\epsilon}'_{f+1}) \geq \mathscr{E}(\boldsymbol{\epsilon}'_f). \tag{33}$$

In short, the results in Table 3 confirm previous tests using different criteria, test matrices, and computer. The direct method always improved the computed inverse matrix except for the 3×3 matrix, which was hardly

affected. The two highest-order cases, especially the 8×8 matrix showed the direct method to be superior to the IR or classical improvement method. The failure of the IR method was in fact predicted on the basis of an operating criterion given by Martin et al.[14], which dynamically tests for convergence in the sense given by

$$\|\epsilon_0\| < 1 \tag{34}$$

without actually computing the norm. The criterion is based on successive iterates and therefore takes into account accumulated numerical errors. As was verified, the test predicted that all computed inverse matrices but those of orders seven and eight would be improved by the IR method such that $\mathscr{E}(\epsilon_f')$ would be less than 10^{-6}, the computer roundoff error. The direct method failed this criterion for orders six, seven, and eight, but gave better results than the IR method for orders seven and eight.

Summary

A simple, direct, matrix-inverse improvement method has been analyzed and demonstrated to be useful for extremely ill-conditioned matrices. The direct method was found to improve the inverses of mildly ill-conditioned matrices (Table 3, orders 4, 5, and 6), but not as much as the classical iteration method. It did, however, give better results than all other methods on the most extreme cases (Tables 1 and 2, order 10 and Table 3, order 8). Experience indicated that double-precision arithmetic should be used in accumulating inner products in all matrix multiplications (Table 1 values vs Table 2 values).

Acknowledgments

Part of the work reported here was done at Rice University under AEC contract AT-(40-1)-2572. I thank the staff of the R1 computer (which has since been retired), and W. A. Murray at the Scientific Center for his advice and help in providing more conclusive numerical examples.

References

1. J. Ullman, "The Probability of Convergence of an Iterative Process of Inverting a Matrix," *Ann. Math. Stat.* **15,** 205 (1944).
2. J. Von Neumann and H. H. Goldstine, "Numerical Inverting of Matrices of High Order," *Bull. Am. Math. Soc.* **53,** 1021 (1947).
3. C. F. E. Satterthwaite, "Error Control in Matrix Calculations," *Ann. Math. Stat.* **15,** 373 (1944).
4. L. B. Tuckerman, "On the Mathematically Significant Figures in the Solution of Simultaneous Linear Equations," *Ann. Math. Stat.* **12,** 307 (1941).
5. J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Her Majesty's Stationery Office, London 1963, pp. 80–84, 24–25, and 121–129.
6. A. M. Turing, "Rounding Off Errors in Matrix Processes," *Quart. J. Mech. Appl. Math.* **1,** 287 (1948).
7. J. H. Wilkinson, "Error Analysis of Direct Methods of Matrix Inversion," *J. ACM* **8,** 281 (1961).
8. E. L. Stiefel, *An Introduction to Numerical Mathematics*, Academic Press Inc., New York 1963, pp. 1–21.
9. L. H. Hotelling, "Some New Methods in Matrix Calculation," *Ann. Math. Stat.* **14,** 1 (1943).
10. R. A. Frazer, W. J. Duncan and A. R. Collar, *Elementary Matrices*, Cambridge University Press, London 1947, pp. 120–121.
11. P. S. Dwyer, *Linear Computations*, John Wiley and Sons, Inc., New York 1951, pp. 252–253 and 255–260.
12. D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra*, W. H. Freeman and Co., San Francisco 1963, pp. 159–161.
13. A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell Publishing Co. (now Ginn Co.), Boston 1964, pp. 94–98.
14. R. S. Martin, G. Peters and J. H. Wilkinson, "Iterative Refinement of the Solution of a Positive Definite System of Equations," *Numerische Mathematik* **8,** 203 (1966).

Received April 9, 1971; revised July 6, 1971

The author is located at the IBM Scientific Center, 6900 Fannin Street, Houston, Texas 77001.